ON LATENT STRUCTURE MODELS FOR MEASURING ACHIEVEMENT ON HIERARCHICALLY RELATED SKILLS

Rand R. Wilcox

and

Jennie P. Yeh

CSE Report No. 124

1979

Measurement and Methodology Program
Center for the Study of Evaluation
UCLA Graduate School of Education
Los Angeles, California 90024

The work upon which this publication is based was performed pursuant to a contract with the National Institute of Education, Department of Health, Education and Welfare. Points of view or opinions stated do not necessarily represent official NIE position or policy.

ABSTRACT

This paper derives explicit estimates of the parameters of latent structure models that are special cases of a model described by Dayton and Macready (1976). These estimates may be used as initial trial values in the iterative estimation technique used by Dayton and Macready and there is some possibility that they might be helpful in avoiding erroneous results when a more general model is preferred. Some implications of these models are discussed and a numerical illustration is described.

On Latent Structure Models for Measuring Achievement on Hierarchically Related Skills

Several writers have described various latent structure models which may be useful when measuring the achievement of examinees (see, e.g., Wilcox, 1977a, 1977b; Macready & Dayton, 1977; Knapp, 1977; Brownless & Keats, 1958; Marks & Knoll, 1967, Duncan, 1974). In these models it is assumed that an examinee either "knows" or "does not know" the correct response to a particular achievement test item. Macready and Dayton (1977) arrive at their models by assuming that learning is all or none. As noted by Wilcox (1979), such a view is not required in order to derive these models. This is not to say that the all-or-none view of learning is incorrect; we are merely pointing out that these models can be derived when some other view of learning is preferred.

The purpose of this paper is to derive explicit estimates of the parameters of a latent structure model when the skills represented by the items are hierarchically related. More specifically, we consider two important special cases of a more general model developed by Dayton and Macready (1976). The models described here may be used as an approximation to the Dayton and Macready model and they also provide initial estimates in an iterative estimation procedure when the more general model is being used. In addition, we discuss illustrations and possible implications of using the latent structure models that are described below.

To further motivate this paper we begin by discussing the difficulties associated with iterative estimation techniques.

Iterative Estimates of Parameters

A common statistical problem is that it is often difficult or impossible to obtain explicit estimates of the parameters in a particular probability model. This is the case with Dayton and Macready's hierarchical model as well as their more recent model which they proposed for describing a mastery test (Macready & Dayton, 1977). One solution is to approximate the maximum likelihood estimates of the parameters in the model by employing some iterative technique. Two popular methods are the Newton-Raphson procedure and the method of scoring. For a detailed explanation of this latter procedure, see Rao (1973, section 5g).

Kale (1962) shows that if the initial estimates used in the scoring method or the Newton-Raphson procedure are consistent, both iterative solutions converge to the maximum-likelihood estimates with probability approaching one as the sample size gets large. Kale concludes, therefore, that these iterative procedures are justified if the initial estimates of the parameters are consistent and when the sample size is large (cf. Barnett, 1966). For latent structure models consistent estimates are often not available. For these cases arbitrary initial values are used which may yield erroneous results. Goodman (1974), for example, discusses a wide class of latent structure models for which an iterative estimation scheme is used. His suggestion (p. 218) is to use various initial trial values as estimates of the latent parameters and then choose the iterative solution which minimizes a chi-squared statistic.

Even if consistent estimates of the parameters can be found, difficulties may still occur. An interesting example of this is the beta-binomial distribution. As shown by Griffiths (1973) maximum-likelihood estimates of the parameters of this distribution may be obtained using the Newton-Raphson procedure. Consistent estimates of the parameters are obtained using the method of moments; yet these estimates may yield inadmissible (negative) values for the parameters when they are used as initial trial values in the Newton-Raphson procedure.

The point of this discussion is that there may be technical problems when using iterative estimation schemes and that it is often difficult to determine whether these procedures yield reasonable results. In fact, it is unclear whether these problems apply to Dayton and Macready's model. If they do not apply, then the results described here might be used in an attempt to increase the rate of convergence of the iterative estimation procedure that is being employed. If they do apply and if one of the models described below is believed to be a reasonable approximation to the Dayton and Macready model, the procedures outlined below might be useful for avoiding erroneous results.

Exact Solutions for Hierarchically Related Test Items

For any achievement test item let α be the probability of a correct response from a randomly chosen examinee who guesses. Correspondingly, let β be the probability of an incorrect response given that the examinee knows. We consider two important special cases of the general probability model considered by Dayton and Macready (1976). The first case is $\alpha>0$ and $\beta=0$. Thus, we are assuming that examinees who "know" are always correct. The case

$\alpha=0$ and $\beta>0$ is discussed below. Both of these cases permit explicit estimates of the parameters in the model and so, for the reasons given above, they are of theoretical interest.

We note that the situation considered here is similar to one described by White and Clark (1973) who emphasized the problem of determining whether a particular hierarchy does in fact hold. Here, however, the emphasis is on deriving estimates of the parameters that characterize a particular model.

We also note that our models are similar to the situation discussed by Proctor (1970) which deals with Guttman scales. The situation is a special case of Dayton and Macready's hierarchical model. Again iterative procedures were used to estimate parameters.

We consider two achievement test items and a population of examinees. We assume that each examinee belongs to one of three mutually exclusive states: The examinee knows the answer to both items, the examinee knows the correct response to the first item but not the second, or the examinee does not know the answer to either of the achievement test items. We let k_1, k_2 and $1-k_1 - k_2$ represent the proportion of examinees belonging to these three mutually exclusive states. We further assume that a randomly sampled examinee who knows the correct response is always correct, an examinee who does not know the correct response to item 1 guesses the correct response with probability α_1 and an examinee who does not know item 2 is correct with probability α_2. Finally, local independence between responses is assumed for a given examinee.

We let P_{11}, P_{10}, P_{01} and P_{00} represent the four probabilities associated

with the four possible outcomes on the two achievement test items. A 1 means a correct and a 0 an incorrect response. Thus, P_{10}, for example, is the probability that a randomly chosen examinee is correct on item 1 and incorrect on item 2. From the assumptions made

$$(1a) \qquad P_{11} = k_1 + k_2\,\alpha_2 + (1 - k_1 - k_2)\,\alpha_1\,\alpha_2$$

$$(1b) \qquad P_{10} = k_2\,(1 - \alpha_2) + (1 - k_1 - k_2)\,\alpha_1\,(1 - \alpha_2)$$

$$(1c) \qquad P_{01} = (1 - k_1 - k_2)\,(1 - \alpha_1)\,\alpha_2$$

$$(1d) \qquad P_{00} = (1 - k_1 - k_2)\,(1 - \alpha_1)\,(1 - \alpha_2)$$

For a multinomial distribution, maximum likelihood estimates of the P_{ij}'s (i=1, 2; j=1, 2) are given by

$$(2a) \qquad \hat{P}_{11} = a/N$$

$$(2b) \qquad \hat{P}_{10} = b/N$$

$$(2c) \qquad \hat{P}_{01} = c/N$$

$$(2d) \qquad \hat{P}_{00} = d/N$$

where a, b, c and d are the observed number of examinees corresponding to the four possible response patterns.

If we could express the latent parameters k_1, k_2, α_1 and α_2 in terms of the P_{ij}'s, we would have maximum-likelihood estimates based on the observations a, b, c and d (see, e.g., Zehna, 1966). This is possible for two of the latent parameters. In particular,

$$\alpha_2 = P_{01}\,(P_{01} + P_{00})^{-1}$$

and

$$k_1 = 1 - \frac{P_{00} + P_{10}}{1 - \alpha_2}$$

It follows that

$$\hat{\alpha}_2 = c(c + d)^{-1}$$

and

$$\hat{k}_1 = 1\text{-}bn^{-1}\,(1 - \hat{\alpha}_2)^{-1} - (c + d)n^{-1}$$

are maximum-likelihood estimates of α_2 and k_1 when they are defined.

As for α_1 and k_2, estimates cannot be made based only on the results for this pair of items. The difficulty is that once α_2 and k_1 have been estimated, we are left with one equation in two unknowns. One solution to this problem is to estimate α_1 by pairing item 1 with some other appropriately chosen item, say item 3. If, for example, items 1 and 3 are equivalent, i.e., each examinee knows the answer to both or to neither one, the model described by Wilcox (1977b) might be used to estimate α_1. Alternatively, item 3 might be chosen so that each examinee knows the answer to both items 1 and 3, or the answer to item 1 but not item 3, or the answer to neither of the achievement test items. In this case the hierarchical model described here might be used to estimate the probability of guessing the correct response to item 1. Still another possibility is to administer item 1 on some other occasion in time and to use Wilcox's modification of the Lazarsfeld-Kendall turnover model (Wilcox, 1977a) or the model proposed by Brownless and Keats (1958).

Numerical Illustration

A preliminary version of a mathematics assessment test was administered to 1875 students in grades 6, 7, 8 and 9. All items were multiple-choice having four alternatives from which to choose. For the purpose of illustration we examine three of the skills which were tested. The first (which we call item 1) involves multiplication of integers having signs. The second (item 2) is multiplication of fractions having signs. The resulting 2x2 table of observed frequencies was

		Item 2: 1	Item 2: 0	
Item 1	1	261	404	665
	0	348	862	1210
		609	1266	1875

The estimate of α_2, the probability of guessing item 2, is

$$\hat{\alpha}_2 = 348/1210$$
$$= .29.$$

The estimate of the proportion of examinees who know both items is

$$\hat{k}_1 = 1 - \frac{404}{1875(1 - .29)} - 1210/1875$$
$$= .05.$$

To estimate the probability of guessing item 1 we need to pair it with some other appropriately chosen skill. As indicated earlier, several possibilities exist. Here it is convenient to choose a skill hierarchically related to item 1, namely, multiplication of positive integers (item 3).

The results were

Item 1

		1	0	
Item 3	1	510	829	1339
	0	155	381	536
		665	1210	1875

The estimate of the probability of guessing item 1 is

$$\hat{\alpha}_1 = .29$$

From the item 1 by item 2 contingency table we see that $\hat{P}_{11} = .139$. Substituting $\hat{P}_{11}$, $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{k}_1$ for P_{11}, α_1, α_2 and k_1 in equation (1a) and solving for k_2, we estimate k_2 to be

$$\hat{k}_2 = .02.$$

This suggests that for the population of examinees under study, item 1 and item 2 are nearly equivalent, i.e., an examinee either knows both or neither of the two items. If we assume the two items are equivalent ($k_2 = 0$), then from results given by Wilcox (1977b) we would estimate the parameters with

$$\hat{\alpha}_1 = b(b + d)^{-1} = .32$$

$$\hat{\alpha}_2 = c(c + d)^{-1} = .29$$

$$\hat{k}_1 = 1 - (c + d)(b + d)/(dN) = .05$$

yielding nearly the same results as those obtained above.

Another Application

In the past, scoring procedures which consider errors due to guessing have met with some criticism, the typical approach being to use a correction for guessing formula. As pointed out by Holzinger (1924) this corrected score is perfectly correlated with the observed scores. Lyerly (1951), for example, criticized a technique proposed by Hamilton (1950) because of this result. Chernoff (1962) argues that Lyerly's criticism is an oversimplification. He goes on to propose a scoring method which minimizes a mean squared error expression based on the assumption that examinees guess at random when they do not know the correct response. A theoretical advantage of the models described here or of Dayton and Macready's model is that they allow the possibility of having two different probabilities of guessing associated with two different populations of examinees. Most importantly, it is possible to detect this difference and correct for it by estimating k_1 and k_2 which may lead to different results than those obtained using observed scores. We illustrate this point for the hierarchical models described above.

Suppose item 1 is constructed to represent the skill of interest and that we want to compare two populations of examinees in terms of the proportion who have the skill. The two populations might represent, for example, two methods of instruction. To make the comparison we consider item 2 which is chosen so that every examinee knows both items 1 and 2, item 2 but not item 1, or neither of the two achievement test items. For any pair of items there are four mutually exclusive classes to which an examinee belongs: The three

classes just described or one in which the examinee knows item 2 but not item 1. The proportion of examinees who know item 1 is equal to the proportion who know both items 1 and 2 plus the proportion who know item 2 but not item 1. This last proportion by assumption is equal to zero and so k_1 in our hierarchical model may also be interpreted as the proportion who know item 1. Thus, by estimating k_1 for both populations of examinees, we may make the desired comparison. The "better" of the two populations might have a lower proportion of corrects on item 1 due to differences in the probability of not knowing and guessing the correct response.

Suppose, for example, the results for the first population are:

		Item 2: 1	Item 2: 0	
Item 1	1	58	16	74
	0	9	17	26
		67	33	100

The estimate of k_1 is

$$\hat{k}_1 = .495.$$

For the second population, suppose the results are:

		Item 2: 1	Item 2: 0	
Item 1	1	61	5	66
	0	5	29	34
		66	34	100

As can be seen, the proportion of examinees giving a correct response to both items 1 and 2 is higher for the first sample of examinees suggesting that they are the better of the two groups. However, for the second group,

$$\hat{k}_1 = .6,$$

which implies that the reverse decision should be made. We see, therefore, that the probability models described here might be important since the number-correct observed scores might yield erroneous results.

n-Item Tests

We consider briefly the case of an n-item test where the items represent n skills which may not be related to each other in any particular way. For each of these skills it is possible to estimate the proportion of examinees who have the skill by using one of the procedures described above. When comparing populations of examinees, we may want to use the number of skills the typical examinee has acquired. If we let k_j represent the proportion of examinees who have the jth skill, we may estimate $k = \sum_j k_j$, the expected number of skills known by a randomly selected examinee, with

$$\hat{k} = \sum_j \hat{k}_j.$$

Let $\hat{k}'$ be the estimate of k for a second group of examinees. From the results given in the previous section, we see that it is possible to have $\hat{k} < \hat{k}'$ implying that the second population is "better" than the first while the observed scores imply that the reverse is true.

The Case $\alpha = 0$, $\beta > 0$.

We modify our hierarchical model for situations which, due to the nature of the test item, rule out the possibility of not knowing and guessing the correct response. Completion items are one example where this assumption might be made. As before we assume that we are given two items and that every examinee knows both, knows the first but not the second, or knows neither one. Again we let k_1, k_2, $(1-k_1-k_2)$ represent these proportions. In contrast to our previous model we assume that a randomly selected examinee who knows might give an incorrect response. If we let β_1 and β_2 represent these probabilities for the two test items, the fourfold table of probabilities is as follows:

Item 2

Item 1		1	0
	1	$P_{11} = k_1(1-\beta_1)(1-\beta_2)$	$P_{10} = k_1(1-\beta_1)\beta_2 + k_2(1-\beta_1)$
	0	$P_{01} = k_1\,\beta_1(1-\beta_2)$	$P_{00} = k_1\beta_1\beta_2 + k_2\beta_1 + 1-k_1-k_2$

It follows that

$$\beta_1 = P_{11}(P_{11} + P_{01})^{-1}$$

and so

$$\hat{\beta}_1 = c/(a + c)$$

is a maximum-likelihood estimate of β_1. As before, there is a difficulty in estimating all of the parameters in the model. Again we can solve this problem by pairing item 2 with some other appropriately chosen item. If, for

example, the third item is equivalent to item 2, a model described by Wilcox (1977b) might be used to estimate β_2. Once β_2 is estimated with say $\hat{\beta}_2$, we have an estimate of k_1 and k_2. In particular,

$$\hat{k}_1 = bn^{-1}\ \hat{\beta}_1^{-1}\ \hat{\beta}_2^{-1}$$

and

$$\hat{k}_2 = [bn^{-1} - k_1(1-\hat{\beta}_1)\ \hat{\beta}_2]/(1-\hat{\beta}_1)$$

Concluding Remarks

As demonstrated above, latent structure models might lead to different conclusions than those obtained using the observed responses of examinees. For situations where the models hold, the possibility of reaching different conclusions than those suggested by observed scores will depend on whether the values of α and β vary for different populations of examinees. In practice, the application of a latent structure model resulting in a reverse ranking of two population of examinees might be a rare event. Nevertheless, latent structure models might be of interest since what appears to be a large difference in terms of observed scores might become a small difference when expressed in terms of the parameter k.

REFERENCES

Barnett, V. D. Evaluation of the maximum-likelihood estimator where the likelihood equation has multiple roots. Biometrika, 1966, 53, 151-165.

Brownless, V. T., & Keats, J. A. A retest method of studying partial knowledge and other factors influencing item response. Psychometrika, 1958, 23, 67-73.

Chernoff, H. The scoring of multiple choice questionnaires. Annals of Mathematical Statistics, 1962, 33, 375-393.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Duncan, G. T. An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 1974, 69, 50-57.

Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 1974, 61, 215-231.

Griffiths, D. A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. Biometrics, 1973, 29, 637-648.

Hamilton, C. H. Bias and error in multiple-choice tests. Psychometrika, 1950, 15, 151-168.

Holzinger, K. J. On scoring multiple response tests. Journal of Educational Psychology, 1924, 15, 445-447.

Kale, B. K. On the solution of likelihood equations by iteration processes. The multiparametric case. Biometrika, 1962, 49, 479-486.

Knapp, T. R. The reliability of a dichotomous test-item: A "correlationless" approach. Journal of Educational Measurement, 1977, 14, 237-252.

Lyerly, S. B. A note on correcting for chance success in objective tests. Psychometrika, 1951, 16, 21-30.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Marks, E., & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.

Proctor, C. H. A probabilistic formulation and statistical analysis of Guttman scales. Psychometrika, 1970, 35, 73-78.

Rao, C. R. Linear statistical inference and its applications. New York: Wiley, 1973.

White, R. T., & Clark, R. M. A test of inclusion which allows for errors of measurement. Psychometrika, 1973, 38, 77-86.

Wilcox, R. R. New methods for studying stability. In C. W. Harris, A. Pearlman and R. Wilcox, Achievement test items: Methods of study. CSE Monograph No. 6. Los Angeles: Center for the Study of Evaluation, University of California, 1977(a).

Wilcox, R. R. New methods for studying equivalence. In C. W. Harris, A. Pearlman, & R. Wilcox, Achievement test items: Methods of study. CSE Monograph No. 6. Los Angeles: Center for the Study of Evaluation, University of California, 1977(b).

Wilcox, R. R. An alternative interpretation of three stability models. Educational and Psychological Measurement, 1979, 39(2), 311-315.

Zehna, P. W. Invariance of maximum likelihood estimation. Annals of Mathematical Statistics, 1966, 37, 744.