MEASURING ACHIEVEMENT WITH LATENT STRUCTURE MODELS

Rand R. Wilcox

CSE Report No. 221
1983

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

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

1. MEASUREMENT PHILOSOPHY

The basic assumption in latent class models designed to measure achievement is that an examinee can be described as knowing or not knowing the answer to a test item, and that inferences about an examinee's ability level should take this notion into account. The goals of an n-item test might be to determine how many of the items an examinee knows, which items are known or which are not known, or what proportion of items among a domain of items are known. The problem is that examinees might give the correct response when they do not know, or they might carelessly give the wrong response when they know. Latent class models are an attempt to measure and correct the effects of these errors when addressing a particular measurement problem. Even if some other model is ultimately preferred, such as latent trait model, latent class models are potentially useful.

Currently it appears that correcting for guessing is more important than might have been expected. Moreover, assuming random guessing seems to be an unsatisfactory solution. Consider, for example, the problem of determining the length of a criterion-referenced test where the goal is to determine whether an examinee's percent correct true score or domain score, p, is above or below some known constant p_0. If $p_0=.8$ and n=29 items are used, the probability of correctly determining whether $p \geq p_0$ is at least .9 when $p \geq .9$ or $p \leq .7$, and when the binomial error model is assumed. If random guessing is assumed, nearly 200 items are needed (van den Brink & Koele, 1980), and if one allows for the possibility that guessing is not at random, over 2,600 items are required to attain the same level of

accuracy (Wilcox, 1980). In some cases guessing might be nearly random, but there is empirical evidence that this is generally not the case (Coombs et al., 1956; Bliss, 1980; Cross & Frary, 1977; Wilcox, 1982a, 1982b).

Another way of describing the measurement philosophy of latent class models is that an examinee's test score is a function, in part, of the distractors that are used, and that it is important to take this effect into account. In the past this problem was ignored, probably because there were no reasonable ways of dealing with it, and because it was not clear just how serious this problem was. Now, however, there are several ways of measuring and correcting the effects of distractors. It might appear that some latent trait models deal with guessing, but in fact latent trait models ignore the errors that are of concern here. Thus, these errors might have a serious effect on how latent trait models are used and interpreted. Wainer and Wright (1980) as well as Mislevy and Bock (1982) examined certain aspects of how guessing affects latent trait models, but the type of guessing examined here is different.

2. THE MODELS AND THEIR ASSUMPTIONS

Generally latent class models are based on assumptions about how examinees behave when responding to an item, or how items are related to one another, or the manner in which tests are administered. While a general description of latent class models is possible, such a description is not given here. Instead attention is focused on those models that seem to have the most practical value.

A Latent Structure Model for Answer-Until-Correct Tests

This section assumes that an examinee responds to a multiple-choice test item according to an answer-until-correct (AUC) scoring procedure. This means that if an examinee chooses an incorrect response, another response is chosen, and this process continues until the correct response is identified.

AUC tests are easily administered in the classroom using especially designed answer sheets where the examinee erases a shield corresponding to a particular alternative. (These answer sheets are available commercially, for example, through Van Valkenburg, Nooger and Neville in New York, N.Y., and they are relatively inexpensive.) If the letter under the shield indicates an incorrect response, the examinee erases another shield, and this continues until the correct shield is erased.

Consider a population of examinees, and let ζ_i be the proportion of the examinees who can eliminate i distractors from consideration. That is, because of partial information, some of the examinees will rule out some of the distractors without knowing the correct response. If there are t alternatives from which to choose, and if the examinee can eliminate t-1 distractors from consideration, the examinee is said to know the correct response. Thus, ζ_{t-1} is the probability that a randomly sampled examinee knows the correct response. Note that no distinction is made between examinees who can eliminate all the distractors via partial information and those that know. In other words, an examinee might choose the correct response, not because the correct answer is known, but because

the test constructor was unable to produce at least one effective distractor. Thus, it is assumed that at least one effective distractor is being used, and presumably this problem can be minimized by choosing t to be reasonably large. Of course the crucial step is finding someone who can write effective distractors.

As alluded to earlier, it is assumed that among the examinees who do not know, some might be able to eliminate one or more distractors from consideration via partial information. It is further assumed that once these distractors are eliminated, the examinee guesses at random among the alternatives that remain. Hence, if p_i is the probability of a correct response on the ith attempt of the item (i=1,...,t),

$$p_i = \sum_{j=0}^{t-i} \zeta_j/(t-j) \tag{2.0}$$

For example, if t=3

$$p_1 = \zeta_0/3 + \zeta_1/2 + \zeta_2$$

$$p_2 = \zeta_0/3 + \zeta_1/2,$$

and

$$p_3 = \zeta_0/3 .$$

In general, the proportion of examinees who know the correct response is

$$\zeta_{t-1} = p_1 - p_2 . \tag{2.1}$$

The model implies that

$$p_1 \geq p_2 \geq \cdots \geq p_t, \tag{2.2}$$

and this can be tested by applying results in Robertson (1978). Empirical investigations (Wilcox, 1982a, 1982b) suggest that (2.2) will usually hold.

The next section describes how one might proceed when (2.2) appears to be unreasonable.

For N randomly sampled examinees, let x_i be the number who get the correct response on the ith attempt. The the x_i's have a multinomial distribution given by $\binom{N}{x} p_1^{x_1} \ldots p_t^{x_t}$ where $\binom{N}{x} = N!/(x_1! \ldots x_t!)$, $\sum x_i = N$, $0 \leq p_i \leq 1$, and $\sum p_i = 1$. An unbiased maximum likelihood estimate of p_i is just x_i/N, and so

$$\hat{\zeta}_{t-1} = (x_1 - x_2)/N \tag{2.3}$$

is a maximum likelihood estimate of ζ_{t-1}, the proportion of examinees who know the correct response. Semantically, if we compute the proportion of examinees who get the item correct on the first attempt, and then subtract the proportion who get it right on the second attempt, we have an estimate of the probability that the typical examinee will know the answer.

Note that ζ_{t-1} given by (2.3) can be negative, but ζ_{t-1} is positive when the model is assumed to be true. This can be corrected by simply estimating ζ_{t-1} to be zero when $\hat{\zeta}_{t-1} < 0$. From Barlow et al. (1972), a maximum likelihood estimate of ζ_{t-1} under the assumption that (2.2) holds can be had by applying the pool-adjacent violators algorithm.

A Misinformation Model

The previous section assumed that the inequality in equation (2.2) is true, but experience indicates that occasionally this will not be the case. In this event a misinformation model may be appropriate. Of course

for some items an investigator might suspect a misinformation model is needed before any test data are collected in which case the results in this section might be applied without testing (2.2).

As will soon become evident, there is no specific misinformation model, but rather a class of models that might be used. The choice from among these models will depend on what seems to be a reasonable assumption about how examinees behave. At the moment there are no empirical procedures to aid a test constructor when choosing from among the various misinformation models. So far, however, this does not seem to be a serious problem.

To better understand how to apply these models, consider the following test item.

> When a block of iron is heated until it is red hot, it gets bigger. If the iron weighs 20 lbs. at room temperature, how much will it weigh when red hot?
>
> 1) 19.8 lbs. 2) 20 lbs. 3) 20.1 lbs. 4) 20.5 lbs. 5) 20.61 lbs.

This item is similar to one investigated in Wilcox (1982b) where the examinees were approximately 14 years old. The point is that it seems reasonable to suspect that some examinees will choose from among the last three alternatives because they believe the iron weighs more when it expands. The goal then is to devise a model that takes this behavior into account.

In this section it is assumed that the examinees belong to one of three mutually exclusive groups: 1) they know the item, 2) they have misinformation, 3) or they do not know, do not have misinformation, and guess at random. For examinees with misinformation, it is also assumed

that they will choose c specific incorrect alternatives before choosing the correct response. At the moment there is no empirical method for choosing c; this must be done based on what seems reasonable for the item being used. For example, in the item described above, c=3 would be considered. In some cases the resulting latent structure model can be checked with a goodness-of-fit test, but as will be seen this is not always the case.

For the population of examinees being tested, let ζ be the proportion of examinees who know, ν_1 be the proportion who do not know, do not have misinformation and guess at random, and let ν_2 be the proportion who have misinformation. If an AUC scoring procedure is used, and if p_i is defined as before, then for c=3 and t=5

$$p_1 = \zeta + \nu_1/5 \tag{2.4}$$

$$p_2 = \nu_1/5 \tag{2.5}$$

$$p_3 = \nu_1/5 \tag{2.6}$$

$$p_4 = \nu_2 + \nu_1/5 \tag{2.7}$$

$$p_5 = \nu_1/5 \tag{2.8}$$

Thus, $\zeta = p_1 - p_2$ as before and ζ is estimated with $(x_1-(x_2 + x_3 + x_5)/3)/N$.. The model can be tested with the usual chi-square test, and it gave a good fit to the data in Wilcox (1982b).

More generally, for arbitrary c,

$$p_1 = \zeta + \nu_1/t \tag{2.9}$$

$$p_{c+1} = \nu_2 + \nu_1/t \tag{2.10}$$

and

$$p_i = \nu_i/t, \; i \neq t, \; c + 1. \tag{2.11}$$

Slight generalizations of the model may be possible. Suppose, for example, c=3 and t=5, as in equations (2.4)-(2.8), but for examinees with misinformation, let ν_3 be the proportion of examinees who choose the correct response once c=3 alternatives are eliminated. Then p_5 and p_4 take the more general form

$$p_4 = \nu_3\nu_2 + \nu_1/t \tag{2.12}$$

and

$$p_5 = (1-\nu_3)\nu_2 + \nu_1/t \tag{2.13}$$

Now, however, a goodness-of-fit test is no longer possible because there are zero degrees of freedom.

Equivalent and Hierarchically Related Items, and Related Latent Structure Models

In recent years, several investigators have proposed models based on the notion of equivalent or hierarchically related items. Two items are said to be equivalent if examinees know both or neither one. If in addition, there are examinees who know the first but not the second, the items are hierarchically related. As argued by Molenaar (1981), clearly there are situations where it may be difficult or impossible to generate equivalent items. However, experience suggests that there are situations where one of these assumptions might be reasonable (e.g., Macready & Dayton, 1977; Harris & Pearlman, 1978; Harris et al., 1980).

It should be mentioned that in some instances a test consisting of hierarchically related items is considered to be desirable and the goal is to measure the extent to which a test has this property. Put another

way, the goal is to determine the extent to which the items on a test form a Guttman scale. One such measure was proposed by Cliff (1977). (See also Harnisch & Linn, 1981.)

The simplest model consists of two equivalent items, and it arises as follows. Let ζ be the proportion of examinees who know both items. In contrast to earlier sections, a conventional scoring procedure is used. That is, examinees get only one attempt at an item, and the item is scored either correct or incorrect. Let p_{ij} be the probability of the response pattern ij (i=0,1; j=0,1) where a 0 means incorrect, and a 1 means correct. Thus, p_{10} represents the probability of a correct-incorrect response for a randomly sampled examinee. If β_1 is the probability of correctly guessing the response to the first item when the randomly sampled examinee does not know, and if β_2 is the corresponding probability on the second item, and if local independence holds (i.e., given an examinee's latent state, the responses are independent) then

$$p_{11} = \zeta + (1-\zeta)\beta_1\beta_2$$

$$p_{10} = (1-\zeta)\beta_1(1-\beta_2)$$

$$p_{01} = (1-\zeta)\beta_2\,(1-\beta_1)$$

$$p_{00} = (1-\zeta)(1-\beta_1)(1-\beta_2).$$

Solving ζ, β_1, and β_2 yields

$$\beta_1 = \frac{p_{10}}{p_{10} + p_{00}}$$

$$\beta_2 = \frac{p_{01}}{p_{01} + p_{00}}$$

it can be seen that

$$p_{12} = p_{21} = p_{22}$$

$$p_{13} = p_{23}$$

and

$$p_{31} = p_{23} .$$

For recent results on testing these equalities, see Smith et al. (1979), and Wilcox (1982e).

Hartke (1978) describes another approach based on latent partition analysis, and an index proposed by Baker and Hubert (1977) might also be useful.

Hierarchically Related Items

Dayton and Macready (1976, 1980) describe very general latent structure models for handling hierarchically related items. Again these models can be used to measure guessing, and they have the advantage of including other errors at the item level such as ζ = Pr(incorrect | examinee knows). The model for AUC tests essentially sets $\zeta = 0$, but the practical implications of this have not been established.

As was the case for equivalent items, estimating the parameters in the model requires iterative techniques. In some instances simple (closed form) estimates exist (e.g., Wilcox, 1980b), but these models make certain assumptions that may be unreasonable in many situations.

3. STRENGTHS AND WEAKNESSES OF LATENT CLASS MODELS

Latent class models have three primary strengths. First, it now

appears that one of two models can be used to explain the observed responses to a multiple-choice test item (Wilcox, 1982b). These models are an oversimplification of reality (as are all models), but they seem to give a good approximation of how examinees behave when taking a test. Of course, future investigations might reveal that more complex models are really needed, but so far this does not appear to be the case.

The second strength is that many measurement problems can now be solved that were previously impossible to address. In particular, these models correct for guessing, or measure the effects of guessing which in turn improves the accuracy of tests and measurement techniques. Note that the nature of guessing in latent class models is different from the guessing parameter in latent trait models (Wilcox, 1982c).

Third, even if some other model is ultimately preferred, a latent class model may be useful, for example, when estimating the item parameters in a latent trait model.

A weakness of latent class models is that certain technical problems still need to be solved. These include better ways of scoring an n-item test, testing the model used in Wilcox (1982e), and finding a strong true-score model that is reasonable when the model in Wilcox (1982a) gives a poor fit to data. Also, some examinees may give an incorrect response when they know, but the seriousness of this problem is not well understood.

4. PRESENT AREAS OF APPLICATION

This section outlines some of the measurement problems that can now be solved with latent class models.

The Accuracy of an Item and the Effectiveness of Distractors

In addition to estimating the proportion of examinees who know the item, the latent structure models for AUC tests can be used to estimate the probability of correctly determining whether a typical examinee knows the item. More specifically, assume it is decided that an examinee knows the correct response if the correct answer is given on the first attempt (i.e., a conventional scoring procedure is used). For a randomly sampled examinee, the probability of correctly determining whether he/she knows is just $\tau = 1-p_2$ (Wilcox, 1981a), and this is estimated with $\hat{\tau} = 1-x_2/N$. Note that when (2.2) is assumed $0 \leq p_2 \leq \frac{1}{2}$, in which case $\frac{1}{2} \leq \tau \leq 1$. The parameter τ is a function of two important quantities. The first is the proportion of examinees who know the answer, i.e., ζ_{t-1}, and the second is the effectiveness of the distractors among the examinees who do not know. To see this more clearly, note that

$$\tau = \zeta_{t-1} + \sum_{i=2}^{t} p_i \, . \tag{4.1}$$

When ζ_{t-1} is close to one the item accurately reflects the true latent state of the examinees because presumably examinees who know will choose the correct response on their first attempt. As ζ_{t-1} moves closer to zero, the accuracy depends more on the effectiveness of the distractors. Thus, it may be important to determine how well distractors are perform-

ing among the examinees who do not know.

It can be shown that the distractors are most effective when guessing is at random which corresponds to

$$p_2 = p_3 = \ldots = p_t \tag{4.2}$$

(Wilcox, 1981a). This suggests (4.2) be tested, and/or we estimate how "far away" the p_i values are from the ideal case where (4.2) holds.

Testing (2.3) can be accomplished by noting that the conditional distribution of $x_2,\ldots,x_t$ given x_1 is multinomial with parameters $N-x_1$ and $p_i/(1-p_1)$, $i = 2,\ldots,t$. Thus, the usual chi-square test can be applied. That is, compute

$$X^2 \quad \sum_{i=2}^{t} \frac{(x_i-(n-x_1)/(t-1))^2}{(N-x_1)/(t-1)} \tag{4.3}$$

If X^2 is greater than or equal to the $100(1-\alpha)$ percentile of the chi-square distribution with t-2 degrees of freedom, reject the hypothesis that (4.2) holds. For recent results on using (4.3), see Chacko (1966), Smith et al. (1979), Wilcox (1982e).

Empirical results indicate that guessing will not be at random. Thus, a more interesting question might be to determine whether the distractors are "close" to the ideal situation where (4.2) holds. The first step in solving this problem is to choose a measure of how unequal the p_i values are ($i = 2,\ldots,t$). Many such measures have been proposed which have similar properties (e.g., Marshall & Olkin, 1979; Bowman et al., 1971). One of these is the entropy function which was used by Wilcox (1982a), and another is Simpson's measure of diversity (Simpson, 1949) given by

$$\sum_{i=2}^{t} [p_i/(1-p_1)]^2 .$$

Writing (4.3) as

$$-(N-x_1) + \frac{t-1}{N-x_1} \sum_{i=2}^{t} x_i^2 ,$$

it is seen that the usual maximum likelihood estimate of Simpson's measure of diversity, namely, $\sum_{i=2}^{t} (x_i/(N-x_1))^2$, is a simple linear transformation of X^2. Since X^2 is better known than Simpson's measure of diversity, X^2 will be used here.

It is helpful to note that the smallest possible value for X^2 is

$$L = \frac{t-1}{n-x_1} [(n-x_1)(2r+1) - (t-1)r(r+1)] - n+x_1 \qquad (4.4)$$

where r is the largest integer satisfying $r(t-1) \leq n-x_1$ (Dahiya, 1971). The maximum value is

$$M = (n-x_1)(t-2) \qquad (4.5)$$

(Smith et al., 1979). The closer X^2 is to M, the more effective are the distractors. Since L and M are known, the relative extent to which X^2 is close to M can be determined. In particular,

$$E = (X^2-L)/(M-L)$$

measures the effectiveness of the distractors being used, where $0 \leq E \leq 1$. If E=0, the distractors are as effective as possible in determining whether an examinee knows the correct response. As E approaches 1, the distractors become less effective.

Comparing Two Items

If the AUC model is assumed, and if independent estimates of the p_i values for two items are available, it is possible to test the hypothesis that one of the items is at least as effective as the second by applying results in Robertson and Wright (1981). The null hypothesis of interest here is that $\sum_{i=2}^{k} p_i/(1-p_1) \geq \sum_{i=2}^{k} p_i'/(1-p_1')$, k=2,...,t-2 where p_i' is the p_i value for the second item. Let τ_1 and τ_2 be the value of τ for two items. Another way of comparing two items is to test whether the first item is better than the second by testing whether $\tau_1 \geq \tau_2$. In effect this approach compares the overall effectiveness of the two items in terms of the population of examinees, while the approach previously described is to compare the effectiveness of the distractors among the examinees who do not know.

Characterizing Tests

Let τ_i be the value of τ for the ith item on an n-item test. A natural way of describing the accuracy of a test is to use $\tau_s = \sum_{i=1}^{n} \tau_i$. This is the expected number of correct decisions about whether a typical (randomly sampled) examinee knows the answer to the items on a test. If, for example, $\tau_s = 7$ and n = 10, then on the average, 7 correct decisions would be made about whether an examinee knows the answer to an item, but for 3 of the items it would be decided that the examinee knows when in fact he/she does not.

Estimating τ_s is easily accomplished using previous results. In

particular, for a random sample of N examinees, let $x_{ij} = 0$ if the jth examinee gets the ith item correct on the second attempt; otherwise $x_{ij} = 1$. Then

$$\hat{\tau}_s = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N} x_{ij}$$

is an unbiased estimate of τ_s.

The k Out of n Reliability of a Test

Once test data is available, the question arises as to how certain we can be that τ_s is large or small. That is, we want to estimate the $Pr(\hat{\tau}_s \geq \tau_0)$(cf. Tong, 1978). This problem is similar to one found in the engineering literature where the goal is to estimate the k out of n reliability of a system. Bounds on this probability can be estimated without assuming anything about $cov(x_{ij}, x_{i'j'})$ (Wilcox, 1982e). The procedure is outlined below.

Let $z_i=1$ if a correct decision is made about whether a randomly sampled examinee knows the ith item on a test; otherwise $z_i=0$. For a randomly sampled examinee $Pr(z_i=1) = \tau_i$. Note that from previous results $Pr(z_i=1) = Pr(x_{ij}=1)$. The k out of n reliability of a test is defined to be

$$\rho_K = Pr(\sum z_i \geq K)$$

This is the probability that for a typical examinee, at least k correct decisions are made among the n items on a test. By a correct decision is meant the event of correctly determining whether the examinee knows an item. Knowing ρ_k yields additional and important information about

the accuracy of a test. An estimate of ρ_k is not available unless $\text{cov}(z_i,z_j)=0$, or n, the number of items, is small. (See Wilcox, 1982g, 1982j.)

For any two items, let p_{km} be the probability that a randomly selected examinee chooses the correct response on the kth attempt of the first item, and the mth attempt of the second. (It is assumed that both items are administered according to an AUC scoring procedure.) Let $\kappa_{ij}(i=0,\ldots,t-1;\ j=0,\ldots,t-1)$ be the proportion of examinees who can eliminate i distractors on the first item and j distractors on the second. Then, under certain mild independence assumptions

$$p_{km} = \sum_{i=0}^{t-k} \sum_{j=0}^{t-m} \kappa_{ij}/[(t-i)(t-j)] \quad .$$

The equation makes it possible to express the κ_{ij}'s in terms of the p_{km}'s which in turn makes it possible to estimate κ_{ij} for any i and j.

Next let ε be the probability that for both items, a correct decision is made about an examinee's latent state. It can be seen that

$$\varepsilon = \kappa_{t-1,t-1} + 1-p_{11}$$

and so ε can also be estimated.

For the ith and jth item on a test, let ε_{ij} be the value of ε, and define

$$S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \varepsilon_{ij}$$

$$U_K = \tau_S - K$$

where τ_S was previously defined to be $\Sigma\tau_i$ and

$$V_K = (2S - K(K-1)/2).$$

Then from Sathe et al. (1980)

$$\rho_K \geq (2V_{K-1} - (K-2))\, U_{K-1}/[n(n-K+1)] .$$

If $2V_{K-1} \leq (n+K-2)U_{K-1}$

$$\rho_K \geq \frac{2((K^*-1)U_{K-1} - V_{K-1})}{(K^*-K)(K^*-K+1)}$$

where $K^* + K-3$ is the largest integer in $2V_{K-1}/U_{K-1}$. Two upper bounds are also available. The first is

$$\rho_K \leq 1 + ((n+K-1)U_K - 2V_K)/Kn$$

and the second is that if $2V_k \leq (K-1)U_K$;

$$\rho_K \leq 1 - 2\,\frac{(K^*-1)U_K - V_K}{(K-K^*)(K-K^*+1)}$$

where $K^* + K - 1$ is the largest integer in $2V_K/U_K$.

What these results mean is that we can estimate quantities that indicate whether ρ_k is large or small. For example, suppose the right side of the third to last inequality is estimated to be .9, and that $2V_{k-1} \leq (n+K-2)U_{k-1}$. This does not yield an exact estimate of ρ_k but it does say that ρ_k is estimated to be at least .9. Thus, this would indicate that the overall test is fairly accurate. If, for example, the above inequalities indicate that $\rho_k \leq .95$ and $\rho_k \geq .1$, this does not give very useful information about whether ρ_k is reasonably large. If $\rho_k \leq .1$ we have a poor test.

Estimating the Proportion of Items an Examinee Knows

It is a simple matter to extend previous results to situations when

a single examinee responds to items randomly sampled from some item domain. For example, let q_i be the probability of a correct response on the ith attempt of a randomly sampled item. Let $\gamma_i (i=0, \ldots, t=1)$ be the proportion of items for which the examinee can eliminate i distractors. It is assumed that each item has at least one effective distractor, so γ_{t-1} is the proportion of items the examinee knows. It follows that

$$q_i = \sum_{j=0}^{t-i} \gamma_j/(t-i)$$

which is the same as equation (2.0) where p_i and ζ_i are replaced with q_i and γ_i. In fact, all previous results extend immediately to the present case.

Criterion-Referenced Tests

A common goal of a criterion-referenced test is to sort examinees into two categories. (See Hambleton et al., 1978a; Berk, 1980; and the 1980 special issue of Applied Psychological Measurement.) Frequently these categories are defined in terms of some true score, and here the true score of interest is γ_{t-1}, the proportion of items in an item domain that an examinee knows. The goal is to determine whether γ_{t-1} is larger or smaller than some predetermined constant, say γ'.

It is known that guessing can seriously affect the accuracy of a criterion-referenced test (van den Brink & Koele, 1980). Moreover, assuming random guessing can be highly unsatisfactory (Wilcox, 1980c). Another advantage of the AUC scoring model is that it substantially reduces this problem (Wilcox, 1982c). For some results on comparing

by the investigator. In this case the joint probability function of $y_1,\ldots,y_t$ is

$$\Gamma(y_0)(\sum y_i \quad I_{[y_i=n_i]}) \prod_{i=1}^{t} p_i^{y_i}/y_i!$$

where I is the usual indicator function given by

$$I_{[y_i=n_i]} = \{1, \text{ if } y_i=n_i \quad ; 0, \text{ if otherwise}\}.$$

For the special case $n_1=n_2=\ldots=n$, the probability function becomes

$$n\Gamma(y_0) \prod_{i=1}^{t} p_i^{y_i}/y_i!$$

which has the same form as the negative multinomial except that for some j, $y_j=n$, and $0 \leq y_i \leq n-1$, $i\neq j$.

The maximum likelihood estimate of q_i is $\hat{q}_i = y_i/y_0$, so the maximum likelihood estimate of γ_{t-1}, the proportion of items an examinee knows, is $\hat{\gamma}_{t-1} = \hat{q}_1 - \hat{q}_2$ (Zehna, 1966). If the model is assumed to hold, $\hat{\gamma}_{t-1}$ may not be a maximum likelihood estimate. Instead one would estimate γ_{t-1} to be zero when $\hat{\gamma}_{t-1} \leq 0$; if the estimates of q_i (i=1,...,t) do not satisfy the inequality $q_1 \geq q_2 \geq \ldots \geq q_t$ apply the pool-adjacent-violators algorithm (Barlow et al., 1972).

Wilcox (1982d) shows that if the goal is to compare γ_{t-1} to the known constant γ', as in criterion-referenced testing, and if $\gamma_{t-1} \geq \gamma'$ is decided if and only if $\hat{\gamma}_{t-1} \geq \gamma'$ the sequential and closed sequential procedures have the same level of accuracy. Moreover, it appears that the closed sequential procedures nearly always improves upon the more conventional fixed sample approach. More recently Wilcox (1982f) proposed two tests of $q_1=\ldots=q_t$, and methods of determining the moments of

the distribution were also described.

A Strong True Score Model

Strong true score models attempt to relate a population of examinees to a domain of items. In many situations an item domain does not exist de facto, in which case strong true score models attempt to find a family of probability functions for describing the observed test scores of any examinee, and simultaneously to find a distribution that can be used to describe the examinees' true score.

Perhaps the best known model is the beta-binomial. If y is the number of correct responses from an examinee taking an n-item test, it is assumed that for a specific examinee, the probability function of y is

$$\binom{n}{y} q^y (1-q)^{n-y}.$$

For the population of examinees, it is assumed that the distribution of q is given by

$$g(q) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} q^{r-1} (1-q)^{s-1}$$

where $r > 0$ and $s > 0$ are unknown parameters that are estimated with observed test scores. Apparently Keats (1951) was the first to consider this model in mental test theory.

The beta-binomial model has certain theoretical disadvantages, but experience suggests that it frequently gives good results with real data. A review of these results is given by Wilcox (1981d). However, the model does not always give a good fit to data, and some caution should be exercised (Keats, 1964). In the event of a poor fit, a gamma-Poisson model might be considered (Wilcox, 1981d).

When the beta-binomial is assumed, many measurement problems can be solved. These include equating tests by the equipercentile method, estimating the frequency of observed scores when a test is lengthened, and estimating the effects of selecting individuals on a fallible measure (Lord, 1965). Other applications include estimating the reliability of a criterion-referenced test (Huynh, 1976a), estimating the accuracy of a criterion-referenced test (Wilcox, 1977c), and determining passing scores (Huynh, 1976b).

A problem with the beta-binomial model is that it ignores guessing. Attempts to remedy this problem are summarized by Wilcox (1981d), but all of these solutions now appear to be unsatisfactory in most situations. This is unfortunate because it means that a slightly more complex model must be used. More recently, however, Wilcox (1982a, 1982b) proposed a generalization of the beta-binomial model that takes guessing into account, and which gives a reasonably good fit to data.

Some Miscellaneous Applications of Latent Structure Models

Several applications of latent structure models have already been described, and there are several other situations where they may be useful. For example, Ashler (1979) derives an expression for the biserial correlation coefficient that includes ς_{t-1}, the proportion of examinees who know an item. Wilcox (1982g) discusses how to empirically determine the number of distractors needed on a multiple-choice test item, and Knapp (1977) discusses a reliability coefficient based on the latent state

point of view. (See also Frary, 1969.) Macready and Dayton (1977) illustrate how the models can be used to determine the number of equivalent items needed for measuring an instructional objective, and Emrick (1971) shows how the models might be used to determine passing scores. Note that Emrick's estimation procedure is incorrect (Wilcox & Harris, 1977), but this is easily remedied using the estimation procedures already mentioned; closed form estimates are given by van der Linden (1981).

5. POSSIBLE EXTENSIONS AND CONTROVERSIAL ISSUES

The AUC models assumed that examinees eliminate as many distractors as they can and then guess at random from among the alternatives that remain. A recent empirical investigation suggests that the random guessing portion of this assumption will usually give a reasonable approximation of reality (Wilcox, 1982k). No doubt there will be cases where this assumption is untenable in which case there are no guidelines on how to proceed.

A theoretical advantage of the latent structure model based on equivalent or hierarchically related items is that they included not only guessing, but errors such as Pr(incorrect response | examinee knows). The practical implications of this are not well understood.

Wilcox (1981a) mentions that under an item sampling model for AUC tests, an examinee with partial information can improve his/her test score by choosing a response, and if it is incorrect, deliberately choose another incorrect response. Thus, if $(y_1-y_2)/n$ is used to estimate ζ_{t-1}, the estimate would be higher for such an examinee because y_2

is lower. Four points should be made. First, this problem can be partially corrected by estimating the q_i's with the pool-adjacent-violators algorithm (Barlow et al., 1972, pp. 13-15). Second, if an examinee is acting as described, it is still possible to correct for guessing by applying the true score model proposed by Wilcox (1982a). If it gives a good fit to data, estimate ζ_{t-1} to be $q_1-(1-q_1)\xi(q_1)$. The third point is that there is no indication of how serious this problem might be. Finally, a new scoring procedure is being examined that might eliminate the problem.

It has been argued (e.g., Messick, 1975) that tests should be homogeneous in some sense. Frequently this means that at a minimum, a test should have a single factor. A sufficient condition for the best known latent trait models (see e.g., Lord, 1980; Wainer et al., 1980; Hambleton et al., 1978b; Choppin, 1983) is that this assumption be met (cf. McDonald, 1981). In general, the latent structure models described in this paper do not require this assumption. One exception is the equivalent item model. (See Harris & Pearlman, 1978.) The point is that in this paper, no stand on this issue is needed, i.e., it is irrelevant whether a test is homogeneous when applying, say, the answer-until-correct scoring procedure, or the corresponding strong true-score model.

Wainer and Wright (1980) and Mislevy and Bock (1982) have studied the effects of guessing on latent trait models, but these investigations do not take into account the results and type of guessing described here. If guessing proves to be a problem, perhaps latent class models can be of use when latent trait models are applied.

References

Ashler, D. Biserial estimators in the presence of guessing. Journal of Educational Statistics, 1979, 4, 325-356.

Baker, F. B., & Hubert, L. J. Inference procedures for ordering theory. Journal of Educational Statistics, 1977, 2, 217-233.

Barlow, R., Bartholomew, D., Bremner, J., & Brunk, H. Statistical inference under order restrictions. New York: Wiley, 1972.

Berk, R. Criterion-referenced measurement. Baltimore: The Johns Hopkins University Press, 1980.

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 1980, 17, 147-153.

Bowman, K., Hutcheson, K., Odum, E., & Shenton, L. Comments on the distribution of indices of diversity. In G. Patil, E. Pielou, & W. Waters (Eds.), International Symposium on Statistical Ecology, Vol. 3. University Park: Pennsylvania State Press, 1971.

Chacko, V. J. Modified chi-square test for ordered alternatives. Sankhya, 1966, Ser. B., 28, 185-190.

Choppin, B. The Rasch model for item analysis. CSE Report No. 218. Los Angeles: UCLA Center for the Study of Evaluation, 1983.

Cliff, N. A theory of consistency of ordering generalizable to tailored testing. Psychometrika, 1977, 42, 375-399.

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial information. Educational and Psychological Measurement, 1956, 16, 13-27.

Cross, L. H., & Frary, R. B. An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 1977, 14, 313-321.

Dahiya, R. C. On the Pearson chi-squared goodness-of-fit test statistic. Biometrika, 1971, 58, 685-686.

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

Dayton, C. M., & Macready, G. B. A scaling model with response errors and intrinsically unscalable respondents. Psychometrika, 1980, 45, 343-356.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Frary, R. B. Reliability of multiple-choice test scores is not the proportion of variance which is true variance. Educational and Psychological Measurement, 1969, 29, 359-365.

Goodman, L. A. On the estimation of parameters in latent structure analysis. Psychometrika, 1979, 44, 123-128.

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

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and application. Review of Educational Research, 1978, 48, 467-510. (b)

Harnisch, D. L., & Linn, R. L. Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. Journal of Educational Measurement, 1981, 18, 133-146.

Harris, C. W., Houang, R. T., Pearlman, A. P., & Barnett, B. Final report submitted to the National Institute of Education. Grant No. NIE-G-78-0085, Project No. 8-0244, 1980.

Harris, C. W., & Pearlman, A. An index for a domain of completion or short answer items. Journal of Educational Statistics, 1978, 3, 285-304.

Hartke, A. R. The use of latent partition analysis to identify homogeneity of an item population. Journal of Educational Measurement, 1978, 15, 43-47.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264. (a)

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78. (b)

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

Keats, J. A. A statistical theory of objective test scores. Melbourne: A.C.E.R., 1951.

Keats, J. A. Some generalizations of a theoretical distribution of mental test scores. Psychometrika, 1964, 29, 215-231.

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

Lord, F. M. A true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, New Jersey: Erlbaum, 1980.

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

Marshall, A. W., & Olkin, I. Inequalities: Theory of majorization and its applications. New York: Academic Press, 1979.

McDonald, R. P. The dimensionality of tests. British Journal of Mathematical and Statistical Psychology, 1981, 34, 100-117.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Mislevy, R. J., & Bock, R. D. Biweight estimates of latent ability. Educational and Psychological Measurement, 1982, 42, 725-737.

Molenaar, I. On Wilcox's latent structure model for guessing. British Journal of Mathematical and Statistical Psychology, 1981, 34, 79-89.

Robertson, T. Testing for and against an order restriction on multinomial parameters. Journal of the American Statistical Association, 1978, 73, 197-202.

Robertson, T., & Wright, F. T. Likelihood ratio tests for and against a stochastic ordering between multinomial populations. Annals of Statistics, 1981, 9, 1248-1257.

Sathe, Y. S., Pradhan, M., & Shah, S. P. Inequalities for the probability of the occurrence of at least m out of n events. Journal of Applied Probability, 1980, 17, 1127-1132.

Simpson, E. Measurement of diversity. Nature, 1949, 163, 688.

Smith, P. J., Rae, D. S., Manderscheid, R., & Silberg, S. Exact and approximate distributions of the chi-square statistic for equiprobability. Communications in Statistics -- Simulation and Computation, 1979, B8, 131-149.

Tong, Y. L. An adaptive solution to ranking and selection problems. The Annals of Statistics, 1978, 6, 658-672.

van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1980, 33, 104-108.

van der Linden, W. Estimating the parameters of Emrick's mastery testing model. Applied Psychological Measurement, 1981, 5, 517-530.

Wainer, H., Morgan, A., & Gustafson, J. A review of estimation procedures for the Rasch model with an eye toward longish tests. Journal of Educational Statistics, 1980, 5, 35-64.

Wainer, H., & Wright, B. D. Robust estimation of ability in the Rasch model. Psychometrika, 1980, 45, 373-391.

Wilcox, R. R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. Journal of Educational Statistics, 1977, 2, 289-307. (c)

Wilcox, R. R. Determining the length of a criterion-referenced test. Applied Psychological Measurement, 1980, 4, 425-446. (a)

Wilcox, R. R. Some results and comments on using latent structure models to measure achievement. Educational and Psychological Measurement, 1980, 40, 645-658. (b)

Wilcox, R. R. An approach to measuring the achievement or proficiency of an examinee. Applied Psychological Measurement, 1980, 4, 241-251. (c)

Wilcox, R. R. Solving measurement problems with an answer-until-correct scoring procedure. Applied Psychological Measurement, 1981, 5, 399-414. (a)

Wilcox, R. R. A review of the beta-binomial model and its extensions. Journal of Educational Statistics, 1981, 6, 3-32. (d)

Wilcox, R. R. Some empirical and theoretical results on an answer-until-correct scoring procedure. British Journal of Mathematical and Statistical Psychology, 1982, 35, 57-70. (a)

Wilcox, R. R. Some new results on an answer-until-correct scoring procedure. Journal of Educational Measurement, 1982, 19, 67-74. (b)

Wilcox, R. R. Determining the length of a multiple-choice criterion-referenced test when an answer-until-correct scoring procedure is used. Educational and Psychological Measurement, 1982, 42, 789-794. (c)

Wilcox, R. R. A closed sequential procedure for answer-until-correct tests. Journal of Experimental Education, 1982, 50, 219-222. (d)

Wilcox, R. R. Approaches to measuring achievement with an emphasis on latent structure models. Technical Report, Center for the Study of Evaluation, University of California, Los Angeles, 1982.

Wilcox, R. R. Bounds on the k out of n reliability of a test, and an exact test for hierarchically related items. Applied Psychological Measurement, 1982, 6, 327-336. (e)

Wilcox, R. R. On a closed sequential procedure for categorical data, and tests for equiprobable cells. British Journal of Mathematical and Statistical Psychology, 1982, 35, 193-207. (f)

Wilcox, R. R. Using results on k out of n system reliability to study and characterize tests. Educational and Psychological Measurement, 1982, 42, 153-165. (g)

Wilcox, R. R. An approximation of the k out of n reliability of a test, and a scoring procedure for determining which items an examinee knows. Center for the Study of Evaluation, University of California, Los Angeles, 1982. (j)

Wilcox, R. R. How do examinees behave when taking multiple-choice tests. Applied Psychological Measurement, 1983, 7, 239-240.

Wilcox, R. R., & Harris, C. W. On Emrick's "An evaluation model for mastery testing." Journal of Educational Measurement, 1977, 14, 215-218.

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