The Nature and Use of State Mastery Models

George B. Macready and C. Mitchell Dayton
University of Maryland

This paper provides a review of a class of probabilistic models that has been developed for use in the assessment of trait or competency acquisition. Consideration is given to the relative merits and limitations of this class of state models, under which trait acquisition is conceived as being "all-or-none," as compared with those occurring under an alternative conceptual framework, in which trait acquisition is assumed to be gradual. In addition, some of the applications of these state models are presented, including the establishment of mastery classification decisions and the assessment of consistency with respect to items and classification. Finally, some extensions to the class of state models, which may be helpful in increasing the applicability of this class of models, are presented.

An important element of the criterion-referenced approach to testing is the assessment of individuals in terms of absolute standards of attainment for traits or competencies of interest; and within this context, the concept of "mastery" has played a central role. To deal with this classification problem, a variety of strategies have been suggested (see Hambleton & Eignor, 1979). Meskauskas (1976) pointed out that most of these strategies can be grouped into two general classes based on the underlying conceptualization of the trait being assessed. Within the first class, called *continuum models*, trait acquisition is assumed to be gradual and mastery is viewed as an interval on a test score scale. Within the second class, called *state models*, trait acquisition is conceived of as an "all-or-none" process and mastery is viewed as the presence of trait acquisition.

Limitations of Continuum Models

From the perspective of continuum models, mastery of a trait is based on "sufficient" partial acquisition of that trait. Thus, mastery classification involves a judgmentally established point or set of points on the trait continuum. An integral part of procedures used to establish rules for mastery classification within the framework of continuum models, mentioned by both Linn (1978) and Glass (1978), is their fundamentally judgmental nature, resulting in rules for classification that are arbitrary. This arbitrariness presents a problem for effective implementation of the procedures; and as might be expected, consistency of mastery classifications both within and across procedures often is

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 4 Fall 1980 pp. 493–516

not high (see Andrew & Hecht, 1976; Meskauskas & Webster, 1975). Yet, there is no clear way to make choices among the available procedures. Presently, there is only limited methodology available to be used in an attempt to provide an effective means of assessing either the absolute or the relative adequacy of some of the more logically well-founded procedures within the class of continuum models. Even though general recommendations and concerns that may be of some use in both selecting and implementing procedures have been presented (e.g., Hambleton & Eignor, 1979; Jaeger, 1979; Shepard, 1979), the available technology provides a less than adequate means for choosing among various procedures.

An equally serious issue related to continuum models, raised by Glass (1978), deals with the logical incompatability between the conception of trait acquisition on a continuous scale and the concept of mastery defined in terms of some "optimal" dichotomy of that continuum. Such an "optimal" dichotomy, Glass argues, does not exist; and it would appear that the use of the concept of mastery under such circumstances may be highly misleading, especially to the general public.

The seriousness of the problems that arise when the concept of mastery is defined within a continuum model framework of trait acquisition suggests the need for a thorough review of the characteristics of state models to determine if they provide a viable alternative for defining the concept of mastery. The purpose of this paper is to provide a review of state mastery models with respect to their strengths and weaknesses relative to each other as well as to continuum models. In addition, a number of model extensions are presented that may be helpful in increasing the applicability of this class of models.

State Mastery Models

General Characteristics

Within the framework of state models, it is assumed that there is a common set of required skills necessary to respond correctly to all items in a domain of interest. Thus, the presence or absence of those skills define the two mutually exclusive and exhaustive states of mastery with the proportion of nonmasters ($k=1$) and masters ($k=2$) being designated Δ_1 and Δ_2, respectively.

Here, masters are those individuals who have acquired the complete set of common skills required to respond correctly to all items within the domain. Thus, for three items from a single domain, a master's true score response pattern on the items would be 111, where a "1" indicates a correct response to an item. Conversely, nonmasters are those individuals who have not acquired the complete set of common skills required to respond correctly to any item within the domain; thus, their true score response pattern would be 000, where "0" indicates an incorrect response to an item. In most real-life situations, however, the true state of an individual will be latent and the individual's observed item response pattern may show deviations from the response pattern representative of the true latent state. These deviations are assumed to be due to the existence of "intrusion" and "omission" errors that are inconsistencies of observed item responses from the expected latent response for nonmasters and masters, respectively.

The likelihood of occurrence of errors for the j^{th} item defines the probabilities of a false positive response, α_j, by a nonmaster and a false negative response, β_j, by a master. Similarly, $1-\alpha_j$ and $1-\beta_j$, respectively, represent the conditional probability for avoiding the above types of errors for some item j.

Previous papers that have considered state mastery models have either implied or explicitly defined the above errors in terms of guessing and forgetting as well as other psychological-environmental factors, such as cheating. This perspective is extremely limiting, since in addition to the long list of

psychological-environmental factors that might result in errors, there is at least one additional set of factors to which errors may be attributed. In many cases, this latter set of factors may be a far more important source of errors. One potentially important factor within the realm of cognitive skills deals with the relation of the sets of necessary and sufficient skills required for an individual to generate positive responses to specific items within the domain of interest.

For some specified item, j, it may be that in addition to the set of "common" required skills for all items, a set of additional required skills may be necessary to generate an appropriate positive response to the item. This would result in a high omission error rate, β_j, for such an item. Alternatively, some items may tend frequently to be initially learned in a rote fashion so that it is possible for a positive response to be generated by an individual who has not acquired the common set of required skills necessary for all items. This would result in a high intrusion error rate, α_j, for such an item. The potential for the above kinds of errors might suggest that a review of the error probabilities related to items would be useful for assessing the adequacy of domain specification rules. This is because generation of domains that are highly homogeneous may be desirable, as has been suggested by Macready and Merwin (1973) and Harris (1974).

Since there is no way to tie the above mentioned error probabilities to any specific closed set of factors, it is here contended that α_j and β_j should be perceived only as conditional probabilities related to the likelihood of item responses conditional on latent mastery state. Under this less restrictive conception of the error probabilities, it also becomes possible to apply what here are called state mastery models to any trait, cognitive or otherwise, which is assumed to exist at two mutually exclusive and exhaustive latent levels (see Lazarsfeld & Henry, 1968). An example within genetics would be the use of phenotypic information about an organism to assess the latent genotype.

With the assumptions (1) that local independence occurs among responses (i.e., the occurrence or nonoccurrence of intrusion or omission errors are assumed to be independent across items), (2) that the latent trait exists at two mutually exclusive and exhaustive levels, and (3) that items are dichotomously scored, the probability for each of 2^n possible response vectors U_r (representing the possible item response patterns for an n-item test) conditional on the latent mastery state may be designated

$$\begin{cases} P(\underline{U}_r \mid k=1) = \prod_{j=1}^{n} \alpha_j^{X_{rj}} (1-\alpha_j)^{1-X_{rj}} \\ P(\underline{U}_r \mid k=2) = \prod_{j=1}^{n} \beta_j^{1-X_{rj}} (1-\beta_j)^{X_{rj}} \end{cases} \qquad [1]$$

where $X_{rj} = \{0,1\}$ is the score of the j^{th} item found within the r^{th} response pattern. Similarly, the probability of an observed response pattern, U_r, across mastery states may be designated as

$$\begin{aligned} P(\underline{U}_r) &= P(\underline{U}_r \cap k=1) + P(\underline{U}_r \cap k=2) \\ &= P(\underline{U}_r \mid k=1)\Delta_1 + P(\underline{U}_r \mid k=2)\Delta_2 \\ &= \left[\prod_{j=1}^{n} \alpha_j^{X_{rj}} (1-\alpha_j)^{1-X_{rj}}\right] \Delta_1 + \left[\prod_{j=1}^{n} \beta_j^{1-X_{rj}} (1-\beta_j)^{X_{rj}}\right] \Delta_2. \end{aligned} \qquad [2]$$

Thus, for three items, the probability of the occurrence of the response vector $U_r = (101)'$ is

$$P[\underset{\sim}{U}_r = (101)'] = [\alpha_1(1-\alpha_2)\ \alpha_3]\Delta_1 + [(1-\beta_1)\beta_2\ (1-\beta_3)]\Delta_2. \quad [3]$$

Equation 2 delineates the most general form of state mastery models (as defined above) that has been presented to date (see Besel, 1973, 1975; Macready & Dayton, 1977, 1980; Roudabush, 1974; van der Linden, 1978). This general model, here called the $\alpha_j\beta_j$ model, subsumes as restricted mathematical forms all other two-state mastery models that have previously been generated. The constraints defining restricted forms of this model deal with restrictions placed on α_j and β_j for $j=(1,\ldots,n)$ such that subsets of these parameters are fixed and/or equated to each other. In addition, these subsumed models differ in terms of item domains of interest, with some models restricting themselves to replications of a single item or equivalent items.

Constrained Forms of Two-State Models

The first subcategory of constrained models are those that equate conditional probabilities (i.e., equating intrusion and/or omission errors). One such model is the $\alpha\beta$ model, which has been considered by Davis, Hickman, and Novick (1973), Emrick (1971), Emrick and Adams (1969), and Macready and Dayton (1977). This model equates all intrusion errors, as well as all omission errors. Thus, the following constraints are imposed on Equation 2: $\alpha_j=\alpha$ and $\beta_j=\beta$ for $j=(1,\ldots,n)$. This model may be particularly viable when test items of interest are replications of a single item or are sets of equivalent items. A second model within this class is the $\alpha_j=\beta_j$ model, presented by Wilcox (1977b) as a means of assessing pairs of items. This model equates omission errors with intrusion errors for each item such that $\alpha_j=\beta_j$ for $j=(1,\ldots,n)$; Wilcox considers only the case of $n=2$. This model would be reasonable if the assumption that false positive and false negative errors are equally likely for a given item. A similar model within this class that was considered by Wilcox (1977a), a constrained form of the $\alpha_j=\beta_j$ model, is the $\alpha=\beta$ model. Here, all errors across items are equated such that $\alpha_j=\alpha_{jj'}=\beta_j=\beta_{j'}$, for $j,j'=(1,\ldots,n)$. However, Wilcox has presented this model as a means for assessing the stability of a single item on two occasions. Thus, this model from an "item" perspective provides a logically compatible counterpart to the $\alpha_j=\beta_j$ model, if error probabilities do not change over occasions as the $\alpha=\beta$ model assumes. Note that for the two-item case, the expected frequencies for the response patterns U_r generated under this model (using maximum likelihood parameter estimates) are identical to those obtained when a McNemar (1947) chi-square test of change for correlated proportions is applied.

The second subcategory of constrained models are those that fix one or more of the conditional errors. The least restrictive of these models are those that set one class of errors equal to zero. They are the α_j model and the β_j model presented by Wilcox (1977b), with the latter model also considered by Harris and Pearlman (1978). The α_j model assumes that omission errors do not occur and sets $\beta_j=0$ for $j=(1,\ldots,n)$, whereas the β_j model correspondingly assumes that intrusion errors do not occur and sets $\alpha_j=0$ for $j=(1,\ldots,n)$. This pair of mathematically equivalent models may be considered as a single model with different interpretations given to the defining parameters. Both of these models have been presented within the context of item pairs; however, as with all of the models presented in this paper, simultaneous consideration of more than two items (or two occasions) is possible. These models provide a potentially viable perspective when individuals in the appropriate latent mastery class always respond in a deterministic fashion.

Harris and Pearlman (1978) have attempted to establish a rationale for the constraints incorporated within the β_j model by arguing against any potential for guessing with free-response-type items. This argument is at best questionable, based on the earlier discussion of conditional probabili-

ties; and for this reason, it is recommended that the fit obtained under this model be statistically compared with that obtained under the unconstrained $\alpha_j\beta_j$ model before it is incorporated for use. Similar arguments might also be developed for all of the constrained models presented in this paper. This will, however, require the simultaneous consideration of at least four items (or occasions), since only then can "fit" of the $\alpha_j\beta_j$ model be statistically assessed.

Two additional mathematically equivalent models (which are also constrained forms of the α_j and β_j models) are the α model and the β model, both of which were presented by Wilcox (1977a), with the latter model also discussed by Harris and Pearlman (1978) and Knapp (1977). Here, in addition to one class of errors being fixed at zero, the errors within the other class are equated. Thus, the α model contains the following constraints: $\alpha_j=\alpha_{j'}$ and $\beta_j=0$ for $j,j'=(1,\ldots,n)$. The β model incorporates similar restrictions: $\alpha_j=0$ and $\beta_j=\beta_{j'}$ for $j,j'=(1,\ldots,n)$. Both of these models were presented for the case in which a single free-response-type item is presented on two occasions. Within that context these models would appear to be logically compatible with their mathematically subsumed counterparts, the α_j and β_j models, if error probabilities do not change over occasions as the models assume. In an attempt to establish a rationale for the assumption $\alpha_j=0$ for the β model, Harris and Pearlman (1978) used an argument similar to the one they advanced regarding the β_j model; thus, the authors' earlier specified concerns and recommendations are again in order.

A final model within this subcategory of constrained models, called the $\alpha_1\beta_1$ model, was considered by Roudabush (1974). This model allows the conditional errors related to a first (criterion-referenced test) item to be free, while fixing those related to a second (criterion) item at zero (e.g., $\alpha_2=\beta_2=0$ are the constraints incorporated within this model). Note that an expanded form of this model would incorporate $n>2$ items of which only the last item would have its conditional errors fixed at zero. The rationale for the constraints imposed within this model is that the last item provides perfect differentiation of masters from nonmasters. This contention may be reasonable in some cases of interest, especially when the trait of interest is noncognitive and the last item is an a posteriori criterion (i.e., job acquisition, organ malignancy, or egg fertility). However, for many traits such criteria may be difficult or impossible to obtain.

An interesting alternative conceptualization of some of the mathematical functions that define mastery models described above (namely, α; β; β_j; and $\alpha=\beta$ models) has been considered by Wilcox 1979a, 1979b). He allowed these models to characterize a single examinee (as opposed to a population of examinees) in terms of a domain of items. Within this context, these models designate the likelihood of the occurrence of omission and intrusion errors for randomly sampled items from the item domain, while Δ_2 designates the proportion of items in the domain that have been acquired by the individual of interest. The conceptual framework that Wilcox considered with the above four models may also be applied to the mathematical equations that define all the aforementioned mastery models. Thus, this conceptual framework allows the mathematical models to be used in situations where examinees are not located within one of two mutually exclusive and exhaustive latent classes. However, the limitations attached to continuum models will present a problem under this conceptual framework.

Multi-state Mastery and Additional Latent State Models

In addition to the mastery models that are constrained forms of the $\alpha_j\beta_j$ model, two other models have been presented in the literature (see Bergan, Cancelli, & Luiten, 1980; Knapp, 1977). Both of these models may be conceptualized as constrained forms of a "multi-state model" in which there are more than two mutually exclusive and exhaustive latent states defining mastery. The unrestricted

multi-state model, which is presented in a broader context by Lazarsfeld and Henry (1968), is defined as follows:

$$P(\underset{r}{U}) = \left[\prod_{j=1}^{n} \alpha_j^{X_{rj}}(1-\alpha_j)^{1-X_{rj}}\right]\Delta_1 + \left[\prod_{j=1}^{n} \beta_j^{1-X_{rj}}(1-\beta_j)^{X_{rj}}\right]\Delta_2 + \sum_{k=3}^{K}\left[\prod_{j=1}^{n} \gamma_{jk}^{X_{rj}}(1-\gamma_{jk})^{1-X_{rj}}\right]\Delta_k \qquad [4]$$

where

$\sum_{k=1}^{k} \Delta_k = 1.0$,

Δ_k is the proportion of individuals in the k^{th} latent state, and

γ_{jk} is the probability of a positive response to the j^{th} item for individuals in the k^{th} latent state.

Note that Equation 4 differs from Equation 2 only in its incorporation of a third term, which relates to the probability of attainment of the observed response pattern U_r by individuals within the latent classes 3 through K. The assumptions underlying this general model, here called the $\alpha_j\beta_j\gamma_{jk}$ model, are comparable to those for the $\alpha_j\beta_j$ model.

The first of the multi-state mastery models was presented by Knapp (1977) for the assessment of replication data for a single selection type item here called the $\beta\gamma_{jk}$ model. This model assumes that there are five levels of mastery, four of which are special types of nonmastery. These four nonmastery states are defined as follows:

k=1 are those nonmasters who guess on neither of two replications of an item;
k=3 are those nonmasters who guess on the first, but not on the second, replication of an item;
k=4 are those nonmasters who guess on the second, but not on the first, replication of an item; and
k=5 are those nonmasters who guess on both of two replications of an item.

Probabilities of intrusion errors are fixed at values that correspond to the likelihood of false positive responses to an item presentation by the various types of nonmasters 'guessing' correctly, while the probability of omission errors are equated across replications. Thus, this model contains the following constraints:

$\beta_1 = \beta_2$;
$\alpha_1 = 0$; $\alpha_2 = 0$;
$\gamma_{31} = 1/d$; $\gamma_{32} = 0$;
$\gamma_{41} = 0$; $\gamma_{42} = 1/d$;
$\gamma_{51} = 1/d$; $\gamma_{52} = 1/d$; and
$\Delta_1 = (1-g)^2(1-\Delta_2)$; $\Delta_3 = \Delta_4 = g(1-g)(1-\Delta_2)$; $\Delta_5 = g^2(1-\Delta_2)$

where

d is the number of item choices,
$1-\Delta_2$ is the total proportion of all types of nonmasters, and
g is the proportion of nonmasters who on any given presentation of the item decide to guess.

Notice that for a completion type item, where d may be assumed to be infinite, this model is equivalent to the β model presented earlier. In addition, there is a similar rationale under this model for fixing conditional probabilities related to nonmasters (based on guessing) as was presented for both the β_j and β models. Thus, similar concerns and recommendations are suggested regarding its use.

The second multi-state mastery model, called the γ_{j3} model, which is also a constrained form of the $\alpha_j\beta_j\gamma_{jk}$ model, was presented recently by Bergan, Cancelli, & Luiten (1980). This model (which is a special case of the "quasi-independence" model presented by Goodman, 1975) incorporates three latent states: nonmasters, (k=1); masters, (k=2); and "intrinsically unscalables," (k=3). The model assumes that individuals within the mastery and nonmastery states respond in a deterministic fashion, with masters always responding correctly and nonmasters always responding incorrectly to all items. Thus, it is only the third class of individuals who are able to attain all of the 2^n possible item response patterns. The restrictions on the general $\alpha_j\beta_j\gamma_{jk}$ model that defines this model are $\alpha_j = \beta_j = 0$ for $j = (1,\ldots,n)$.

In a sense, this model may be conceived as an "additional latent state model" that falls somewhere between two-state mastery models and continuum mastery models. This is because continuum models assume an infinite number of states of trait acquisition; thus, Equation 4 provides an approximate representation of such a model as K becomes large. This may suggest that if better fit to data is obtained by increasing the number of latent states, consideration of continuum models may be in order; however, lack of improved fit does not necessarily negate the viability of continuum models. In any case, comparisons of the above type are recommended as standard procedure for assessing the adequacy of state mastery models. An appropriate model to compare with a given state mastery model may be a restricted form of the $\alpha_j\beta_j\gamma_{jk}$ model with equivalent restrictions to the state mastery model being assessed but with one additional unconstrained latent class. This may be preferable to using the γ_{j3} model because only the $\alpha_j\beta_j\gamma_{jk}$ model subsumes all state mastery models presented, and thus an assessment of relative fit is possible. In addition, Dayton and Macready (1980) have presented some conceptual problems with quasi-independence models of which the γ_{j3} model is a special case.

Independence Models

Models that are constrained forms of the general models $\alpha_j\beta_j$ and $\alpha_j\beta_j\gamma_{jk}$ but that themselves are not mastery models form the so-called class of independence models. These models assume that there is only one class of individuals, in other words, $\Delta_1 = 1.0$. Since there is only one class of individuals within these models, the assumption of local independence among items becomes an assumption of independence among items. The first model within this class is the Iα_j model, which has as its only constraint that $\Delta_1 = 1.0$. This model is viable in those cases in which there is only one level of the (latent) trait being assessed. The second model within this class is the Iα model, which is a constrained form of the Iα_j model with equal probabilities of positive responses for all items and is defined by the following constraints: $\Delta_1 = 1.0$ and $\alpha_j = \alpha$ for j=$(1,\ldots,n)$. Notice that one or both of these independence models is subsumed as a mathematically constrained form of every mastery model considered. These independence models are important to consider, since they provide simpler means for conceptualizing data. Thus, unless a mastery model can provide statistically better fit than the most complex independence model that it subsumes, a question may be raised as to whether that mastery model (from the perspective of parsimony) provides an appropriate framework for describing the relation between the latent trait and the observed data. This is true even if that mastery model provides adequate absolute fit to the data.

Implementation and Use of State Models

The state models that are described in this paper have been discussed by previous authors with respect to a number of different issues. These issues deal with strategies for selection and implementation of models as well as their applications. Some of the important areas that have been considered include parameter estimation, assessment of model fit, implementation of mastery classification, and assessment of consistency.

Parameter Estimation

The first of these areas is of particular importance, since without adequate procedures for parameter estimation, the models are of little practical use. These models contain $nK + K - w - 1$ independent parameters, where w is the number of independent constraints placed on the general model. Maximum likelihood estimates of these parameters can be obtained (given a sufficient number of items for an identifiable model) by means of the Newton-Raphson iterative procedure presented in Rao (1965) or the iterative proportional fitting procedure described by Goodman (1974). It is frequently desirable to utilize these procedures, since it is not always feasible to obtain explicit formulas in terms of sufficient statistics. In addition, computer programs that incorporate these procedures are available (see Clogg, 1977; Dayton & Macready, 1977). Furthermore, the Newton-Raphson procedure has the added advantage of providing estimates of sampling variances and covariances of the maximum likelihood estimators. However, for the case in which $n = 2$, explicit formulas for maximum likelihood estimates have been generated for many mastery models that are identifiable for that case (see Wilcox, 1977a, 1977b).

A number of other estimation procedures that are not maximum likelihood have been developed (see Blischke, 1964; Emrick, 1971; Houang & Harris, 1980; van der Linden, 1980). Some of these procedures, however, are of limited use because they require one parameter value to be specified or restricted to an assumed amount. This has been pointed out by Wilcox and Harris (1977) and van der Linden (1980), respectively, for the "Emrick" and "Endpoint" methods of estimation. In addition, there are other procedures that have been found to provide parameter estimates that are highly biased and/or have large standard errors when used with data sets where number of respondents is as large as 100 (see Blischke, 1964; van der Linden, 1980).

For the $\alpha\beta$, the β, and the β_j models (which are the only models that have thus far been empirically evaluated), however, there are one or more procedures available, including some that are based on rather simple noniterative techniques that have been shown to provide accurate parameter estimates with minimal bias (see Blischke, 1964; Houang & Harris, 1980; van der Linden, 1980). If the number of items is sufficiently large, this appears to be true even when the number of respondents is as small as 20. For small numbers of respondents, minimally sufficient numbers of items appear to be somewhere between 5 and 10, depending on the model in question. In contrast, Hayek (1978) has found for the $\alpha\beta$ model that when insufficient numbers of items are considered (i.e., five or fewer items) maximum likelihood estimation procedures provide biased estimates of the proportion of masters, even when the number of respondents is in the hundreds.

Assessment of Model Fit

A second important area that must be considered if an adequate mastery model is to be identified deals with assessment of model fit. The effectiveness with which mastery models provide an acceptable representation of individuals' true states of mastery may be based on both absolute and relative

statistical assessments as well as judgmental assessments of the estimated parameters. Assuming that the necessary parameters have been estimated, a chi-square goodness-of-fit test may be used to assess a specified model's absolute fit by utilizing Equation 2 or 4 to obtain expected frequencies for the 2^n response patterns. These expected frequencies may be used, in turn, to generate a likelihood ratio statistic that is asymptotically distributed as chi-square with $2^n - Kn - K + w$ degrees of freedom, where w is the number of independent parameter restrictions for a given model.

Macready and Dayton (1977) have pointed out that a potential source of lack of fit is the untenability of the assumption that mastery is "all-or-none." Thus, they suggest that when fit is not obtained, a possible strategy is to subdivide the item domain into two or more subdomains and assess these subsets for fit. In combination with statistical tests of fit, they also recommend that judgmental assessment of parameter estimates be considered. This is because parameter estimates, along with their estimated standard errors, may reveal logical inadequacies in the values generated and thus may raise a question regarding the adequacy of the model.

Finally, as Dayton and Macready (1976) suggest, it is possible to compare the statistical fit provided by a given mastery model with other models that (1) it subsumes as a constrained form or that (2) subsume it as a constrained form. This may be done by taking the difference between the likelihood ratio statistic for the two models in question. This difference is asymptotically distributed as chi-square with degrees of freedom equal to the difference in degrees of freedom for the two statistics on which it is based. This "difference" chi-square statistic may be used to assess whether the more constrained model provides poorer fit than the subsuming model to which it is compared. Thus, it is possible to make comparisons among a number of different mastery models. Such comparisons may be useful in determining what, if any, restrictions might be imposed on general unconstrained models. In addition, it is also possible to use the above difference chi-square statistic to compare fit provided by mastery models relative to subsuming additional latent state models and identity models, as was suggested earlier in this paper.

Implementation of Mastery Classification

A major potential application of state mastery models is providing a method of mastery classification that is nonjudgmental in nature. One such decision rule was generated by Emrick and Adams (1969) and has more recently been presented by Emrick (1971) and Davis, Hickman, and Novick (1973). This decision rule, which is appropriate for use with the $\alpha\beta$ model and all models it subsumes, results in the location of a cutoff score that minimizes expected loss due to misclassification. The cutoff score, which defines the lower bound of total test scores for individuals classified as masters, is defined as follows:

$$\pi_{01} = \frac{\left[\ln\frac{\beta}{1-\alpha} + \frac{1}{n}\ln\frac{\lambda\Delta_2}{\Delta_1}\right]n}{\ln\frac{\alpha\beta}{(1-\alpha)(1-\beta)}} \qquad [5]$$

where λ is an established loss ratio of false negative loss over false positive loss.

Macready and Dayton (1977) and Bergan et al. (1980) have developed similar optimizing decision rules, which in the generalized form presented here may be used with the $\alpha_j\beta_j$ model, the $\alpha_j\beta_j\gamma_{jk}$ model, or any of the models they subsume to minimize expected loss due to misclassification. However, this procedure implements mastery classification by separate assessment of each observed response

pattern rather than by assessment of total scores. This is because under these general models the joint probability between a specified mastery state k and a response pattern U_r is not necessarily constant for all response patterns that result in the same total score. Here, classification of individuals obtaining a special response pattern U_r to some mastery state k' is based on that level of k' that minimizes the following quantity:

$$\sum_{k=1}^{K} l_{kk'} P(U_r \cap k) = \text{minimum}, \quad \text{for } k' = (1,\ldots,K) \qquad [6]$$

where

$l_{kk'}$ is the loss associated with assigning an individual who is a member of mastery state k to mastery state k',

$l_{kk} = 0$, and

$P(U_r \cap k)$ is the joint probability of the occurrence of response U_r with mastery state k.

Another issue that involves mastery classification, which has been addressed with respect to state models, deals with the minimally sufficient number of items required for mastery decision rules to attain an acceptable expected level of correct classification. In order to obtain this information, it is necessary to designate the maximum acceptable proportion of misclassified individuals, the loss ratio λ to be used in making classification decisions, as well as the identification of the assumed underlying mastery model. It is then possible to apply procedures presented by Macready and Dayton (1977) for estimating the number of items required under the $\alpha\beta$ and $\alpha_j\beta_j$ models or any models they subsume. In addition, they present tabled values generated for use under the $\alpha\beta$ model.

Assessment of Consistency

A final area of analytic development that relates to state models deals with the assessment of item reliability for dichotomous platonic (i.e., all-or-none) true scores. Early work in the area of item reliability for platonic true scores (see Klein & Cleary, 1967, 1969; Levy, 1969) dealt with the tenability of classical test theory assumptions for platonic true scores, which present a dilemma for obtaining acceptable item reliabilities. However, later work by Werts, Linn, and Jöreskog (1973) has shown that by incorporating a congeneric model for use with platonic true scores and dichotomously scored items, it is possible to generate an equation for estimating item reliabilities while maintaining the tenability of the usual underlying assumptions.

The function defining item reliability which was generated by Werts et al. (1973) for some given item j is

$$r_j = \frac{(1-\alpha_j-\beta_j)^2 \Delta_1 \Delta_2}{[(\alpha_j \Delta_1)+(1-\beta_j)\Delta_2][(1-\alpha_j)\Delta_1+\beta_j \Delta_2]} . \qquad [7]$$

Note that the definition for reliability given in Equation 7 is compatible with the $\alpha_j\beta_j$ model or any of the state mastery models it subsumes.

In addition to the work of Werts et al. (1973) based on a congeneric model, a number of indices related to consistency between manifest and latent state have been developed. Knapp (1977) has presented three indices for designating level of item consistency over two occasions, which may be used

within the framework of his multi-state mastery model as well as (if appropriately extended) within any of the two-state model frameworks for which parameter estimates can be obtained. Although Knapp considered these indices for use with data based on a single item on two occasions, they might also be used with data based on two different items.

One class of the indices that Knapp presents designates the proportion of individuals within a specified level of mastery who obtain a manifest two-item response pattern U_r, which is consistent with the "ideal" response pattern for their latent state. The ideal pattern for a given mastery state is the manifest pattern that would be obtained by all individuals within that state if item classification errors did not occur. The following two equations define extended forms of Knapp's indices for non-masters and masters, respectively, which may be used with the $\alpha_j\beta_j\gamma_{jk}$ model or any model it subsumes, given $n=2$,

$$r_{k\neq 2} = \left[\prod_{j=1}^{2}(1-\alpha_j)\Delta_1 + \sum_{k=3}^{K}\left(\prod_{j=1}^{2}(1-\alpha_{jk})\Delta_k\right)\right]/[1-\Delta_2] \qquad [8]$$

and

$$r_{k=2} = \prod_{j=1}^{2}(1-\beta_j). \qquad [9]$$

Notice that for two-state models, the last term in Equation 8 is equal to zero.

In addition to the index of consistency for each level of mastery, Knapp (1977) has also presented a weighted average index of consistency across levels of mastery. This index may be defined as follows:

$$\bar{r}_k = (1-\Delta_2)r_{k\neq 2} + \Delta_2 r_{k=2}. \qquad [10]$$

Harris and Pearlman (1978) have suggested a slightly modified form of the index defined in Equation 10 that designates consistency between observed and ideal scores for a single item on one occasion. This modified index is obtained by simply restricting "j" in Equations 8 and 9 to a single level. It should be pointed out that although these indices presented by Knapp are useful for the one- or two-item (or occasions) case, the extensions of these indices allowing for more than two items or occasions (obtained by replacing "2" with "n" in Equations 8 and 9) is troublesome, since the magnitude of such indices are negatively related to number of items.

An index similar to those presented by Knapp, which was alluded to by Macready and Dayton (1977), specified the expected proportion of correctly classified individuals. This index is defined as follows:

$$r_c = \sum_{r=1}^{2^n}\left[P(\underline{U}_r \cap k=1)b_r + P(\underline{U}_r \cap k=2)(1-b_r)\right] \qquad [11]$$

where $b_r = \begin{cases} 0 \text{ if individuals with response pattern } U_r \text{ are classified as nonmasters,} \\ 1 \text{ otherwise.} \end{cases}$

This index may be used with any of the two-state models and with minor modification is applicable for all state models. In addition, it has the desirable property of being positively related to number of items (or occasions).

A general comment regarding the indices defined in Equations 8 through 11, which has not been previously set forth, is that they are not necessarily affected by variability in observed scores. This means that the usual prerequisite of variability in observed scores required for using most reliability indices is not necessary. Thus, these indices may be helpful in dealing with the potential problem involved in assessing consistency for criterion-referenced tests due to lack of variability, which was raised by Popham and Husek (1969).

Limitations of State Models

Conceptual Issues

State mastery models, similar to continuum models, have certain attributes that, under some circumstances, present limitations to their applicability. One major limitation for state models is that they provide an unreasonable representation of trait acquisition for traits that are highly heterogeneous in content. For cases in which traits are so defined, it may be preferable to incorporate a conception of trait acquisition on a continuous scale. Heterogeneous trait definitions may occur either because such definitions are consistent with the task at hand (e.g., assessment of general reading ability as a single entity) or because definitions that result in instructionally meaningful homogeneous traits sometimes may not be easily established.

Thus, if it is not possible for an investigator to identify one or more homogeneously defined traits of interest, this class of state models would appear to be conceptually incompatible with trait acquisition. This is because it is less than reasonable to assume that there is a meaningful set of common skills that all items in a heterogeneous domain have in common as necessary prerequisites for the occurrence of an "appropriate" positive response. In addition, it is suspected that the statistical fit attained under state models for heterogeneously defined traits will be frequently less than adequate. This problem due to heterogeneously defined traits has been avoided by some model builders, such as Wilcox (1977a) and Knapp (1977), by considering items within the framework of single items presented on two or more occasions. Such a scope may be useful for answering questions about the characteristics of specific items; however, for many educational questions, it will provide a highly limiting framework.

This is not to say, however, that homogeneously defined traits are not desirable. In fact, within the instructional setting, homogeneously defined traits may be far more useful to educators than heterogeneous ones for identifying what it is that students do and do not know. In those cases in which it is both possible and desirable to define traits in a homogeneous fashion, there is evidence from the field of learning that provides some empirical support for both conceptions of trait acquisition (see Gazda & Corsini, 1980). Thus, under such circumstances, it is at least reasonable to empirically investigate the adequacy of state models as well as continuum models.

A second limitation for state models, pointed out by van der Linden (1978), relates to the assumed nature of intrusion and omission errors. More specifically, all nonmasters are assumed to have equal intrusion error rates per item. A similar constraint is also present for masters and their omission error rates. This means that irrespective of other potentially relevant factors such as experience, aptitude, and creativity, the probability of incurring an inappropriate response is the same for all individuals within a given mastery class (e.g., fifth-grade students who know how to solve square root problems have an equal chance of incurring an omission error as their math teacher, who also presumably has

acquired this skill). This may place limitations on the kinds of populations of individuals with which state models, as they are presently defined, should be used. The populations considered should be homogeneous enough that the assumption of equality of error rates for all individuals within a given state of mastery on any specified item is reasonable.

Practical Issues

In addition to the above-mentioned conceptual limitations for state models, there are two practical issues that present limitations for state models. The first of these practical issues deals with the application of these models in situations in which there are large numbers of items (or item replications). For any of the restricted forms of the model that have no unique item parameters (e.g., the $\alpha\beta$ model), large numbers of items present no problem, since the number of parameters to be estimated remains small and total score frequencies provide sufficient information for parameter estimation. However, for the less constrained forms of state models, the simultaneous consideration of large numbers of items (e.g., $n > 15$) is problematic. This is because of the large number of parameters to be estimated and because estimation accuracy becomes increasingly low for samples of individuals that are realistic in size. The limitation is, however, of small practical consequence, since for most uses of state models, small numbers of items prove to be sufficient (see Macready & Dayton, 1977).

The second practical issue related to state models deals with the requirement that the observed item scores are dichotomous in nature, since in many cases it may be desirable to differentiate among more than two categories of responses for a given item. This may happen in a number of commonly occurring situations in testing including

1. Situations in multiple-choice testing where differentiation among response alternatives is desired,
2. Testing situations in which it is desirable to differentiate between incorrect responses and omission of responses, and
3. Testing situations in which it is desirable to give partial credit to items.

Under the present framework for state models, multichotomously scored data must be dichotomized in some logically reasonable fashion if these models are to be applied.

Arbitrariness

A final criticism that has been lodged against all mastery classification procedures (and thus by inclusion must be considered an indictment against classification strategies based on state models) is that they are intrinsically arbitrary in nature. The rationale for this criticism is that all classification strategies contain one or more inherently judgmental steps resulting in classification decisions that are suggested to be arbitrary. This contention has been made by many applied measurement specialists including Glass (1978), Jaeger (1979), Linn (1978), and Hambleton and Eignor (1979). Nonetheless, the criticism is at best inaccurate and, in our view, clearly incorrect, since once a model has been selected, the only step in classification that requires any judgment for a state model procedure is the establishment of a loss ratio. This by itself is not enough to reasonably call the decision-making procedure arbitrary, since all decision-making implicitly or explicitly must incorporate a consideration of relative losses. However, explicit judgment may be eliminated from the classification process simply by using "unweighted" misclassification as a criterion for decision-making while "ignoring" relative loss (i.e., by setting the loss ratio at 1.0).

This improper inclusion of state model procedures within the class of appropriately criticized procedures may in part be explained by the lack of a clear distinction of these procedures from other "mathematically" based procedures. This is suggested in a comment by Glass (1978) about decision theoretic approaches (within which he has included some state model procedures) that investigators who use these procedures "eschew questions of how any particular 'criterion score' is justified or how it is selected. Rather, they proceed from the point at which someone . . . has determined a criterion" (p. 251). This criticism is clearly incorrect for state model procedures.

Extensions of State Models

In an attempt to deal with some of the limitations of previously developed classes of state mastery models, in this section a number of model extensions are proposed, in some sense addressing the criticisms in question. One criticism that is easily addressed is the requirement that observed variables are dichotomously scored. This may be dealt with by simply redefining the most general form of the model specified by Equation 4 in a way that allows for each item j to have $S_j \geq 2$ observable outcomes

$$P(\underset{r}{U}^*) = \sum_{k=1}^{K} \left[\prod_{j=1}^{n} \tau_{X_{rj}|k} \right] \Delta_k \qquad [12]$$

where

U_r^* is the r^{th} response pattern across the multichotomously scored items;

X_{rj} is the score on the j^{th} item found within the r^{th} multichotomous response pattern; and

$\tau_{x_{rj}|k}$ is the probability of the score X_{rj} conditional on the k^{th} level of mastery.

This model, which has been considered in a broader context by Lazarsfeld and Henry (1968), has assumptions comparable to the previous class of models, with the exception that the items are not constrained to be dichotomous. It also subsumes, as constrained forms, all previously discussed state models. In addition, for multichotomous data, the incorporation of appropriate restrictions on Equation 12 results in constrained models that are similar to those found in models considered earlier in this paper. Notice that K may be set at one, two, or some larger integer depending on whether an independence, two-state, or multi-state/additional latent state model is desired.

One means of attempting to deal with the restriction that error rates for observed item responses are constant across individuals within a specified mastery state is to consider a more general model that incorporates "covariate" information. Such information may relate to one or more discrete variables with underlying scales that are nominal or higher. However, these "covariates" should be related to the level of error rates for observed item responses attained by either masters and/or nonmasters. Because covariate information will be treated as nominal data, it is possible to combine one or more multichotomously scored covariates to form a single covariate with C levels such that any level c of that variable corresponds to a unique combination of levels of the original variables. Given such a covariate, it is possible to define a general model (here called the $\alpha_{jc}\beta_{jc}\delta_{kc}$ model), which is specified as follows:

$$P(\underset{r}{U}^{**}) = \sum_{c=1}^{C} \left\{ \delta_{c-1c} \left[\prod_{j=1}^{n} (\alpha_{jc})^{X_{rj}} (1-\alpha_{jc})^{1-X_{rj}} \right] \Delta_{1c} + \delta_{c-2c} \left[\prod_{j=1}^{n} (\beta_{jc})^{1-X_{rj}} (1-\beta_{jc})^{X_{rj}} \right] \Delta_{2c} \right\} \qquad [13]$$

where

U_r** is the r^{th} response across the n dichotomous items on the traits of interest followed by observed covariate score, c',

$\delta_{c'1c}$ is the probability of observed response c' on the covariate, which is contained within U_r**, conditional on nonmastery (k=1) and true level c on the covariate,

$\delta_{c'2c}$ is the probability of observed response c' on the covariate, which is contained within U_r**, conditional on mastery (k=2) and true level c on the covariate,

α_{jc} is the probability of a false positive response to item j conditional on nonmastery and true level c on the covariate,

β_{jc} is the probability of a false negative response to item j conditional on mastery and true level c on the covariate,

Δ_{1c} is the proportion of individuals who are nonmasters and have a true level on the covariate of c, and

Δ_{2c} is the proportion of individuals who are masters and have a true level on the covariate of c.

The $\alpha_{jc}\beta_{jc}\delta_{kc}$ model subsumes as restricted forms all previously discussed two-state models for dichotomous items. Notice that by setting C=1 and $\delta_{c'1c} = \delta_{c'2c} = 1$ Equation 13 equals Equation 2. As might be expected, this new model is based on comparable assumptions to those for two-state models, plus similar assumptions for the covariate except that it is not constrained to be dichotomous. The additional parameters that are incorporated within Equation 13 permit differentiation among individuals at specified levels of mastery. More specifically, the framework of the present model allows for differential error rates (per item, j) for individuals within a given state of mastery who are at a specified level of the covariate (i.e., α_{jc} and β_{jc}). It also incorporates parameters that designate probabilities for errors in specification of levels on the covariate (i.e., $\delta_{c'kc}$ when $c' \neq c$), which are conditional on both individuals' true level c on the covariate and their mastery state k. In addition, differential relative proportions of masters to nonmasters for each level of the covariate are allowed. This is accomplished through the use of separate nonmastery and mastery classes for each level c of the covariate (i.e., Δ_{1c} and Δ_{2c}).

Just as with other classes of state mastery models, there are a variety of constrained forms of the model defined by Equation 13 that might profitably be considered. One subcategory of constrained models related to this general model places no restrictions on latent state proportions. Two constrained models contained within this subcategory that are of particular interest are the $\alpha_{jc}\beta_{jc}$ model and the $\alpha_c\beta_c$ model. Both of these models are similar, not only in their incorporation of covariate information, but also in their assumption that response specification of level on the covariate is an error-free process. This assumption for many covariates may at least be closely approximated. This assumption accounts for all the parameter constraints that define the $\alpha_{jc}\beta_{jc}$ model, namely, that $\delta_{c'1c} = \delta_{c'2c} = 1$ if $c'=c$, or 0 otherwise. However, the $\alpha_c\beta_c$ model is a further constrained form of the $\alpha_{jc}\beta_{jc}$ model and requires that intrusion errors and omission errors for a specified level of the covariate do not differ across items. Thus, the $\alpha_c\beta_c$ model requires in addition to the above specified restrictions on $\delta_{c'kc}$, the following parameter restrictions: $\alpha_{jc} = \alpha_{j'c}$ and $\beta_{jc} = \beta_{j'c}$ for a $j, j' = (1,\ldots,n)$. It is interesting to note that the maximum likelihood parameter estimates obtained under these two models are the same, respectively, as those obtained under the correspondingly constrained two-state models with no covariate when parameter estimates are obtained separately for subgroups of individuals falling at each level of the covariate.

A second subcategory of models (here called the c=1 models), which are constrained forms of the $\alpha_{jc}\beta_{jc}\delta_{kc}$ model, are those that restrict the number of latent mastery and nonmastery classes to one each (i.e., there are no unique parameters for levels of mastery at each level of the covariate). These

models provide a means for assessing whether covariate level is related to the likelihood of occurrence of omission and intrusion errors (i.e., whether a multi-state covariate model provides better fit to the data than a two-state covariate model that ignores levels of the covariate). This may be accomplished by assessing the difference in likelihood ratio chi-square statistics of models from this subcategory with correspondingly restricted models from the general class of covariate models without the above restrictions on the number of latent states. The least restricted model from this subcategory is the $\alpha_{jc}\beta_{jc}\delta_{kc}$: c=1 model. The constraints defining this model are simply that the parameters of nonmasters at each level of the covariate are the same and that the parameters of masters at each level of the covariate are equal. Two other more constrained models from this subcategory are the $\alpha_{jc}\beta_{jc}$: c=1 model and the $\alpha_c\beta_c$: c=1 model. These models, in addition to the above restriction on the number of latent states, also have constraints that are equivalent to those found in the $\alpha_{jc}\beta_{jc}$ and the $\alpha_c\beta_c$ models, respectively.

Other models within this class are in the subcategory of independence models for covariate data. These models have only one state (i.e., K=1); and the most general form within this category will be designated $I\alpha_{jc}$, which will occur if the only restriction imposed on the general model is the restriction to a single state. An additional constraint may be imposed by equating all error probabilities (this additional restriction defines the $I\alpha_c$ model). Thus, any of the models considered within other subcategories of this general class of covariate models will subsume at least one of these two models. This allows for comparable statistical comparisons between state mastery models and independence models, as were suggested for the models with no covariate response data.

It is possible to address simultaneously the above-mentioned limitations dealing with constraints on both the number of item responses and the error rates among individuals within latent states (as well as allowing for "additional latent states" beyond those defining mastery). This may be accomplished by redefining the number of items as "$n-1$," and letting the last element "n" be the covariate. With the incorporation of these modifications, Equation 12 provides a definition for such a model. This new model is more general than any of the previously presented models and in fact subsumes as constrained forms all other models mentioned in this paper. For this overall subsuming class, depending on whether K for a given model equals one, two, or $K>2$, that model will fall within the subcategory of independence, two-state, or multi-state/additional latent state models, respectively.

A Framework for Classifying State Models

There are three model attributes defining all classes of latent state models which have either been discussed or implied in this paper. These attributes are

1. *Level of item response* (for which there are two categories: dichotomous and multichotomous);
2. *Model type* (for which there are four categories: identity models, two-state mastery models, multi-state mastery models, and "additional latent state" models); and
3. *Presence of covariate* (for which there are two levels: no covariate and covariate).

Table 1 provides a model classification matrix designating all possible combinations of levels of these attributes. In the upper left corner of each cell, the general model representing that cell is designated; in the lower portion of the cell, specific subsumed models that were mentioned in the paper are designated (or for the $\tau_{x_r|k}$ models, differentiating attributes are specified). Notice that for cells in any column, a given model within a specified cell subsumes all corresponding models that are in any cell above it. It might also be pointed out that for a number of cells, there are no constrained models

Table 1
Classification of Latent State Models

Model Type	No Covariate		Covariate	
	Dichotomous Items	Multichotomous Items	Dichotomous Items	Multichotomous Items
Independence Model	$\frac{I\alpha_j}{I\alpha}$	$\frac{T_{X_{rj}\|k}}{}$ where K=1 n=no. of items $S_j > 2$, for at least one j	$\frac{I\alpha_{jc}}{I\alpha_c}$	$\frac{T_{X_{rj}\|k}}{}$ where K=1 n-1=no. of items $S_j > 2$, for at least one j
Two-State Mastery Model	$\frac{\alpha_j\beta_j}{\alpha\beta\ ;\alpha_j=\beta_j;\ \alpha=\beta;\alpha_1\beta_1;\ \alpha_j \text{ (or) } \beta_j;\ \alpha \text{ (or) } \beta}$	$\frac{T_{X_{rj}\|k}}{}$ where K=2 n=no. of items $S_j > 2$, for at least one j	$\frac{\alpha_{jc}\beta_{jc}\delta_{kc}:c=1}{\alpha_{jc}\beta_{jc}:c=1;\ \alpha_c\beta_c:c=1}$	$\frac{T_{X_{rj}\|k}}{}$ where K=2 n-1=no. of items $S_j > 2$, for at least one j
Multi-State Mastery Model	$\frac{\alpha_j\beta_j\gamma_{jk}}{\beta\gamma_{jk}}$	$\frac{T_{X_{rj}\|k}}{}$ where K > 2 n=no. of items K=no. of mastery states $S_j > 2$, for at least one j	$\frac{\alpha_{jc}\beta_{jc}\delta_{kc}}{\alpha_{jc}\beta_{jc};\ \alpha_c\beta_c}$	$\frac{T_{X_{rj}\|k}}{}$ where K > 2 n-1=no. of items K=no. of mastery states $S_j > 2$, for at least one j
Additional Latent State Model	$\frac{\alpha_j\beta_j\gamma_{jk}}{\gamma_{j3}}$	$\frac{T_{X_{rj}\|k}}{}$ where K > 2 n=no. of items K > no. of mastery states $S_j > 2$, for at least one j	$\frac{T_{X_{rj}\|k}}{}$ where K > 2 n-1=no. of items K > no. of mastery states $S_j=2$, for all j	$\frac{T_{X_{rj}\|k}}{}$ where K > 2 n-1=no. of items K > no. of mastery states $S_j > 2$, for at least one j

specified, since none was formally discussed. However, based on those models that were considered, a number of additional constrained models might be easily generated and would be located in those empty cells. In addition, it should be pointed out that by no means have all "relevant" constrained models for any cell in the table been considered.

Applicability of New State Mastery Models

The extended models presented in this paper are presently feasible for use in the same areas of application as previously developed state models. A major factor contributing to that applicability is the availability of generalized computer programs (e.g., Clogg, 1977) with the capability of obtaining maximum likelihood parameter estimates by iterative procedures for all of the classes of models here discussed. It should be cautioned, however, that for cases in which item sample size is insufficient, models will not be identified, negating the existence of such maximum likelihood estimates. As Goodman (1974) has pointed out, nonnegative degrees of freedom are a prerequisite for identifiability; thus, in general, the greater the complexity of the model being considered, the greater the number of items that will be required for parameter estimation. Note that, in general, the degrees of freedom for the classes of models specified by Equations 12 and 13 are, respectively,

$$\prod_{j=1}^{n} (S_j) - K \left[\sum_{j=1}^{n} (S_j - 1) \right] - K + w \qquad [14]$$

and

$$2^n \cdot C - 2C(n+2) + W + 2, \qquad [15]$$

Using similar strategies to those discussed earlier in this paper, it is possible for an investigator to choose empirically from among a number of possible state models, including those new models presented here. This may be accomplished by assessing both absolute and relative fit of subsuming models both within and across classes of state models. Once an acceptable model is identified (if such occurs), it is then possible to use that model for a variety of purposes including:

1. Establishing mastery classification rules and implementation of decisions based on those rules;
2. Establishment of minimum acceptable item sample size for mastery classification; and
3. Establishment of item reliability indices of consistency that may be used in the assessment of items, domains, and classification strategies.

These applications are all possible through simple extensions of similar procedures established for previously developed state models considered in this paper.

An Example Comparison among State Mastery Models

Macready (1975) presents a study in which the relations among items across domains from a criterion-referenced test were explored. In this example, that portion of the data from two separate item domains was used to explore the fit provided by several different covariate state mastery models. The two domains considered in this example contain items involving integer multiplication of free-response-type items presented within a vertical format. For the domain defining the trait that was as-

sessed with respect to mastery, the item form rules require that all items have a three-digit multiplier and a three-digit multiplicand, and may require one or more "carry" operations for a correct solution. A second item domain that defines the covariate in this example has item form rules that require that all items have a two-digit multiplier and a three- and four-digit multiplicand, and do not require any "carry" operations for a correct solution. Note that the selected covariate might also be expected to be related to the magnitudes of intrusion and omission errors as well as to the proportion of students who are masters. The data considered in this study were based on 5 and 10 dichotomously scored items that were randomly selected from the domains defining the trait and covariate, respectively. These items were administered to 285 fourth-grade students in the Minneapolis Public Schools.

Total scores on the covariate were dichotomized to reduce the number of model parameters that would need to be estimated. This dichotomization was based on a classification strategy that minimized the expected misclassification and was itself based on an $\alpha\beta$ model. In this case, the cutoff score for mastery classification on the covariate was 6 correct items out of 10. As was expected, the mean score on the assessed trait was far lower for individuals classified as nonmasters on the covariate than those classified as masters. In addition, the distribution of trait scores conditional on covariate level, which is presented in Table 2, shows large differences in variability with almost all individuals who were classified as nonmasters on the covariate attaining a trait score of zero.

Table 2
Conditional Frequency Distributions of Observed Scores

Observed Score	Level of Covariate 1	2	Combined Levels
0	57	37	94
1	2	17	19
2	0	29	29
3	1	48	49
4	1	61	62
5	0	32	32
Total	61	224	285
Mean	.15	2.78	2.22

The covariate data were assessed by means of a number of different covariate state models including the $\alpha_c\beta_c$, the $\alpha_{jc}\beta_{jc}$, and the $\alpha_{jc}\beta_{jc}\delta_{kc}$ models. These models were considered under two conditions: (1) that in which the covariate was considered in the model (here called the c=1,C models and (2) that in which no differentiation was made with respect to the covariate (here called the c=1 models), which is a two-state model. The estimated parameters (with the exception of δ_{kc}) that were obtained under each of these models are presented in Table 3. Notice that due to the lack of differentiation with respect to the covariate occuring within the c=1 models, all corresponding parameter estimates for the two levels of the covariate are the same. However, these corresponding estimates differ across levels of the covariate for the c=1,C models. These differences seem to be such that, in general, the intrusion errors are larger at Level 2 of the covariate (where intrusion errors might be expected to

Table 3
Parameter Estimates for Mastery Models With and Without Covariates

	Mastery Model											
	$\alpha_c\beta_c$				$\alpha_{jc}\beta_{jc}$				$\alpha_{jc}\beta_{jc}\delta_{kc}$			
	c=1: Covariate Level		c=1,2: Covariate Level		c=1: Covariate Level		c=1,2: Covariate Level		c=1: Covariate Level		c=1,2: Covariate Level	
Estimated Parameter	1	2	1	2	1	2	1	2	1	2	1	2
$\hat{\Delta}_2\|_c$	.62	.62	.03	.77	.62	.62	.03	.77	.64	.64	.34	.78
$\hat{\alpha}_{1c}$	.03	.03	.01	.07	.00	.00	.00	.01	.00	.00	.00	.15
$\hat{\alpha}_{2c}$	.03	.03	.01	.07	.03	.03	.02	.06	.02	.02	.01	.40
$\hat{\alpha}_{3c}$	.03	.03	.01	.07	.01	.01	.00	.03	.00	.00	.00	.00
$\hat{\alpha}_{4c}$	.03	.03	.01	.07	.05	.05	.02	.10	.04	.04	.02	.35
$\hat{\alpha}_{5c}$	.03	.03	.01	.07	.05	.05	.00	.13	.03	.03	.00	.48
$\hat{\beta}_{1c}$	.30	.30	.32	.30	.38	.38	.50	.38	.40	.40	.44	.31
$\hat{\beta}_{2c}$	.30	.30	.32	.30	.24	.24	.50	.24	.26	.26	.37	.19
$\hat{\beta}_{3c}$	.30	.30	.32	.30	.28	.28	.00	.28	.30	.30	.00	.29
$\hat{\beta}_{4c}$	.30	.30	.32	.30	.27	.27	.50	.26	.29	.29	.62	.10
$\hat{\beta}_{5c}$	.30	.30	.32	.30	.32	.32	.00	.33	.34	.34	.40	.30

be more likely) and omission errors are larger at Level 1 of the covariate (where omission errors might be expected to be more likely). In addition, these c=1,C models have a greater proportion of masters at Level 2 of the covariate (i.e., $\Delta_2|_c = \Delta_{2c} / (\Delta_{1c}+\Delta_{2c})$ is larger for c=2), as would be expected.

In order to identify an acceptable state mastery model for the trait in question, likelihood ratio chi-square tests of fit were considered. Information related to both absolute and relative fit provided by the various models is presented in Table 4. These analyses were based on the U_r** response pattern data described earlier, where scores on the covariate were dichotomized, as well as an alternative form of that data in which four levels of the covariate were considered. (These covariate scores were obtained by establishing four intervals on the covariate total score scale which contained comparable numbers of individuals.)

It is interesting to note that when a conventional level of significance of .05 was used for assessing statistical fit, the conclusions related to relative and absolute fit attained under a specified model were the same for cases in which the number of levels of the covariate considered in the response patterns was either two or four.

The results presented in Table 4 for the c=1 models (under both forms of the response pattern data) suggest that reasonable fit was obtained only under the $\alpha_{jc}\beta_{jc}\delta_{kc}$: c=1 model. This model was also found to provide superior fit to the other c=1 models, yet it provided no worse fit than the corresponding c=1,C model, which is more complex. A comparable assessment of the c=1,C models resulted in reasonable fit under all models. Comparisons of the relative fit of the c=1,C models resulted in the $\alpha_c\beta_c$ model providing no worse fit than either of the more complex $\alpha_{jc}\beta_{jc}$ or $\alpha_{jc}\beta_{jc}\delta_{kc}$

Table 4
Assessment of Absolute and Relative Model Fit

Levels of Covariate Considered		Mastery Model								
In the Mastery Model	In the Response Patterns	(a) $\alpha_c\beta_c$			(b) $\alpha_{jc}\beta_{jc}$			(c) $\alpha_{jc}\beta_{jc}\delta_{kc}$		
		χ^2	df	p[a]	χ^2	df	p	χ^2	df	p
1	2	178.43	60	.000	161.98	52	.000	35.45	50	.940
2	2	42.76	56	.903	19.53	40	.997	17.21	36	.997
1	4	266.77	124	.000	250.32	116	.000	111.46	108	.392
4	4	105.44	112	.657	61.96	80	.932	51.06	56	.661
Difference Between Models	2	135.67	4	.000	119.22	12	.000	18.24	14	.196
	4	161.33	12	.000	188.35	36	.000	60.39	52	.120
		Differences Between Models								
		(a) vs (b)			(a) vs (b)			(b) vs (c)		
		χ^2	df	p	χ^2	df	p	χ^2	df	p
1	2	6.46	8	.036	143.02	10	.000	126.57	2	.000
2	2	23.23	16	.108	25.55	20	.182	2.32	4	.677
1	4	16.46	8	.036	155.32	16	.000	138.86	8	.000
4	4	43.48	32	.085	54.38	56	.537	10.90	24	.990

[a] Probability of error in rejecting null hypothesis for χ^2 statistic.

models. Thus, within the c=1,C models the $\alpha_c\beta_c$ model is seen as being most appropriate for use with the trait in question. Unfortunately, no direct statistical comparison of the relative fit provided by the $\alpha_c\beta_c$ and the $\alpha_{jc}\beta_{jc}\delta_{kc}$: c=1 models is possible, since neither model subsumes the other. Therefore, some other means must be used to select between these two "best fitting" models.

In addition to the above comparisons for fit, the mastery models considered in this example were also compared to corresponding independence models, as well as corresponding additional latent state models (which incorporate one additional unconstrained latent state). In all cases considered, comparisons with independence models resulted in substantially better fit under the mastery models, suggesting that model simplification to a single state model is not appropriate. Comparisons of fit between the mastery models and the additional latent state models did not always result in the same outcome, as can be seen from the results presented in Table 5. However, comparable fit to the additional latent state models was attained by the two best fitting $\alpha_c\beta_c$ and $\alpha_{jc}\beta_{jc}\delta_{kc}$: c=1 models.

Overall, there is considerable evidence that provides support for the contention that both of these best fitting models provide a reasonable framework for the trait in question. Which of these two models is finally selected from a perspective of expected minimum misclassification for mastery state of the present sample of fourth graders may not be of major consequence. This is because only six students or 2% of the total sample would be differentially classified under the two best fitting models

Table 5
Improvement in Fit Provided by the
Incorporation of an Additional Latent Class

Levels of Covariate in Models	Number of Latent Classes Compared	Mastery Model								
		$\alpha_c\beta_c$			$\alpha_{jc}\beta_{jc}$			$\alpha_{jc}\beta_{jc}\delta_{kc}$		
		χ^2	df	p[a]	χ^2	df	p	χ^2	df	p
c=1	2(vs)3	77.403	7	.000	70.874	7	.000	12.879	7	.075
c=1,2	4(vs)5	11.726	7	.110	.030	7	.999	3.003	7	.885

[a] Probability of error in rejecting null hypothesis for χ^2 statistic.

using the covariate response pattern data with C=2. Similarly for the covariate data with C=4, there was only 3.5% differential classification of students obtained under these two models.

References

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement*, 1976, *36*, 45-50.

Bergan, J. R., Cancelli, A. A., & Luiten, J. W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. *Journal of Educational Statistics*, 1980, *5*, 65-81.

Besel, R. *Using group performance to interpret individual responses to criterion-referenced tests*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 1973.

Besel, R. R. *Mixed group validation and the problem of mastery-learning decisions*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April 1975.

Blischke, W. R. Estimating the parameters of mixtures of binomial distributions. *Journal of the American Statistical Association*, 1964, *59*, 510-528.

Clogg, C. C. *Unrestricted and restricted maximum likelihood latent structure analysis: A manual for users* (Working Paper No. 1977-09). Unpublished manuscript, Pennsylvania State University, 1977.

Davis, C. E., Hickman, J., & Novick, M. R. *A primer on decision analysis for individually prescribed instruction* (ACT Technical Bulletin No. 17). Iowa City, IA: The American College Testing Program, 1973.

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

Dayton, C. M., & Macready, G. B. Model3G and Model5: Programs for the analysis of dichotomous, hierarchic structures. *Applied Psychological Measurement*, 1977, *1*, 412.

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.

Emrick, J. A., & Adams, E. N. *An evaluation model for individualized instruction* (Report RC 2674). Yorktown Hts., NY: IBM, Thomas J. Watson Research Center, 1969.

Gazda, G., & Corsini, R. *Theories of learning: A comparative approach*. Itasca, IL: F. E. Peacock Press, 1980.

Glass, G. V. Standards and criteria. *Journal of Educational Measurement*, 1978, *15*, 237-261.

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

Goodman, L. A. A new model for scaling response patterns: An application of the quasi-indepen-

dence concept. *Journal of the American Statistical Association*, 1975, *70*, 755-768.

Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard-setting. In R. M. Jaeger & C. Tittle (Eds.), *Minimum competency achievement testing*. Berkeley, CA: McCutcheon Publishing Co., 1979.

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Harris, C. W., & Pearlman, A. P. An index for a domain of completion or short answer items. *Journal of Educational Statistics*, 1978, *3*, 285-303.

Hayek, L. C. *Properties of maximum likelihood estimators for a class of probabilistic models*. Unpublished doctoral dissertation, University of Maryland, 1978.

Houang, R. T., & Harris, C. W. *Sampling variance of parameter estimates for a domain referenced latent class model*. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.

Jaeger, R. M. Measurement consequences of selected standard-setting models. M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency based measurement*. Washington, DC: National Council on Measurement in Education, 1979.

Klein, D. F., & Cleary, T. A. Platonic true scores and error in psychiatric rating scales. *Psychological Bulletin*, 1967, *68*, 77-80.

Klein, D. F., & Cleary, T. A. Platonic true scores: Further comment. *Psychological Bulletin*, 1969, *71*, 278-280.

Knapp, T. R. The reliability of a dichotomous test item: A 'correlationless' approach. *Journal of Educational Measurement*, 1977, *14*, 237-252.

Lazarsfeld, P. F. & Henry, N. W. *Latent structure analysis*. Boston: Houghton Mifflin, 1968.

Levy, P. Platonic true scores and rating scales: A case of uncorrelated definitions. *Psychological Bulletin*, 1969, *71*, 276-277.

Linn, R. L. Demands, cautions, and suggestions for setting standards. *Journal of Educational Measurement*, 1978, *15*, 301-308.

Macready, G. B. The structure of domain hierarchies found within a domain referenced testing system. *Educational and Psychological Measurement*, 1975, *35*, 583-598.

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

Macready, G. B., & Dayton, C. M. A two-stage conditional estimation procedure for unrestricted latent class models. *Journal of Educational Statistics*, 1980, *5*, 129-156.

Macready, G. B., & Merwin, J. C. Homogeneity within item forms in domain-referenced testing. *Educational and Psychological Measurement*, 1973, *33*, 351-360.

McNemar, Q. Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika*, 1947, *12*, 153-157.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research*, 1976, *46*, 133-158.

Meskauskas, J. A., & Webster, G. W. The American Board of Internal Medicine recertification examination process and results. *Annals of Internal Medicine*, 1975, *82*, 577-581.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1-9.

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

Roudabush, G. E. *Models for a beginning theory of criterion-referenced tests*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1974.

Shepard, L. A. Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency based measurement*. Washington, DC: National Council on Measurement in Education, 1979.

van der Linden, W. J. Forgetting, guessing and mastery: The Macready and Dayton models revisited and compared with a latent trait approach. *Journal of Educational Statistics*, 1978, *3*, 305-318.

van der Linden, W. J. *Estimating the parameters of Emrick's mastery testing model*. Unpublished manuscript, Twente University of Technology, Enschede, The Netherlands, 1980.

Werts, C. E., Linn, R. L. & Jöreskog, K. A congeneric model for Platonic true scores. *Educational and Psychological Measurement*, 1973, *33*, 311-318.

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

Wilcox, R. R. New methods for studying equivalence. In C. W. Harris, A. P. Pearlman, & R. R. Wilcox (Eds.), *Achievement test items: Methods for study* (CSE Monograph No. 6). Los Angeles: University

of California, Center for the Study of Evaluation, 1977. (b)

Wilcox, R. R. Achievement tests and latent structure models. *British Journal of Mathematical and Statistical Psychology*, 1979, *32*, 61-71. (a)

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

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

Authors' Address

George Macready and C. Mitchell Dayton, Department of Measurement and Statistics, College of Education, University of Maryland, College Park, MD 20742.