Binomial Test Models and Item Difficulty

Wim J. van der Linden
Twente University of Technology

In choosing a binomial test model, it is important to know exactly what conditions are imposed on item difficulty. In this paper these conditions are examined for both a deterministic and a stochastic conception of item responses. It appears that they are more restrictive than is generally understood and differ for both conceptions. When the binomial model is applied to a fixed examinee, the deterministic conception imposes no conditions on item difficulty but requires instead that all items have characteristic functions of the Guttman type. In contrast, the stochastic conception allows non-Guttman items but requires that all characteristic functions must intersect at the same point, which implies equal classically defined difficulty. The beta-binomial model assumes identical characteristic functions for both conceptions, and this also implies equal difficulty. Finally, the compound binomial model entails no restrictions on item difficulty.

In educational and psychological testing, binomial models are a class of models increasingly being applied. For example, in the area of criterion-referenced measurement or mastery testing, where tests are usually conceptualized as samples of items randomly drawn from a large pool or domain, binomial models are frequently used for estimating examinees' mastery of a domain and for determining sample size. Despite the agreement among several writers on the usefulness of binomial models, opinions seem to differ on the restrictions on item difficulties implied by the models. Millman (1973, 1974) noted that in applying the binomial models, items may be relatively heterogeneous in difficulty. Wilcox (1976, 1977) adopted the same position for both the binomial model and the beta-binomial model. Huynh (1976a) stated that it is the exchangeability of all domain items that is automatically assumed in the binomial model, implying similarity of item difficulties; he observed that the beta-binomial model is suitable when a separate sample of items is given to each examinee (Huynh, 1976b, 1977). Lord and Novick (1968, chap. 23) as well as Hambleton, Swaminathan, Algina, and Coulson (1978) made the same observation.

In another paper, Huynh (1976c) does not even mention assumptions of the beta-binomial model. The condition of equal item difficulty for applying the beta-binomial model is also mentioned by Mellenbergh, Koppelaar, and van der Linden (1977). No item difficulty restrictions are mentioned by Fhanér (1974). Kriewall (1972) first says that for a given examinee all items are of equal difficulty by assumption (p. 6); but when giving assumptions for applying the binomial model to item sampling

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 3 Summer 1979 pp. 401–411

(pp. 10-12), he does not mention any restriction on item difficulty. Subkoviak (1976) does not impose explicit restrictions on item difficulties but says instead that a constant probability of a correct response across items for a fixed person is a condition for the binomial model.

This paper carefully examines the restrictions on item difficulties that must be met when binomial models are applied to domain-referenced testing. This is done for both a deterministic and a stochastic conception of item responses. In brief, the former supposes that for a given domain an examinee responds successfully over repeated independent trials (replications) with a probability equal to 1 for some items and to 0 for the remaining part of the domain. The stochastic conception is based on the idea that responding to items is a stochastic process. The probabilities associated with successful outcomes of this process may have values between 0 and 1. The distinction between these two conceptions is justified by the finding that both lead to different restrictions regarding item difficulties. First, however, there must be a consideration of the formal assumption of binomial models and of some definitions and aspects of item-sampling theory.

Binomial Models and Item Sampling

Whenever the formal definition of Bernoulli trials applies to a series of experiments or trials with the outcomes "success" and "failure," the binomial model offers the correct probability distribution for the number of successes, X, in a series of n trials. Bernoulli trials are defined as trials that (1) have two possible outcomes, "success" and "failure"; (2) have a probability of success constant for all trials; and (3) are stochastically independent. The first assumption is evident. The second and third assumptions allow the derivation of the binomial model with only the aid of the product and sum rule for probabilities (see, for instance, Hogg & Craig, 1972, p. 87). Denoting the probability of success at any trial by λ, the binomial density can be written as

$$P(X|n,\lambda) = \binom{n}{X}\lambda^X(1-\lambda)^{n-X}. \qquad [1]$$

When trials conform to the first and the third property, but not to the second, they are said to be Poisson trials (Feller, 1968, p. 218). In that case, the compound binomial model offers the correct description of the number of successes, X, in a series of n trials. This probability distribution is given by the following generating function:

$$\prod_{g=1}^{n}(Q_g+P_gt), \qquad [2]$$

where $Q_g = 1 - P_g$ and P_g denotes the probability of success at the g^{th} trial (Lord, 1965; Lord & Novick, 1968, p. 525).

In some applications of the binomial model, it also makes sense to consider a probability distribution of the binomial parameter λ. Representing the probability density of λ by $f(\lambda)$, it follows that the number of successes, X, is distributed according to

$$h(X) \equiv \int_0^1 P(X|n,\lambda)f(\lambda)d\lambda$$

$$= \binom{n}{X}\int_0^1 f(\lambda)\lambda^X(1-\lambda)^{n-X}d\lambda. \qquad [3]$$

Because of its flexible form and mathematical advantages, a choice is often made of the two-parameter beta density,

$$f(\lambda) = \lambda^{v-1}(1-\lambda)^{w-n}/B(v,w-n+1), \qquad [4]$$

with the complete beta function in the denominator, and $v > o$, and $w > n - 1$ (Lord & Novick, 1968, p. 520). The result obtained in this way is known as the beta-binomial model. From Equation 3 it is clear that $h(X)$ may be considered a mixture of independent binomial distributions, each weighted by $f(\lambda)$. Therefore, the beta-binomial model can only apply to situations in which the conditional binomial distributions $P(X|n, \lambda)$ are realized independently for all values of λ.

In test theory, the item-sampling model is well known. In this model, score X_a of person a is considered the score on a test of n items randomly drawn from a population or domain. Sampling may be real or hypothetical, with or without replacement, stratified or simple, and from a finite or infinite domain. When persons are also randomly sampled, matrix sampling is possible. The starting point for matrix sampling is a matrix $\|Y_{ga}\|$ that arises by taking the Cartesian product of the populations of items and persons, where Y_{ga} is a stochastic variable with possible values 1 and 0 representing the responses of person a to item g. Since Y_{ga} is random over replications for a given g and a, it is in fact the classical test theory propensity distribution at item level (Lord & Novick, 1968, pp. 29–30). A matrix sample is now a realization of the stochastic submatrix obtained by drawing randomly and independently a number of rows and columns of $\|Y_{ga}\|$. As a consequence, the same sample of items is administered to each person in a sample of persons. Sampling plans in which each randomly selected person is given a separately drawn sample of items are also possible.

In item sampling theory, it is usual to define

$$\zeta_a \equiv E_g Y_{ga} \qquad [5]$$

as the person parameter of interest. In terms of the matrix $\|Y_{ga}\|$, ζ_a is the expectation for a given person or column across rows and replications. (Note that Equation 5 should, in fact, have contained two expectation signs, one across items and the other across replications. In this paper, however, the notation proposed in Lord and Novick, 1968, pp. 34–35, has been followed and the latter has been omitted.) It can readily be shown that ζ_a may also be interpreted as the expected relative test score across samples of items for person a. Since the items are not necessarily parallel, randomly drawn samples from a domain are to be considered nominally parallel tests and ζ_a is the relative generic true score.

For nominally parallel measurements, it is usual to assume the following analysis of variance decomposition:

$$Y_{ga} = \mu + (\zeta_a - \mu) + (\pi_g - \mu) + (\alpha_{ga} + e_{ga}), \qquad [6]$$

where μ is the generic true score expected across persons,

π_g is the classical item difficulty,

α_{ga} is the person × test interaction, and

e_{ga} is the specific measurement error (Lord & Novick, 1968, p. 176).

This model is of importance because it demonstrates that the generic measurement error

$$\varepsilon_{ga} \equiv Y_{ga} - \zeta_a$$

$$= e_{ga} + \alpha_{ga} + (\pi_g - \mu) \qquad [7]$$

behaves differently from the specific measurement error of the classical test model. For instance, it is known that the generic measurement errors for two randomly drawn persons, *a* and *b*, are not independent across nominally parallel measurements but have a covariance equal to the variance of item difficulties

$$E_a E_b \sigma(\varepsilon_{*a}, \varepsilon_{*b}) = E_g(\pi_g - \mu_\zeta)^2 = \sigma_\pi^2 \qquad [8]$$

(Lord & Novick, 1968, p. 181). In terms of matrix sampling, this means that a matrix sample yields correlated errors when used for estimating the person parameters (Equation 5), unless all samples possess equal difficulty in the classical meaning of the word.

A Deterministic Conception of Item Responses

Consider first the assumptions regarding item difficulty implied by the binomial models for a deterministic conception of item responses, i.e., when it is assumed that a person produces correct answers to some items of the domain with a probability equal to 1 and to the others with a probability equal to 0. The population matrix from item-sampling theory now contains, not the stochastic variables Y_{ga}, but the deterministic values y_{ga}. Sampling from this matrix means sampling from a pool of 0's and 1's and is entirely equivalent to sampling from the well-known vase with red and white marbles. (An explicit reference to this analogy can sometimes be found in the literature on mastery testing, e.g., see Kriewall, 1972). In classical test theory, a true score is defined as the observed score expected across replications. Since the expected value of a constant is the constant itself, the deterministic conception of item responses entails the equality of observed and true item scores. There is no measurement error (in the classical meaning of the word), and the only error possibly involved is estimation error when using a sample of item responses to estimate the proportion of items a person has correct in a given domain.

The idea of item responses as deterministic events, only having a probability of success equal to 1 or 0, is akin to the hypothesis of learning as an all-or-none process. According to this hypothesis, a student is not able to produce any correct response up to a certain point in the learning process; however, having passed this point, the situation has fully changed and he/she will always produce the correct answer. In the parlance appertaining to this hypothesis, knowledge is treated as an all-or-none state: A student who has passed the critical point in the learning process "knows" the item; the others "do not know it."

The deterministic view also underlies the so-called state models for mastery testing (see Besel, 1973; Dayton & Macready, 1976; Emrick, 1971; Emrick & Adams, 1969; Macready & Dayton, 1977; Meskauskas, 1976). For instance, Macready and Dayton (1977), in formulating their models for mastery testing, took the idea that a person is either a master with a true item score vector of 1's or a nonmaster with a true item score vector of 0's. This is clearly a deterministic starting point, since true item scores equal to 1 or 0 imply probabilities of correct responses having the same values of 1 and 0, respectively. According to Macready and Dayton, however, occasional disturbances, such as forgetting and guessing, prevent this true state from showing itself fully; they next adopted parameters to correct these probabilities to make them better fit real test data.

Lord and Novick (1968, p. 236), in their discussion of matrix-sampling theory, indicated that they did not consider the item response variable Y_{ga} as random across replications and replaced it by the constant y_{ga}. Although they stated that this was done to simplify their thinking and that they did not expect to arrive at incorrect conclusions, this amounts to replacing a stochastic view of item re-

sponses by a deterministic view. It is the purpose of this paper, however, to show that this choice is not without consequences and leads to different conditions for applying binomial models to domain-referenced testing or to item-sampling situations.

Latent Trait Conceptualization

Since the concern of this paper is in the implications of binomial models not only for item difficulty defined according to classical test theory (i.e., as the expected item response across replications and persons) but also for the difficulty parameter from latent trait theory, it is worth noting how the deterministic conception entails a special form for the characteristic curves of all items of the domain. To show this, a latent trait point of view has been adopted and the discussion has been restricted to domains for which an underlying continuum can be assumed and the probability of a successful response is a nondecreasing function for each item in the domain. Items of this type are sometimes called monotonic items; it seems safe to assume them here, since nonmonotonic items are only occasionally found in the area of attitude measurement and not in achievement testing.

From a latent trait point of view, a deterministic conception of item responses amounts to the idea that these characteristic curves are degenerated to a Guttman form: Up to a certain point, the probability of a correct response is equal to 0; thereafter, it is equal to 1. Denoting the latent continuum by θ, the Guttman item characteristic curve for item g is defined as

$$P_g(1|\theta) = \begin{cases} 0 & \text{for } \theta < b_g \\ 1 & \text{for } \theta \geq b_g, \end{cases} \qquad [9]$$

where b_g is the point at which the curve shows its jump. In latent trait theory, b_g is interpreted as the difficulty parameter of item g. For the sake of clarity, it has not been shown that Equation 9 is always true. What has been pointed out is that whenever a latent trait point of view is appropriate, the deterministic conception of item responses must take the form of Equation 9. This is necessary in order to analyze whether the use of binomial models entails any restrictions on the latent trait theory item difficulty parameter, b_g; The results of such an analysis would have practical meaning only when the latent trait point of view is indeed appropriate.

Some authors (e.g., Millman, 1974) have advocated that item domains need not necessarily be homogeneous and that binomial models are excellently suited to analyze sampling from heterogeneous domains. Therefore, unless otherwise indicated, it will be assumed that θ from Equation 9 is a vector representing the complete collection of all latent variables underlying the domain. In that case, Equation 9 shows a multidimensional generalization of a Guttman characteristic curve. Multidimensional item characteristic curves are usually called item characteristic functions, and this custom will be adopted to indicate when and when not to consider the multidimensional case.

Domain Sampling for a Fixed Person

An analysis will now be made of the situation of a fixed person, a, with a latent vector θ_a and items randomly drawn with replacement from a finite domain or without replacement from an infinite one. For this person, the number of items for which he/she has a correct response can, in principle, be counted; and a proportion of correct responses may be defined as

$$\tau_a \equiv \sum_{g=1}^{N} y_{ga}/N, \qquad [10]$$

where N denotes the domain size and, in case of an infinite domain, the limit for N to infinity should be added. Note how τ_a results from applying the definition of relative generic true score Equation 5 to deterministic values y_{ga} instead of to random variables Y_{ga}. Although the population matrix $\|Y_{ga}\|$ and the parameter τ_a defined on it are deterministic quantities, sampling creates a chance mechanism generating the probability distribution of the number of 1's in a sample of size n. It will be clear that this distribution is the binomial Equation 1 with λ replaced by τ_a, since the responses to randomly drawn items can be considered outcomes of Bernoulli trials: There are only two possible outcomes, and random sampling guarantees equal probabilities for these outcomes at each trial and stochastic independence between all trials.

From this conclusion it follows that for a deterministic view of item responses, a fixed person, and item sampling of the above type, the binomial model does not impose any restriction on the item difficulties. The Guttman item characteristic functions may display their jump anywhere in the latent space without invalidating the description of the item responses as outcomes of a Bernoulli process. And the individual parameters in the difficulty vector b_g, which represent the item difficulty for each separate dimension of the complete latent space and together indicate the place where this jump occurs, are not restricted in their possible values. For a fixed person and a deterministic conception, the classical definition of item difficulty as the expected item response across replications and persons degenerates to the expected value of the constant y_{ga}, which is equal to y_{ga} itself. In the previous reasoning, no assumptions were made regarding these constants. Therefore, although this definition has a degenerate meaning here, it may be said that applying the binomial model in the present case does not impose any restriction on classical item difficulty either. Applying the compound binomial model is never a requirement. The reason for applying this model to sampling deterministic responses is not variation in item difficulties but stratified, instead of simple, random sampling.

Matrix Sampling

The case of a population of persons and a domain of items with random sampling from both, assuming that sampling takes the form of matrix sampling, will be treated next. The persons and items are drawn independently, and the same test is administered to all persons. An extension of the binomial model Equation 1 to the beta-binomial model in Equations 3 and 4, with λ replaced by τ from Equation 10, seems to be obvious. It was seen earlier, however, that the beta-binomial model only applies to situations in which the conditional distributions $P(X|n, \tau)$ are realized independently for all values of τ; and this requirement is not met when matrix sampling is used. From the equivalence of τ to the relative generic true score ς and Equations 7 and 8, it follows that the distributions

$$\begin{aligned} P(X|n,\tau) &= P(X-n\tau|n,\tau) \\ &= P(\varepsilon|n,\tau) \end{aligned} \qquad [11]$$

are not independent and that the beta-binomial model does not apply. Equation 8 shows that the covariance between any two distributions is equal to the variance of the classical item difficulties and that this covariance only vanishes when all items of the domain have equal difficulty. For this reason, Lord and Novick (1968, p. 524) have observed that the beta-binomial model may be applied to matrix sampling only if all items in the domain are of equal difficulty. It is important, however, to add that this is only a necessary condition for applying the beta-binomial model. A covariance equal to zero by no means implies that the distributions are independent. Note, too, that the combination of equal

item difficulty (in the classical meaning of the word) and Guttman characteristic functions imposes a restriction on the difficulty parameters of the latter as well.

In order for a domain of items with these characteristic functions to yield equal classical difficulties, the following inequality,

$$\theta_a < b_g \leq \theta_g \qquad [12]$$

must hold for $g = 1, 2, \ldots, N$, θ_a and θ_b being, respectively, the largest person vector below and the smallest vector not below the difficulty vector of any item g from the domain. In practice, it follows from Equation 12 that all item characteristic functions must be identical when the beta-binomial model is applied to a large population of examinees or to a small population with a slight dispersion in θ.

The necessary condition in applying the beta-binomial model to matrix sampling—that all items must possess (approximately) identical characteristic functions—is stringent and in practice will never be met. There is a sampling plan, however, that deviates somewhat from the idea of matrix sampling but offers in principle the same information and permits the application of the beta-binomial model without this condition, namely, independence between the distributions of Equation 11 is guaranteed when a separately drawn sample of items is administered to each person in the sample (see Lord & Novick, 1968, p. 524). It is therefore possible to avoid the necessity of fulfilling conditions that in practice can never be met by using this relatively simple experimental procedure.

A Stochastic Conception of Item Responses

Suppose that a stochastic conception of item responses (i.e., for a given person and item) is now adopted. Item responses are seen as the outcomes of a stochastic process dependent upon several person and item characteristics. The probability of a correct response is not restricted to the possible values 0 and 1 but may adopt every value on the (real) interval from 0 to 1. Earlier, it was seen that the deterministic conception is akin to the view of learning as an all-or-none process and knowledge as an all-or-none state. The stochastic conception, on the contrary, seems to be allied to the view that learning is a process in which a student improves his/her knowledge gradually. It is not so that a student either "knows" or "does not know" the items; but according to this view, it seems more natural to conceive of knowledge as a continuum underlying the item responses on which a student can take several positions, which represents the amount of mastery the student possesses with regard to the cognitive skills needed for solving the items and influences his/her probability of a successful response.

In terms of matrix-sampling theory, this means that the population matrix is no longer viewed as a deterministic matrix with cells (a, g) containing one element of $\{0, 1\}$ but as a stochastic matrix of which the cells contain *probability distributions* on $\{0, 1\}$ or, equivalently, probabilities $P_g(0|\theta_a)$ and $P_g(1|\theta_a)$ with possible values in the interval $\{0, 1\}$ and $P_g(1|\theta_a) = 1 - P_g(0|\theta_a)$. As a consequence, sampling from this matrix is to be seen, not as sampling correct responses, but as probabilities of correct responses. The item characteristic functions are not necessarily the Guttman type but may adopt any monotonic increasing form. For the sequel, it is superfluous to assume a model to explain the probability $P_g(1|\Theta_a)$. One of the logistic or normal ogive models could illustrate the unidimensional case; and one of the multidimensional generalizations, the multidimensional case, but the conclusions apply to any model that describes the probability of success as a function of the complete latent space.

Domain Sampling for a Fixed Person

First, consider the case in which randomly drawn items are administered to a fixed person with vector θ_a. Under the deterministic conception, the population matrix degenerated to a matrix with 1's and 0's representing the items which the student will always have correct (items he/she "knows") and not correct (items he/she "does not know"), respectively, over replications. It was possible to count the items of the former type and define a proportion of correct responses, as has been done in Equation 10. This can not be repeated for the stochastic conception, inasmuch as the population matrix now contains probability distributions and there are no correct responses which can be counted. Therefore, it is meaningless to define for this person a proportion of items he/she "knows." It makes sense, however, to use the probability distributions in the population matrix to introduce instead an *expected* proportion of items ν_a, responded to correctly when the entire domain is administered to examinee a:

$$\nu_a \equiv \sum_{g=0}^{N} \{P_g(0|\theta_a).0 + P_g(1|\theta_g).1\}/N$$

$$= \sum_{g=0}^{N} P_g(1|\theta_a)/N. \qquad [13]$$

(Again, the limit for N to infinity must be added when the domain is infinite). Equation 13 defines the proportion correct true score (Lord & Novick, 1968, chap. 23) for the whole population matrix, which is not surprising, since the proportion correct true score is an expected proportion correct according to the classical true score definition.

A second conspicuous difference from the deterministic conception is that simple random sampling of items is a superfluous chance mechanism, because item responses are already stochastic events. In addition, the probability that a certain item is successfully responded to by a person is equal to the joint probability of drawing this item and a successful response to it. Since the probability of being drawn is equal for all items in simple random sampling, it is a scale factor and can be excluded from consideration.

In order to be allowed to consider the responses to randomly drawn items for a person with vector θ_a as outcomes of a series of Bernoulli trials, all items of the domain must possess equal success probabilities, that is,

$$P_g(1|\theta_a) = P(1|\theta_a), \qquad [14]$$

or using Equation 13,

$$P_g(1|\theta_a) = \nu_a \qquad [15]$$

for $g = 1, 2, \ldots, N$.

Since a series of Bernoulli trials is a necessary and sufficient condition for the binomial distribution, the number of successes, X, in a sample of size n only follows a binomial distribution, which can be written as

$$P(X|n,\nu_a) = \binom{n}{X} \nu_a^X (1-\nu_a)^{n-X}, \qquad [16]$$

if the restriction formulated in Equations 14 and 15 is met. This restriction says that for a fixed examinee with vector θ_a, all item characteristic functions of the domain must intersect each other for $\theta = \theta_a$.

Assume for a moment that θ may be considered unidimensional and analyze the consequences for the difficulty parameters of the logistic models known from latent trait theory. It is clear that Equations 14 and 15 involve the restriction that all difficulty parameters of the Rasch model must be equal. According to the Rasch model, the logistic curves can only vary in location; and to meet the condition that they intersect each other for $\theta = \theta_a$, they must all have the same location or value for their difficulty parameter. The situation differs, however, for the two-parameter Birnbaum model. According to this model, the logistic curves can vary both in location and slope (or discrimination), and these curves have identical forms only if the restriction of not one but two intersections is imposed. Therefore, applying the binomial simultaneously to persons with scores θ_a and θ_b implies that for the two-parameter Birnbaum model, all items must have equal values not only for their difficulty parameters but for their discrimination parameters as well. An example of this application is found in mastery testing with an indifference zone (Fhanér. 1974; Kriewall, 1972; Wilcox, 1976), where optimal cutting scores and test length are determined by analyzing simultaneously the binomially distributed measurement errors for persons with a minimum mastery level $\nu_a = \nu(\theta_a)$ and a maximum nonmastery level $\nu_b = \nu(\theta_b)$. The same analysis shows that in order to apply the binomial model simultaneously to three persons with different latent scores, the item characteristics of the three-parameter logistic model must be identical.

Since the classical definition of item difficulty boils down to the probability $P_g(1|\theta_a)$ for a fixed person and a stochastic conception of item responses, it follows from Equation 14 that applying the binomial model in this case requires all items of the domain to be of the same classical difficulty. This is different from the deterministic conception, for which no assumptions of classical item difficulty are involved.

Matrix Sampling

Finally, suppose that persons are randomly sampled and that the question of the assumptions of item difficulty implied by the beta-binomial model is raised. Since the conditions for applying the binomial model are necessary conditions for applying the beta-binomial model, the conditions indicated above must be met. An extra requirement is now that Equation 14 must be in force not only for one point $\theta = \theta_a$ but for all points of the latent space θ. All item characteristic functions must, therefore, intersect each other for all points of θ. This implies that all item characteristic functions must be identical, regardless of the number of dimensions that make θ complete or the model that may be adopted to describe these functions.

Since identical characteristic functions mean that the items are completely equivalent, the beta-binomial model implies that all items have equal difficulty, according to the classical, or any other, definition. In this case, a stringent condition for applying a binomial model is again encountered. Note that now the necessity of fulfilling this condition cannot be avoided by leaving matrix sampling and the administering of a separately drawn sample of items to each person. The reason is that the equality of the item characteristic functions follows from the requirement that each person must have the same probability of success for each item, and this cannot be reached by changing the sampling plan. The compound binomial model allows items to vary in their probability of success for each person, and it seems wise to use this less stringent model for the conditional part in Equation 3 when confronted with a domain of unequal item characteristic functions.

Discussion

Although binomial models are widely used for solving testing problems, the item difficulty assumptions are not generally understood. The argument of this paper has shown that binomial models involve rather strong assumptions, which may be summarized as follows:

1. *Binomial Model (One-Person Case).* Adopting the deterministic conception, it is not required that all items of the domain have equal difficulties when using both the classical and the latent trait theoretic definition of this parameter. The stochastic conception, however, entails the condition of equal values for the classical, as well as the Rasch, item difficulty parameter.
2. *Beta-Binomial Model (Group-of-Persons Case).* If the sample of items is administered to all persons in the sample, either conception involves equal item difficulty for both the classical and latent trait definitions. However, if the deterministic conception is adopted, this condition can be avoided by giving separate samples of items to each person, whereas it can not if the stochastic conception is adopted.

It is clear that the most important difference is between a deterministic and a stochastic conception of item responses. Both lead to differences in the condition under which binomial models can be applied. The present writer believes that the stochastic view has greater validity than the deterministic view. As indicated earlier, the latter is akin to the conception of learning as an all-or-none process and to state-models for mastery testing, which has been criticized in another paper (van der Linden, 1978). Here, it suffices to say that delusions, fatigue, fluctuations in intellectual capacity and attention, reading errors, slips of the pen, guessing, and the like compel a conception of item responses as outcomes of a stochastic process. This idea is present in classical test theory when it uses the so-called propensity distributions to define true score. Only in modern test theory is it thoroughly explored: Latent trait theory conceives of item responses as stochastic events and, in fact, applies the propensity distribution at the item level. It also provides models that can be used for explaining the distributions.

As indicated earlier, defining a proportion of the domain an examinee knows does not make sense for the stochastic conception of item responses. In the literature about criterion-referenced measurement, it is, however, usual to consider this proportion as a typical criterion-referenced measure and to place it opposite norm-referenced measures like percentile scores (for a classical paper, see Ebel, 1962). Defining a proportion instead of an expected proportion corresponds to the idea that sampling from a domain of items is equivalent to sampling from a vase with red and white marbles. The proportion of items an examinee knows is comparable with the proportion of red marbles that vase contains and can be estimated accordingly. As noted above, unlike the color of marbles, item responses are not deterministic events; a response to an item is therefore not comparable with the color of a marble. The important difference is that the former is liable to all kinds of stochastic influences, whereas the latter is free of this. Consequently, a distinction between a proportion, which could in principle be observed when the student responds to the total domain, and a true or expected proportion must be made. Only the latter is the person parameter in which criterion-referenced measurement should be interested.

The robustness of the binomial models with respect to assumptions of item difficulty has not been considered in this paper. It is conceivable that the conditions for applying binomial models are, in practice, less strong, because numerical results are relatively independent of the degree to which these conditions are violated. Analysis of robustness should further clarify this point.

References

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, LA, March 1973. (EDRS No. ED 076 658)

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

Ebel, R. L. Content standard test scores. *Educational and Psychological Measurement*, 1962, *22*, 15–25.

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, October 1969.

Feller, W. *An introduction to probability theory and its applications* (Vol. 1). New York: John Wiley & Sons, Inc., 1968.

Fhanér, S. Item sampling and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 172–175.

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, *40*, 1–47.

Hogg, R. V., & Craig, A. T. *Introduction to mathematical statistics.* New York: MacMillan, 1972.

Huynh, H. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 1976, *13*, 263–264. (a)

Huynh, H. Statistical consideration of mastery scores. *Psychometrika*, 1976, *41*, 65–79. (b)

Huynh, H. *On mastery scores and efficiency of criterion-referenced tests when losses are partially known.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976. (c)

Huynh, H. Two simple classes of mastery scores based on the beta-binomial model. *Psychometrika*, 1977, *42*, 601–608.

Kriewal, T. E. Aspects and applications of criterion-referenced tests. *Illinois School Research*, 1972, *9*, 5–18.

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

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

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

Mellenbergh, G. J., Koppelaar, H., & van der Linden, W. J. Dichotomous decisions based on dichotomously scored items: A case study. *Statistica Neerlandica*, 1977, *31*, 161–169.

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

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205–216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education.* Berkely, CA: McCutchan, 1974.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 1976, *13*, 265–276.

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.

Wilcox, R. R. A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, 1976, *1*, 359–364.

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.

Acknowledgment

Thanks are due to Fred N. Kerlinger, Gideon J. Mellenbergh, and both reviewers for their most useful comments and to Betsy Becker for purifying some parts of the English text.

Author's Address

Send requests for reprints or further information to Wim J. van der Linden, Onderafdeling Toegepaste Onderwijskunde, T. H. Twente, Postbus 217, 7500 AE Enschede, The Netherlands.