Determining the Length of a Criterion-Referenced Test

Rand R. Wilcox
University of California, Los Angeles

When determining how many items to include on a criterion-referenced test, practitioners must resolve various nonstatistical issues before a particular solution can be applied. A fundamental problem is deciding which of three true scores should be used. The first is based on the probability that an examinee is correct on a "typical" test item. The second is the probability of having acquired a typical skill among a domain of skills, and the third is based on latent trait models. Once a particular true score is settled upon, there are several perspectives that might be used to determine test length. The paper reviews and critiques these solutions. Some new results are described that apply when latent structure models are used to estimate an examinee's true score.

When trying to determine how many items to include on a criterion-referenced test, perhaps the most fundamental problem is that there are at least three conceptualizations, or models, of an achievement test that might be used. Each of these conceptualizations is based on a different type of true score. The first deals with the number of items an examinee would get correct if he/she were to respond to every item in some item domain. The second is concerned with the proportion of skills among a domain of skills that an examinee has acquired. Because of errors at the item level, such as guessing, this conceptualization is different from the first. The final approach is based upon latent trait models. In some cases, one model might yield substantially different results from another in terms of test length, and so the choice of a model can be crucial.

Once one of the above conceptualizations is settled upon, a variety of other issues must be resolved. For example, when comparing an examinee's true score to a standard, should it be assumed that the standard is known, or should the process by which it was determined be taken into account? Should the test length problem be formulated in terms of a single examinee, a "typical" examinee, or both? How certain do we want to be of making a correct decision (classification) of an examinee? Are we willing to use a Bayesian solution?

This paper has several goals. The first is to give a brief review and critique of the three general approaches that might be used when determining the length of a criterion-referenced test. The second is to describe new results on test length, and the third is to indicate possible directions for future research.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 4 Fall 1980 pp. 425–446

The Purpose of the Test

Consistent with earlier test length solutions, it is assumed that the purpose of the test is to sort the examinees into one of two mutually exclusive groups. In addition, it is assumed that it is possible to define these two groups in terms of some notion of true score, say π, that characterizes a particular examinee. For the moment, the exact nature of the true score, π, need not be specified.

Let π_0 be a constant that may or may not be known; π_0 is referred to as the standard. An examinee is said to belong to the first group, which is designated as S_G if his/her true score is greater than or equal to π_0. If $\pi<\pi_0$, the examinee is said to belong to group S_B. The problem is to determine how many items to include on the test so that there can be a reasonable certainty of correctly determining whether an examinee belongs to S_G or S_B.

Hambleton, Swaminathan, Algina, & Coulson (1978) have described two primary uses of criterion-referenced tests, namely, estimating domain scores and classifying examinees. In this paper the concern is only with the latter use of the test scores.

Solutions Using Domain Scores, with π_0 Known

In the context of a criterion-referenced test, perhaps the most frequently used notion of true score is based on the concept of an item-sampling model (e.g., Harris, 1974; Huynh, 1976; Novick & Lewis, 1974; Wilcox, 1977). Consider, for example, a domain of dichotomously scored items. In some cases the item pool actually exists, and in other cases the notion of an item domain is a convenient conceptualization. For this model, π represents the proportion of items an examinee would answer correctly if he/she were to answer every item in the item pool. For a *single* examinee responding to a random sample of n items, the probability of getting x correct is assumed to be

$$\binom{n}{x} \pi^x (1-\pi)^{n-x}, \qquad [1]$$

the binomial probability function. This is justified when items are randomly sampled from an infinite item pool or a finite pool with replacement, and π remains constant. Using Equation 1 is also appropriate when it gives a good fit to the observed scores of an examinee. Gelfand and Thomas (1976), Katz (1963), Tarone (1979), and Cochran (1954) have discussed the problem of determining when a good fit is obtained.

One difficulty with this model occurs when more than one examinee is considered. If $g(\pi)$ represents the distribution of true scores over a population of examinees, Equation 1 implies that the marginal distribution of observed scores is given by

$$\int_0^1 \binom{n}{x} \pi^x (1-\pi)^{n-x} g(\pi)\, d\pi \qquad [2]$$

Lord and Novick (1968, sec. 23.8) have shown that Equation 2 implies that the correlation between observed scores and true scores is given by the KR21 reliability formula. If every examinee takes the same n items, the implication is that every item has the same level of difficulty. This result prompted Lord and Novick to replace Equation 1 with a two-term approximation to the more general compound binomial model, the approximation being given by

$$P_n(x) + d\pi(1-\pi)C(x) \qquad [3]$$

where

$$P_n(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad [4]$$

and

$$C(x) = \sum_{v=0}^{2} (-1)^{v+1} \binom{2}{v} P_{n-2}(x-v) \quad [5]$$

The parameter d is equal to

$$\frac{n^2(n-1)\sigma_\rho^2}{2\{\mu_x(n-\mu_x)-\sigma_x^2 - n\sigma_\rho^2\}} \quad [6]$$

where μ_x^2 and σ_x^2 are the mean and variance, respectively, of the marginal distribution of observed scores and where σ_p^2 is the variance of the item difficulties. It should be noted that if every examinee takes a different random sample of n items, the simpler binomial probability function is theoretically justified and Equation 2 is correct. Further comments concerning Equation 3 are made below.

Apparently, the earliest attempt at providing a solution to the test length problem was made by Millman (1973) using the binomial probability function. Shortly thereafter, Fhanér (1974) gave a more formal approach, again using the binomial probability function but with an indifference zone built into the analysis. This means that in addition to the known constant π_0, a constant $\delta^* > 0$ is specified with the idea that if the examinee's proportion correct true score is less than or equal to $\pi_0 - \delta^*$ or greater than or equal to $\pi_0 + \delta^*$, there should be a reasonable certainty of making a correct decision. If $\pi_0 - \delta^* < \pi < \pi_0 + \delta^*$, any decision is said to be correct.

The goal can be stated more precisely as follows: Let n_0 be a specified passing score, i.e., if the examinee's observed score is greater than or equal to n_0, the decision $\pi \geq \pi_0$ is made; otherwise, the decision is $\pi < \pi_0$. The problem is to determine the smallest n, so that regardless of the actual value of π, the probability of a correct decision, $P(CD)$, is reasonably high, say greater than or equal to P^*. More briefly, it is desired that

$$P(CD) \geq P^*, \quad \tfrac{1}{2} < P^* < 1 \quad [7]$$

The reason for requiring $P^* > \frac{1}{2}$ is that it can be guaranteed that the $P(CD)$ is at least .5 without any observations at all, simply by randomly deciding whether π is above or below π_0.

For $\pi < \pi_0$,

$$P(CD) = \sum_{x=0}^{n_0-1} \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad [8]$$

and for $\pi \geq \pi_0$,

$$P(CD) = \sum_{x=n_0}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x} \quad [9]$$

Moreover, it can be shown that for $\pi \leq \pi_0 - \delta^*$, Equation 8 is minimized at $\pi = \pi_0 - \delta^*$ for any n and that for $\pi \geq \pi_0 + \delta^*$, Equation 9 is minimized at $\pi = \pi_0 + \delta^*$. Thus, to satisfy Equation 7 for any π, it is sufficient to find the smallest n so that for $\pi = \pi_0 - \delta^*$, Equation 8 exceeds P^*; and simultaneously for $\pi = \pi_0 + \delta^*$, Equation 9 also exceeds P*. Note that for $\delta^* = 0$, both Equations 8 and 9 approach .5. Thus, $\delta^* > 0$ is a necessary condition for ensuring that an n exists satisfying Equation 7.

It is also of interest to observe that it is relatively easy to incorporate virtually any loss function into the above framework. Suppose, for example, $L_1(\pi) \geq 0$ is the loss associated with misclassifying an examinee for whom $\pi < \pi_0$ and let $L_2(\pi) \geq 0$ be the loss when $\pi \geq \pi_0$. The risk or expected loss is

$$\sum_{x=n_0}^{n} L_1(\pi) \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad \pi < \pi_0 \qquad [10]$$

$$\sum_{x=0}^{n_0-1} L_2(\pi) \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad \pi \geq \pi_0 \qquad [11]$$

Using numerical procedures, the values of π maximizing Equations 10 and 11 are readily determined. Moreover, if there is an open interval around π_0 such that $L_1(\pi) = L_2(\pi) = 0$, and if L_1 and L_2 are bounded above, both Equations 10 and 11 can be made arbitrarily small.

Before concluding this subsection, it should be noted that Wilcox (1979a) has generalized Fhanér's solution in two directions. In particular, Wilcox's solution applies to any model involving some notion of true score π (not necessarily domain scores) for which there exists a statistic $\hat{\pi}(x)$ for estimating π such that the cumulative distribution function of $\hat{\pi}(x)$, say $F(\hat{\pi}(x)|\pi)$, is stochastically increasing. In other words, it is assumed that $\pi < \pi'$ implies that $F(\hat{\pi}(x)|\pi') \leq F(\hat{\pi}(x)|\pi)$ for all x. Consider any group of k examinees and let g be the number of examinees for whom $\pi \geq \pi_0$. It follows that over all possible configurations of true score, the minimum probability of a correct decision is given by

$$\prod_{i=1}^{k-g} F(\pi_0 | \pi_i = \pi_0 - \delta^*) \times \prod_{j=k-g+1}^{k} 1 - F(\pi_0 | \pi_j = \pi_0 + \delta^*) \qquad [12]$$

where $\pi_1, \ldots, \pi_{k-g}$ are the true scores of the $k-g$ examinees for whom $\pi < \pi_0$ and $\pi_{k-g+1}, \ldots, \pi_k$ are the true scores for the examinees having $\pi \geq \pi_0$. It has been shown that in terms of g, Equation 12 is minimized at $g=0$ or $g=k$ if $\hat{\pi}_i(x)$, the statistic for estimating π_i, is independent of $\hat{\pi}_j(x)$, $i \neq j$. Thus, by examining these two cases and choosing n accordingly, Equation 7 can be guaranteed no matter what the values of the π_i's happen to be. Wilcox's (1979a) solution contains the binomial error model, Poisson process models, and normal distributions as a special case. Finally, in the case of proportion correct true score, Wilcox (1979a) has indicated that the simpler binomial model appears to give a conservative solution when the conditional distribution of observed scores is given by a two-term approximation to the compound binomial model, as described by Equation 3. In other words, the binomial error model appears to result in a longer test length than would be obtained using Equation 3, all other things be equal. However, a rigorous proof that this is the case has not been derived.

An important feature of the test length solution proposed by Fhanér (1974) and extended by Wilcox (1979a) is that it is conservative in the sense that it makes no assumption about the value of the examinee's true score, π. Furthermore, it is assumed that there is no information beyond an examinee's observed score for making the decision about whether π is above or below π_0. For a single examinee, Fhanér's solution might result in the use of a moderate number of items on the test. Suppose, for example, it is assumed that the binomial error model holds, that $k=1$, $P^*=.9$, $\delta^*=.1$, $\pi_0=.8$ and that the passing score n_0 is chosen to be the smallest integer such that $n_0/n \geq .8$. It follows that $n=26$ is the shortest test length satisfying Equation 7.

In practice, however, it is often desired to simultaneously make a decision about $k>1$ examinees. If it is insisted that a conservative approach to the test length problem should be used and if the decision-making process is viewed in terms of k examinees, the minimum probability of a correct decision decreases rapidly as k gets large (see Wilcox, 1979a).

For large k, it might be argued that it is conservative but unrealistic to consider the case where the value of every examinee's true score is equal to $\pi_0-\delta^*$ or $\pi_0+\delta^*$. Thus, some other perspective might be deemed more appropriate when judging the adequacy of the length of the test. The remainder of this section considers how this might be done.

A Bayesian Approach

Novick and Lewis (1974) have described a Bayesian approach to determining n that applies to the case of a single examinee whose conditional distribution of observed scores is given by the binomial probability function. As is typically done for the binomial case, an examinee's true score is viewed as a random variable with a distribution belonging to the beta family. More specifically, it is assumed that the probability density function of π is given by

$$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\,\pi^{r-1}\,(1-\pi)^{s-1} \qquad [13]$$

where $r>0$ and $s>0$ are unknown parameters and Γ is the usual gamma function. If r and s were known, it might be possible to justify a shorter test length than would be required if the approach used by Fhanér (1974) were employed. Another appealing feature of the Bayesian solution is that the probability of a correct decision would be known, given the examinee's observed score.

More specifically, the test length solution is formulated as follows: Given Equation 13, and assuming Equation 1 holds, it is known that the conditional distribution of π, given an observed score x, is

$$h(\pi \mid x) = \frac{\Gamma(n)\pi^{x}(1-\pi)^{n-x}}{\Gamma(x+1)\Gamma(n-x+1)} \qquad [14]$$

This is a beta distribution with parameters $x+1$ and $n-x+1$. Thus, the probability of $\pi \geq \pi_0$ when $x=n_0$ can be computed. The test length is determined by increasing n until this probability is believed to be reasonably close to 1.0.

Novick and Lewis (1974) also illustrate how to incorporate a simple loss function into their analysis. In particular, let a be the "loss" of passing an examinee who should fail, and let b be the cost of failing an examinee who should pass. A student is advanced if $bP(\pi \geq \pi_0 | x, n) \geq aP(\pi < \pi_0 | x, n)$. Note that only the ratio a/b needs to be specified, not the actual values of a and b, since the ratio a/b is being compared to $Pr(\pi \geq \pi_0 | x, n)/Pr(\pi < \pi_0 | x, n)$.

Morgan (1979) has extended the work of Novick and Lewis (1974) to situations where guessing and carelessness are incorporated into the analysis. Novick (1973) has discussed the specification of the prior distribution; and Novick and Lewis (1974) have stated that the specification of the prior must be done carefully, suggesting that the book by Novick and Jackson (1974) and the paper by Novick, Lewis, and Jackson (1973) might aid in this process.

Most issues in statistics are controversial, at least to some degree. Consider, for example, the problem of estimating the mean of a distribution. The sample mean has various optimal properties under certain circumstances (e.g., normality), but a variety of alternative estimates might be used instead (Andrews, Bickel, Hampel, Huber, Rogers, & Tukey, 1972). When discussing Bayesian solutions to a problem, it seems prudent to remind the reader that this area of statistics is a bit more controversial than others. Bartlett (1971) noted the following:

> I would say that the statisticians' model is different in principle from a prior distribution in that it can be tested. Where it cannot be tested this is to me unsatisfactory. Prior distributions are, as I understand it, in general untestable. What does Professor Lindley mean when he says that "the proof of the pudding is in the eating"? If he has done the cooking it is not surprising if he finds the pudding palatable, but what is his reply if we say that we do not. If the Bayesian allows some general investigation to check the frequency of errors committed, or even real losses, this might be set up; but if the criterion is inner consistency, then to me this is not acceptable. (p. 447)

Despite Bartlett's comment, the importance of Bayesian statistics should not be underestimated. Even if one insists on the classical approach, Bayesian methods may prove to be valuable (e.g., Murray, 1977). For further favorable comments toward the Bayesian approach to statistical inference, the reader is referred to Kendall & Stuart (1973, pp. 159–161).

An Alternative Approach

There is an approach to statistical inference developed by Dempster (1966, 1967), which might be applied to the test length problem. Apparently, this approach has not been discussed in terms of the problem at hand; and thus, for completeness, it is described here. Dempster's results are quite general; but for the sake of clarity, the discussion is limited to the case of a single examinee for whom the binomial error model applies.

Suppose $\pi_0=.7$, $n=10$, $n_0=7$ and that an examinee's observed score is $x=6$. Thus, the decision $\pi<\pi_0$ would have been made. If the observed score had been $x=1$, say, the same decision would have been made but one might "feel" more certain that the correct decision has been reached. Assuming that one should feel more certainty about the decision when $x=1$ versus $x=6$, the question arises as to how to express this certainty in some meaningful way. For the Bayesian statistician, the problem is relatively straightforward, since once the beta prior has been specified, $P(\pi<\pi_0|x=1)$ and $P(\pi<\pi_0|x=6)$ can be calculated.

Dempster's theory does not give an exact value for $P(\pi<\pi_0|x)$; rather it yields two values, say P_1 and P_2, which are interpreted as lower and upper bounds, respectively, on $P(\pi<\pi_0|x)$. Bounds on $P(\pi\geq\pi_0|x)$ can also be derived.

From Dempster (1968a) it can be seen that for $P(0\leq\pi\leq\pi_0|x)$,

$$P_2 = \sum_{y=x}^{n} \binom{n}{y} \pi_0^y \quad (1-\pi_0)^{n-y} \qquad [15]$$

and

$$P_1 = \sum_{y=x}^{n-1} \binom{n}{y} \pi_0^y \quad (1-\pi_0)^{n-y} \qquad [16]$$

As for $P(\pi_0 \leqslant \pi \leqslant 1|x)$

$$P_2 = \sum_{y=0}^{x} \binom{n}{y} \pi_0^y \quad (1-\pi_0)^{n-y} \qquad [17]$$

$$P_1 = \sum_{y=0}^{x-1} \binom{n}{y} \pi_0^y \quad (1-\pi_0)^{n-y} \qquad [18]$$

In terms of test length, one might choose an n so that P_1 is reasonably close to 1.0 for $P(0 \leqslant \pi < \pi_0 | x < x_0)$ and $P(\pi_0 \leqslant \pi \leqslant 1 | x \geqslant x_0)$. If, however, an examinee's observed score is x_0, it can be seen that such an n may not exist. One way out of this dilemma is to incorporate an indifference zone into the analysis but in a slightly different fashion than was done in the earlier portion of this paper. Here the bounds on $P(0 \leqslant \pi \leqslant \pi_0 + \delta^* | x)$ and $P(\pi_0 - \delta^* \leqslant \pi \leqslant 1 | x)$ might be used. In this case, an n would be chosen so that

$$\sum_{y=n_0-1}^{n-1} \binom{n}{y} (\pi_0 + \delta^*)^y \quad (1-\pi_0-\delta^*)^{n-y} \qquad [19]$$

the lower bound on $P(0 \leqslant \pi \leqslant \pi_0 - \delta^* | x = n_0 - 1)$, and

$$\sum_{y=0}^{n_0-1} \binom{n}{y} (\pi_0 - \delta^*)^y \quad (1-\pi_0+\delta^*)^{n-y} \qquad [20]$$

the lower bound on $P(\pi_0 - \delta^* \leqslant \pi \leqslant 1 | x = n_0)$, are reasonably close to 1.0.

In practice, it would seem that this approach might yield nearly the same results as those obtained with the classical methods described earlier. However, Dempster's approach might be preferred over the usual Bayesian solution because it is possible to incorporate into the analysis prior beliefs about whether π is above or below π_0 without specifying a specific form for the prior. Dempster (1968a) has illustrated how this might be done. (For further comments on this approach to making inferences, the reader is referred to the discussion following the paper by Dempster, 1968b.)

Error Rates for the Typical Examinee

In addition to considering the adequacy of the test length in terms of a single examinee or $k>1$ specific examinees, examinees being tested might be considered to be a random sample from some larger population of examinees and the question might arise as to whether n is sufficiently large for the typical examinee being tested (Wilcox, 1977). (See also Huynh, 1980; Livingston & Wingersky, 1979). It is not being suggested that the analysis presented in this subsection replace the approaches described above. Rather, the results reviewed and outlined here might be used to give additional insight into whether or not there are an adequate number of items on the test.

Assume that observed test scores x_i (i=1,. . .,k) are available for k examinees and consider the estimation of α=$P(x \geq x_0, \pi < \pi_0)$, the probability of a false-positive decision, and β=$P(x < x_0, \pi \geq \pi_0)$, the probability of a false-negative decision for a randomly selected examinee. Since there are observations for estimating α and β, the present approach might be termed a retrospective study. This is in contrast to Fhanér (1974) and Wilcox (1979a), where it is assumed that no information is available concerning an examinee's true score, and so the problem is more along the lines of designing an experiment.

Note that π is again an unknown fixed constant for a specific examinee; no prior distribution is considered for an examinee as is done in the Bayesian solution. In referring to a distribution for π, say $w(\pi)$, the distribution of π is now over a population of examinees.

Let $h(x|\pi)$ be the conditional distribution of observed scores for an examinee having true score π. It follows that

$$\alpha = \sum_{x=n_0}^{n} \int_0^{\pi_0} h(x|\pi)w(\pi)\,d\pi \qquad [21]$$

and

$$\beta = \sum_{x=0}^{n_0-1} \int_{\pi_0}^{1} h(x|\pi)w(\pi)\,d\pi \qquad [22]$$

To define α and β in terms of an indifference zone, simply replace π_0 with $\pi_0-\delta^*$ in Equation 21 and replace π_0 with $\pi_0+\delta^*$ in Equation 22. If $h(x|\pi)$ and $w(\pi)$ were known, α and β would be known. If α and β were judged to be too large, their values could be decreased by increasing the test length.

For most situations neither $h(x|\pi)$ nor $w(\pi)$ is known. Suppose, however, following Lord and Novick (1968, chap. 23), it is assumed that $h(x|\pi)$ is some approximation to the compound binomial distribution. Further suppose that the moments of the true score distribution can be estimated using the x_i's. Lord (1965) has described how to do this when the two-term approximation to the compound binomial given by Equation 3 is deemed to be appropriate. Once these estimates are available, the methods described below for estimating $w(\pi)$ can be employed.

Perhaps the most frequently used approach to estimating the true score distribution is to assume that $w(\pi)$ belongs to the family of beta distributions (e.g., Keats & Lord, 1962; Lord, 1965). If the true score distribution is unimodal, a good approximation to it may be possible with a beta distribution (Springer, 1979, p. 268). If the true score distribution is multimodal, it is not clear when or even if a good fit to the true score distribution can be obtained. One possible problem is that (excluding U-shaped distributions) beta distributions can have at most one mode. To complicate matters, there is no satisfactory method of detecting a poor fit to the true score distribution using the observed x_i's. The difficulty is that even if the first m moments of π over the population of examinees (m being any integer) are given, the true score distribution is not uniquely determined. There are alternative approaches to estimating the distribution of π (e.g., Blischke, 1964; Lord, 1969; Maritz, 1970; von Mises, 1964, pp. 384–401), but the circumstances under which these procedures give more accurate results appears to be unknown.

Because of the difficulty in determining the accuracy of point estimates of α and β, there is some doubt as to when such estimates should be relied upon when judging the adequacy of the test length. An alternative approach that might be used is to estimate bounds on α and β that make no assumptions about the shape of the true score distribution. Wilcox (1979c) has indicated how this can be

done. The solution is based on estimating the first two moments of π and applying results by Skibinsky (1977) that yield bounds on the probability that π is in a particular interval. It should also be pointed out that an earlier paper by Skibinsky (1976) has described how to use the first three moments of the true score distribution to obtain upper bounds to the probability that π is in the interval $(0, \pi_0)$ or $(\pi_0, 1)$. Following Wilcox (1979c), these bounds might also be used to determine bounds on α and β.

The beta-binomial model appears to give a good estimate of α and β, even when the conditional distribution of observed scores are generated according to the two-term approximation of the compound binomial given by Equation 3 (Wilcox, 1977). Thus, when investigating the adequacy of the test length, the beta-binomial model would seem to suffice when the true score distribution belongs to the beta family. Moreover, a moderate number of examinees usually gives a reasonably accurate estimate of α and β when the beta-binomial model holds. However, there are occasions when observed scores on a moderate number of examinees can result in extremely inaccurate estimates of the parameters of the beta distribution (Wilcox, 1979e). This is a highly unusual event, but it seems prudent to keep this fact in mind when considering test length.

Solutions Using Domain Scores, with π_0 Unknown

Thus far a very brief outline and review of procedures for judging test length have been given, all of which assume that the cutoff score, π_0, is known. In reality the cutoff score is not known; rather, it is determined by some process. Huynh (1976), for example, has described a method for determining π_0 when an external criterion exists.

One important aspect of a criterion-referenced test is the effect the process of determining π_0 has on the test length, n. In other words, it may be of interest to incorporate this process into the analysis. This is done by Wilcox (1979b) for the case where π_0 is the unknown parameter of some distribution. The examples given are based on the notion that the control (the distribution characterized by π_0) is a population of examinees. In this section a variation of this situation is considered.

In practice the cutoff score is often specified by a panel of judges. In an attempt to better approximate reality, the following conceptualization is used. A total of k judges have specified a cutoff score, π_{0i} $(i=1,\ldots,k)$. Furthermore, these k judges are viewed as a random sample from some population of individuals who are qualified for specifying π_0. In particular, it is assumed that the realization of π_{0i} is independent of π_{0j}, $i \neq j$ and that π_0 is the mean of the cutoff scores that would be specified by the population of judges. Since π_0 is unknown, it is estimated with $\bar{\pi}=k^{-1}\sum_{i=1}^{k}\pi_{0i}$. Accordingly, if an examinee's true score is estimated to be greater than or equal to $\bar{\pi}_0$, the decision $\pi \geq \pi_0$ is made; otherwise, the reverse is said to be true.

Let $G(\bar{\pi})$ be the cumulative distribution function of $\bar{\pi}$ and consider the case of a single examinee for whom the binomial error model holds. It follows that the probability of a correct decision is given by

$$\int_0^1 \sum_{x=[n\bar{\pi}_0]}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x} \, dG(\bar{\pi}) \qquad [23]$$

if $\pi \geq \pi_0$

or by

$$\int_0^1 \sum_{x=0}^{[n\bar{\pi}-1]} \binom{n}{x} \pi^x (1-\pi)^{n-x} \, dG(\bar{\pi}), \qquad [24]$$

$$\text{if } \pi \geq \pi_0$$

where $[\bar{n}]$ is the smallest integer greater than or equal to $\bar{n}$.

Again, it is necessary to specify an indifference zone (i. e., a $\delta^* > 0$) to be certain that there exists an n so that Equation 7 is satisfied. From Wilcox (1979b) it follows that the minimum probability of a correct decision is given by

$$\int_0^1 \sum_{x=0}^{[n\bar{\pi}-1]} \binom{n}{x} (\pi_0 - \delta^*)^x (1-\pi_0+\delta^*)^{n-x} \, dG(\bar{\pi}) \qquad [25]$$

or

$$\int_0^1 \sum_{x=[n\bar{\pi}]}^{n} \binom{n}{x} (\pi_0 + \delta^*)^x (1-\pi_0-\delta^*)^{n-x} \, dG(\bar{\pi}) \qquad [26]$$

whichever is smallest.

There remains the technical problem that $G(\bar{\pi})$ is unknown. One approach would be to assume that $G(\bar{\pi})$ is the distribution that minimizes the $P(CD)$ so that no matter what the distribution of $G(\bar{\pi})$ happens to be, n can be chosen so that $P(CD) \geq P^*$. A method of deriving this distribution is unknown to the author. However, Wilcox (1979b) suggested that if it is assumed that

$$P(\pi_{0i} = 0) = P(\pi_{0i} = 1) = 1/2 \qquad [27]$$

a reasonably conservative solution to the test length problem will be obtained. Note that this distribution is the limiting form of a noninformative beta prior used in Bayesian statistics. (See, e.g., Aitchison & Dunsmore, 1975, chap. 2.)

There are three reasons for suspecting that the distribution given by Equation 27 will give a conservative solution when specifying n. The first is that this distribution has the maximum possible variance of any distribution in the closed interval [0, 1]. The second reason stems from considering the asymptotic case. Finally, familiarity with variational methods (e.g., Rustagi, 1976) suggests that the minimum of Equations 25 and 26 over all possible distributions $G(\bar{\pi})$ occurs when $G(\bar{\pi})$ is a step function.

Note that when Equation 27 holds, $G(\bar{\pi})$ is a binomial distribution. Thus, the approximate solution for specifying n that is given by Wilcox (1979b, Expression 7) can be applied to the present situation. Suppose, for example, $P^*=.9$ and $\delta^*=.1$. From Wilcox (1979b, Table 1), $n=k=84$ is required. In other words, if $n=84$ items are administered to an examinee and if $k=84$ judges specify a criterion score, there is (approximately) at least a 90% chance of correctly classifying the examinee.

In practice, there might be at least two objections to the procedure just given. The first is that it might be too conservative, in the sense that it is unrealistic to expect (or perhaps even to allow) a judge

to specify $\pi_0 = 0$ or 1. If it is assumed that every judge will specify a π_0 that is between .5 or .9, perhaps a fewer number of judges would be required. The second objection (related to the first) is that the variance of π_0 over judges might be small relative to the variance of the observed score of an examinee so that there is no need for sampling as many judges as items, as was done for convenience in the illustration given above.

When comparing a single examinee's proportion correct true score π to π_0, the goal may be viewed as determining which of two populations has the larger mean (i.e., attempting to determine whether π is larger or smaller than π_0). Thus, in the asymptotic case, the results given by Bechhofer (1954) may be applied. How this might be done is presented in the following illustration.

For any random variable y having mean μ and variance σ^2 defined on the closed interval $[a, b]$, $\sigma^2 \leqslant (\mu - a)(b - \mu)$ with equality holding when $P(y=a) = (b-\mu)/(b-a)$ and $P(y=b) = 1 - P(y=a)$ (e.g., Skibinsky, 1977). Suppose, for the sake of illustration, it is assumed (or required) that $.5 \leqslant \pi_0 \leqslant .9$. It follows that the maximum possible variance of π_0 is .04, which occurs when $\pi_0 = .7$ and $P(\pi_0 = .5) = P(\pi_0 = .9) = .5$. Thus, in an attempt to find a conservative choice for n and k, the case in which $\pi_0 = .7$ and the variance of π_0 is .04 is considered.

Suppose $P^* = .9$ and $\delta^* = .1$. Via the central limit theorem, the solution proposed by Bechhofer (1954, p. 24) may be applied. In particular, the required number of judges is $k = d(.04)/(\delta^*)^2$, where d is read from Bechhofer's Table 1 (the column headed, in Bechhofer's notation, with $k=2$ and $t=1$). For $P^* = .9$, $d=1.8124$, and after rounding, $k=13$. As for n, first observe that in the example, $\pi_0 - \delta^* = .6$ and $\pi_0 + \delta^* = .8$ (i.e., under the assumptions made, the $P(CD)$ is minimized either when the examinee's true score is .6 or .8). Since a binomial distribution with probability of .6 has a larger variance than when $\pi = .8$, consider the case $\pi = .6$. Thus, for this special case, the variance of a binomial distribution for a single observation is .24, and so $n \approx (1.8124)^2(.24)/.01 \approx 78$. This result also follows from Bechhofer's Equation 34. (For related comments on the actual $P(CD)$ in the case of normal distributions, see Lam & Chiu, 1976; Tong & Wetzell, 1979.)

As a partial check on the accuracy of the approximate solution for k and n, monte carlo procedures were used to estimate the $P(CD)$ with $k=13$, $n=78$, $P(\pi_{0i} = .5) = P(\pi_{0i} = .9) = .5$ and $\pi = .6$. The resulting estimate was .912. As a further check, the approximate solution for $P^* = .75$ and .95 was used. The corresponding values of (k, n) were (4, 22) and (22, 130), respectively. The estimated $P(CD)$'s were .76 and .94.

As a final comment, note that when π_0 is known, the above illustrations indicate that the required number of items on the test is reduced considerably. For instance, suppose it is known that $\pi_0 = .7$. In this case, for a given P^* and δ^*, the minimum required test length is approximately equal to

$$\lambda^2 .7(.3)/(\delta^*)^2 \qquad [28]$$

where λ is the P^* quantile of the standard normal distribution (Wilcox, 1979a). Thus, the values of n corresponding to $P^* = .75$, .9, .95 are approximately 10, 34, and 57, respectively. It can be seen, therefore, that having precise information regarding π_0 can have a substantial effect on the test length.

Bayesian Solutions When π_0 is Unknown

It is possible to transform a binomial distribution to a normal distribution having known variance (e.g., Freeman & Tukey, 1950). If the distribution of π_0 is assumed to be normal with known variance, and if the Freeman-Tukey transformation is applied to the observed score of the examinee, the Bayesian approach described by Huang (1975) might be applied. However, the main results reported by Huang are concerned with finding optimal decision rules (Bayes procedures) for determining whether

π is greater than or less than π_0; no discussion is given on finding the smallest n so that Equation 4 is attained.

Solutions in Terms of Proportion of Skills Acquired

In this section it is assumed that it is meaningful to say that an examinee either "knows" or "does not know" the answer to a particular item on a test. Alternatively, it might be said that an examinee has or has not acquired the skill that is represented by a particular test item. Still another description of the approach taken here would be to say that the probability of a correct response to an item is a function of a dichotomized latent trait (Harris & Pearlman, 1978).

It should be stressed that when an examinee is described as either knowing or not knowing the correct response to an item, no implication is being made that learning is all or none. Consider any model, for example, a latent trait model (see, e.g., Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978) or classical test theory, in which the probability of a correct response is a function of some continuous unobservable variable. Either this variable has a value at which the examinee has a tendency to answer the item correctly (the probability is greater than or equal to .5) or the examinee has a tendency to answer it incorrectly. In some cases, it might be desirable to make inferences about this tendency, as is the case in the Lazarsfeld-Kendall "turnover" model as described by Goodman and Kruskal (1959). No insistence is being made that such a continuous unobservable variable exists. The point is that describing an examinee as knowing or not knowing does not rule out or have implications about some other continuous latent trait variable or model, since it is always possible to go from any latent trait to a latent state model. If should be noted, however, that the latent class point of view used in this section is deterministic in the sense that if an examinee's latent state is known, and *if* there were no errors at the item level, the examinee's observed response could be predicted. This special case, from the point of view of latent trait theory, results in Guttman item characteristic curves (cf. van der Linden, 1979). A discussion and further clarification of the relative merits of using latent classes in mental test theory are presented in Reulecke (1977) and the references cited therein, and so further comments are omitted.

There are two different but highly related approaches that have been considered, based on the framework just described. The first, which seems to have received the most attention in the literature, is to consider a specific skill in terms of a population of examinees (e.g., Harris & Pearlman, 1978; Marks & Noll, 1967). Macready and Dayton (1977) illustrated how this point of view can be used, among other things, to determine the number of items to be used when making a mastery/nonmastery decision concerning a particular skill. Their solution was recently extended by Bergan, Cancelli, & Luiten (1980).

This section concentrates on the second point of view, which considers a single examinee in terms of a domain of skills. Let ξ be the proportion of skills that the examinee knows. Consistent with previous sections, the goal is to determine whether ξ is above or below some known cutoff score, ξ_0. The problem is to find a minimum value for n, the test length, so that regardless of the actual value of ξ, there is a reasonable certainty of making a correct decision whenever $\xi \leqslant \xi_0 - \delta^*$ or $\xi \geqslant \xi_0 + \delta^*$. If ξ is in the open interval $(\xi_0 - \delta^*, \xi_0 + \delta^*)$, i.e., the indifference zone, any decision is said to be correct.

One reason for considering this conceptualization of testing is that it occurs in real-life situations. For example, certain state-wide testing programs designed to determine a student's eligibility for graduation from high school have taken this view. A second reason is that it provides an interesting perspective on the test length problem. As alluded to earlier, formulating the problem in terms of ξ rather than the domain score π, can have a dramatic effect on the value of n. Finally, when measuring

achievement, it seems reasonable to formulate the problem in terms of ξ, the proportion of skills an examinee has acquired.

Consider two errors at the item level. They are

$$\gamma = P(\text{incorrect} \mid \text{examinee knows}) \quad [29]$$

and

$$\varepsilon = P(\text{correct} \mid \text{examinee does not know}). \quad [30]$$

From a frequentist point of view, ε can be interpreted as the proportion of correct responses an examinee would get among all the items in the item pool he/she does not know. Alternatively, ε might be defined as the probability of a correct response to the *same* item over independent trials. This is similar to using the propensity distribution in classical test theory (Lord & Novick, 1968, chap. 2) except that here the distribution is defined in terms of an item an examinee does not know. To avoid the estimation problems noted by Wilcox (1979e), the former definition of ε is used. Of course, γ can be defined in an analogous fashion.

A fundamental problem with this approach to testing is deciding whether additional errors at the item level should be included in the analysis. Duncan (1974), for example, argued that in some cases a misinformation model should be used. That is, there is allowance for the possibility that an examinee chooses an incorrect response to a multiple-choice test item because he/she believes it is indeed correct. Here, however, only the errors represented by Equations 29 and 30 are considered.

Wilcox (1979d) has given some consideration to the relationship between γ, ξ, and n, the test length. It was found that if one item per skill was used, an extremely large number of items might be needed to satisfy Equation 7. Suppose, for example, $P^*=.9$, $\delta^*=.1$ and $\xi_0=.8$. Further assume that $.1 \leqslant \xi \leqslant .3$ and $0 \leqslant \gamma \leqslant .1$. In this case, over 2,600 items would be required to guarantee that $P(CD) \geqslant P^*$! The problem is that by allowing γ and ε to have positive values, the indifference zone is being shrunk in terms of the domain score π. One approach to this problem, which is considered by Wilcox (1979d), is to use more than one item per skill. It was found that this might lower the overall number of items on the test; however, a large number of items might still be required. Some alternative solutions can be considered.

In the case of multiple-choice test items, one possible approach is to use the usual correction for guessing formula score. (For a Bayesian formula score, see Molenaar, 1977.) Assuming that one item per skill is used, and that each item has m alternatives from which to choose, the formula score is $x-(n-x)/(m-1)$, where, as before, x is the observed (number correct) score. This suggests that ξ be estimated with

$$\hat{\xi} = n^{-1}[x-(n-x)/(m-1)] \quad [31]$$

(See also van den Brink & Koele, 1980.)

Note that $\hat{\xi}$ can be negative, in which case ξ is estimated to be zero.

Suppose that it is inferred that ξ is less than $\xi_0=.8$ if $\hat{\xi}<.8$, and if $\hat{\xi} \geqslant .8$ it is decided that $\xi \geqslant .8$. Let x_0 be smallest integer, such that $\hat{\xi} \geqslant .8$. When $\xi=\xi_0-\delta^*$, the examinee's domain score is given by

$$\pi_1 = (1-\gamma)(\xi_0-\delta^*) + \varepsilon(1-\xi_0+\delta^*) \quad [32]$$

and for $\xi = \xi_0 + \delta^*$, π is equal to

$$\pi_2 = (1-\gamma)(\xi_0+\delta^*) + \varepsilon(1-\xi_0-\delta^*) \quad . \qquad [33]$$

Thus, for $\xi = \xi_0 - \delta^*$

$$P(CD) = \sum_{x=0}^{x_0-1} \binom{n}{x} \pi_1^x (1-\pi_1)^{n-x} \qquad [34]$$

and for $\xi = \xi_0 + \delta^*$

$$P(CD) = \sum_{x=x_0}^{n} \binom{n}{x} \pi_2^x (1-\pi_2)^{n-x} \qquad [35]$$

To give some indication of the properties of using Equation 31, the smallest test length was determined so that simultaneously Equations 34 and 35 are at least $P^*=.9$ with $\delta^*=.1$. The results are reported in Table 1 for $m=4,5$ and various values of γ and ε.

Table 1
Minimum Test Lengths Using
Correction for Guessing Formula Scores,
$\xi_0=.8$, $\delta^*=.1$ and $P^*=.9$

γ	ε	n(m=4)	n(m=5)
0	.15	35	35
0	.30	65	93
0	.40	159	281
.02	.15	54	44
.02	.30	51	56
.02	.40	113	205
.05	.15	419	188
.07	.15	>1500	>1400

In some cases, when the test length is formulated in terms of ξ, using Equation 31 can substantially reduce the value of n over what would otherwise be required because, in essence, the passing score is being adjusted in a manner appropriate for what the value of γ and ε happen to be. This result is to be expected. The important point made by the values of n in Table 1 is that the solution to the test length problem is highly sensitive to the values of γ and ε. Moreover, for the cases considered, the closer γ and ε are to zero, the smaller is the resulting value of n, but it can be verified that this is not always the case. In practice, it is frequently assumed that $\gamma=0$ and that guessing is at random. It is generally conceded that this assumption is unrealistic, but it is often made anyway (e.g., Duncan, 1974). Weitzman (1970) proposes a procedure for ensuring guessing is at random. If this procedure is successfully implemented, the number of items that would otherwise be required might substantially be reduced.

Solutions Using Latent Structure Models

Since the test length is sensitive to the values of γ and ε, it would be helpful to have some method of estimating γ and ε or to have an estimate of ξ that does not assume guessing is at random. Under certain circumstances, such estimates are available (e.g., Anderson, 1954; Goodman, 1974, 1979; Lazarsfeld & Henry, 1968; McHugh, 1956). In this section test length is considered when these methods are applied to estimate ξ.

Consistent with the previous section, it is assumed that an examinee is responding to a random sample of sets of equivalent items. Only situations involving pairs or triplets of equivalent items are considered. (For a general approach to the case of items representing hierarchically related skills, see Dayton & Macready, 1976. For an approach to determining the equivalency of item pairs the reader is referred to Baker & Hubert, 1977, as well as Hartke, 1978; see also Wilcox, 1980a.) Note that the usual method of judging the adequacy of the latent structure model is via the chi-square goodness-of-fit test, as illustrated by Macready & Dayton (1977).

Given the willingness to make the assumptions necessary for the application of latent structure models, there are at least two technical problems with determining test length. The first is that it can no longer be certain that the $P(CD)$ is minimized at either $\xi=\xi_0-\delta^*$ or $\xi=\xi_0+\delta^*$ unless perhaps an asymptotic argument is used or numerical techniques are employed. The second is that there is no convenient method of determining, or even approximating, the smallest n so that Equation 7 holds. No attempt is made to solve these problems; to be thorough, it is necessary to indicate that these difficulties exist. In this section, brief consideration is given to whether latent structure models might be useful in reducing the number of items on the test that would otherwise be needed.

Consider the case in which two items per skill are used. For this situation, it is necessary to assume that one of the parameters γ or ε is known, since otherwise the parameters are not uniquely determined and cannot be estimated. For present purposes assume that $\gamma=0$ when estimating ε.

As a comparison with the results on using the correction for guessing formula score, monte carlo methods were used to estimate the $P(CD)$ using the values of n reported in Table 1 for the case $m=4$. The total number of skills on the test was set at $n/2$ or $(n+1)/2$, whichever gave an integral result. In each case ξ_0 was set to .8, the estimates of the $P(CD)$ were made with $\xi=\xi_0-\delta^*=.7$, and then with $\xi=\xi_0+\delta^*=.9$. The method used to estimate ξ has been described by Wilcox (1979e). The results are reported in Table 2. As can be seen, the latent structure model performed satisfactorily for $\gamma=0$; but even for γ slightly larger than zero, the results do not support this approach when $\xi=.9$.

Table 2
Estimated Probability of a Correct Decision, P(CD), Using Two Items per Skill, $\xi_0=.8$

			P(CD)	
γ	ε	n	$\xi=.7$	$\xi=.9$
0	.15	18	.86	.88
0	.30	33	.92	.94
0	.40	80	.94	.93
.02	.15	27	.92	.75
.02	.30	26	.88	.78
.02	.40	57	.98	.71
.05	.15	210	.99	.23
.07	.15	150	.99	.04

Next, the use of three items per skill was considered, for a total of 30 skills, and hence 90 items. In this case, the iterative procedure described by Goodman (1979) was used to approximate the maximum likelihood estimates of the parameters of the model. (It was no longer assumed that $\gamma=0$.)

For the specific examinee being tested, let p_{ijk} be the probability of a particular pattern of responses on a randomly sampled triplet of equivalent items where a subscript of 0 or 1 corresponds to an incorrect and correct response, respectively. For example, p_{011} denotes the probability of an incorrect response on the first item and a correct response on the other two. When applying Goodman's estimation procedure, the p_{ijk}'s that are the cell probabilities of a multinomial distribution which must be estimated. Two estimation procedures were used. The first was the usual sample mean; the other was an estimate proposed by Fienberg and Holland (1973, Equation 2.13).

It should be noted that when the sample mean is used to estimate the p_{ijk}'s, the solution to Goodman's Equations 7, 8a,. . . ,8d are maximum likelihood estimates of the parameters in the latent structure model. However, when the Fienberg-Holland estimate of the p_{ijk}'s are used, maximum likelihood estimates are no longer being obtained.

The results of the monte carlo studies are reported in Table 3. The columns headed MLE are the values of the $P(CD)$ using Goodman's estimation procedure. The columns headed FH are the modified estimates based on the Fienberg-Holland estimate of the cell probabilities of a multinomial distribution. All indications are that the $P(CD)$ is at least .9 when $\varepsilon=0$ or .15, even for $\gamma=.05$ or .07. For all previous approaches the $P(CD)$ was considerably below .9 for $\gamma=.05$ and .07. However, for $\varepsilon=.3$, and particularly for $\varepsilon=.4$, the $P(CD)$ is not very large. Note that the Fienberg-Holland estimate of the parameters in the multinomial distribution nearly always yielded better results than those obtained with the sample mean estimate of the p_{ijk}'s.

Table 3
Estimated Probability of a Correct Decision Using Three Items Per Skill for a Total of 30 Skills, $\xi_0=.8$, $\delta^*=.1$

		MLE		FH	
γ	ε	$\xi=.7$	$\xi=.9$	$\xi=.7$	$\xi=.9$
0	0	.97	.85	.92	.88
0	.15	.92	.78	.91	.93
0	.30	.78	.74	.69	.93
0	.40	.55	.68	.65	.95
.02	0	.94	.92	.92	.90
.02	.15	.86	.89	.82	.92
.02	.30	.76	.73	.79	.94
.02	.40	.54	.79	.69	.97
.05	.15	.94	.92	.82	.90
.07	.15	.91	.90	.92	.88

No generalizations should be drawn from the monte carlo results reported in this section. The goal was a more modest one, namely, to suggest what results might be obtained with latent structure models. The point is that it might be possible to take into account the errors γ and ε when comparing ξ to ξ_0 with a realistic, though perhaps large, number of items on the test. The results reported here

are intended to motivate a more extensive investigation of the application of these models as well as of the modified estimation procedure described above.

Solutions in Terms of ξ and a Population of Examinees

As was the case with the proportion correct true score, it is possible to formulate the test length problem in terms of ξ, the proportion of skills known by an examinee, where now concern is with the typical examinee among a population of examinees being tested. Note that when determining mastery of a single skill, rather than for a domain of skills, the solution described by Macready and Dayton (1977) might be applied.

For the special case $\varepsilon>0$ and $\gamma=0$ (or $\varepsilon=0$ and $\gamma>0$) the model proposed by Wilcox (1979e) might be used to obtain a point estimate of the probability of a false-positive or false-negative decision on the test. As before, if these two errors are judged to be too high, the length of the test might be increased. Wilcox (1979e) has illustrated how this might be done for the case of a single item per skill, and so the details of the procedure will not be discussed here.

It is important to realize that in certain circumstances, the case of $k>1$ items per skill can be accomodated. Suppose, for example, that for each skill there are k items and that in each case the same decision rule (for instance, all k items correct) is used to determine mastery. For a specific examinee and a sample of t skills (for a total of $n=tk$ items), the probability of x mastery decisions is

$$\binom{t}{x} p^x (1-p)^{t-x} \qquad [36]$$

where p is the unknown probability of a mastery decision for a randomly sampled skill. Note that for this special case, $p=\xi+(1-\xi)\varepsilon_1\varepsilon_2\ldots\varepsilon_k$, where ε_i is the probability of guessing the i^{th} item used to measure the skill. Let $\phi=(1-\xi)\varepsilon_1\ldots\varepsilon_k$ and assume that ξ and ϕ arise from a bivariate Dirichlet distribution. In other words, it is assumed that the joint probability density function of ξ and ϕ over the population of examinees is given by

$$\frac{\Gamma(v_1+v_2+v_3)}{\Gamma(v_1)\Gamma(v_2)\Gamma(v_3)} \xi^{v_1-1} \phi^{v_2-1} (1-\xi-\phi)^{v_3-1} \qquad [37]$$

where the v_i $(i=1,2,3)$ are unknown parameters.

If t, the number of skills, is sufficiently large, as might be the case in a preliminary investigation, the ξ and the ε_i's can be estimated for each examinee. (The effect of having a small number of skills appears to be unknown, cf. Wilcox, 1980b). Once ξ and ϕ have been estimated for a random sample of examinees, v_1, v_2 and v_3 can be estimated in the manner described by Wilcox (1979e). Substituting these estimates into the right-hand side of

$$f(x,\xi)=\binom{n}{x} B^{-1}(v_1,v_2,v_3) \times$$

$$\sum_{w=0}^{x} \binom{x}{w} B(w+v_2,\ n-x+v_3)\xi^{x-w+v_1-1}(1-\xi)^{n-x+w+v_2+v_3-1} \qquad [38]$$

where $B(.,.,.)$ is the usual beta function, yields an estimate of the joint probability density function of x and ξ. If x_0 is the passing score of the test, the two possible errors are simply

$$\sum_{x=0}^{x_0-1} \int_{\xi_0}^{1} f(x,\xi)\, d\xi \qquad [39]$$

and

$$\sum_{x=x_0}^{n} \int_{0}^{\xi_0} f(x,\xi)\, d\xi \qquad [40]$$

which can be evaluated with subroutine BDTR in the IBM (1971) scientific subroutine package. Thus, the test length $n=tk$ can be determined by adjusting t (and x_0) until Equations 39 and 40 are sufficiently small.

Bounds on the Probability of an Error

If γ, the conditional probability that an examinee gives an incorrect response, given that he/she knows the skill, is greater than zero and if ξ is estimated with a latent structure model, it is no longer clear how to estimate the probability of a false-positive and false-negative decision. However, if $\hat{\xi}$ is any estimate of ξ, the $P(\hat{\xi}<\xi_0)$ can be estimated for a randomly selected examinee. This estimate is simply the proportion of examinees for whom $\hat{\xi}<\xi_0$. Moreover, if $\hat{\xi}$ is consistent, the mean and variance of ξ over the population of examinees can be estimated (Wilcox, 1979e). Thus, following Wilcox (1979c), bounds on the two error types can be estimated. As previously explained, these bounds provide information about whether there are enough items on the test.

Solutions Using Latent Trait Models

The third general approach to the test length problem is based on a latent trait model. For a specific examinee the probability of a correct response to a test item, say $p_\tau(\zeta)$, is viewed as a function of ζ, $-\infty < \zeta < \infty$, the examinee's "ability" level, and the vector τ, which consists of parameters that characterize the item. Lord (1974) interprets $p_\tau(\zeta)$ as a relative frequency over randomly selected test questions, all having the same vector of τ values.

Several forms for $p_\tau(\zeta)$ have been proposed (see, e.g., Hambleton & Cook, 1977). For a recent review of latent trait models, the reader is referred to Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978). Familiarity with this review is assumed henceforth.

Birnbaum (1968) considers the classification of examinees in some detail. As was the case with the two types of true score previously considered, it is assumed that two ability levels have been specified, say ζ_1 and ζ_2, with the idea that if $\zeta \leqslant \zeta_1$ or if $\zeta \geqslant \zeta_2$, it is desired that there be a reasonable certainty of making a correct decision about whether the examinee's true score is large or small.

Rather than formulate the test length problem in terms of the probability of a correct decision, Birnbaum chooses n so that the probabilities associated with the two possible errors do not exceed prespecified values. That is, n is chosen so that simultaneously

$$P(x \geq x_0 \mid \zeta = \zeta_1) \leq \alpha^* \qquad [41]$$

and

$$P(x < x_0 \mid \zeta = \zeta_2) \leq \beta^* \qquad [42]$$

where x_0 is the passing score and α^* and β^* are preassigned constants. Note that x need not be a number-correct score. Birnbaum (1968, Equation 19.5.13) derives an approximation to the minimal n satisfying Equations 41 and 42 given by

$$n^{1/2} = \left[p_\beta(\zeta_2) - p_\beta(\zeta_1)\right]^{-1} \times$$

$$\left[\Phi^{-1}(1-\alpha^*)[p_\beta(\zeta_1)(1-p_\beta(\zeta_1))]^{1/2}\right.$$

$$\left. - \Phi^{-1}(\beta^*)[p_\beta(\zeta_2)(1-p_\beta(\zeta_2))]^{1/2}\right] \qquad [43]$$

where Φ^{-1} is the inverse function of the standard normal cumulative distribution. (Actually this expression for n differs slightly from Birnbaum's, which apparently has a typographical error.) To apply this solution, an estimate of the function $p_\tau(\xi)$ must already be available. The solution also makes the highly restrictive assumption of equivalent items, i.e., every item has the same values for τ.

As with all the probability models in this paper that attempt to make inferences about what an examinee knows beyond the observed responses on a test, there are several technical issues that remain to be resolved, not the least of which is a guideline on when one form of $p_\tau(\xi)$ is to be preferred over another. Some of these issues are discussed by Hambleton, Swaminathan, Cook, Eignor, and Gifford (1978).

Certainly, latent trait models deserve careful study and consideration. When measuring achievement, there are at least two issues that deserve a special comment. The first, which has been raised by Baker (1977), is whether latent trait models are even appropriate at all. As Baker has stated, latent trait theory is the culmination of the work on the measurement of ability begun by Binet that was the major focus of psychometrics in the 1920s, 1930s and 1940s, but the educational problems of an earlier era are not the problems of the 1970s and 1980s. The major trend in educational measurement today is one of instructionally related testing. Moreover, the problems arising from the individualization of instruction are very different from those of ability measurement.

A more specific problem with latent trait models that needs to be considered is what to do with the items that do not fit the model. From comments made by Gustafsson (1979), it would seem that many such items might exist when the Rasch model is assumed to hold. If these items really do represent a skill associated with a particular instructional program, it may be of interest to determine whether an examinee has mastered the skill, even if the item does not fit a particular latent trait model. Hambleton pointed out that this problem should be addressed at the test development stage; if there is evidence that the items measure the objectives of interest, and if the model does not fit, the model should be thrown out, not the items.

In terms of the present paper, the question is if one chooses not to ignore the items that do not fit a latent trait model, what should be done with these items, and how should our actions be related to the problem of test length? It should be mentioned, however, that under certain circumstances, an argument has been made in favor of latent trait models over item-sampling models (Wood, 1976). In addition, Messick (1975, p. 957) has argued that all measurement should be construct referenced and that a measure should estimate how much of a trait an individual possesses. Nevertheless, the issues of what to do, if anything, with items that do not fit a latent trait model has yet to be satisfactorily resolved.

Conclusions

Perhaps the most important point of this paper is that there is no magic number or even magic formula for determining test length. Even within the seemingly narrow problem of comparing an examinee's true score to some constant, there are many approaches to the problem. Moreover, in terms of which true score to use, it is not at all clear as to what extent the three types considered here are in competition with one another. For the moment, the best that can be done is to be very precise about what we want to determine, consider what assumptions we are willing to make, and act accordingly.

References

Aitchison, J., & Dunsmore, I. R. *Statistical prediction analysis*. London: Cambridge University Press, 1975.

Anderson, T. W. On estimation of parameters in latent structure analysis. *Psychometrika*, 1954, *19*, 1-10.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. *Robust estimates of location*. Princeton, NJ: Princeton University Press, 1972.

Baker, F. B. Advances in item analysis. *Review of Educational Research*, 1977, *47*, 151-178.

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

Bartlett, M. S. Untitled comment on "The estimation of many parameters" by D. V. Lindley. In V. P. Godambe & D. A. Sprott, *Foundations of statistical inference*. Toronto: Holt, Rinehart, & Winston, 1971.

Bechhofer, R. E. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, 1954, *25*, 16-39.

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.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

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

Cochran, W. G. Some methods for strengthening the common χ^2 tests. *Biometrica*, 1954, *10*, 417-451.

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

Dempster, A. P. New approaches for reasoning towards posterior distributions based on sample data. *Annals of Mathematical Statistics*, 1966, *37*, 355-374.

Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 1967, *36*, 325-339.

Dempster, A. P. Upper and lower probabilities generated by a random closed interval. *Annals of Mathematical Statistics*, 1968, *39*, 957-966. (a)

Dempster, A. P. A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Ser. B*, 1968, *30*, 205-232. (b)

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

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

Fienberg, S. E., & Holland, P. W. Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, 1973, *68*, 683-691.

Freeman, M. F., & Tukey, J. W. Transformations related to the angular and the square root. *The Annals of Mathematical Statistics*, 1950, *21*, 607-611.

Gelfand, A., & Thomas, D. Discrimination between the binomial and hypergeometric models. *Communications in Statistics A, Theory and Methods*, 1976, *18*, 225-240.

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

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

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications II: Further discussion and references. *Journal of the American Statistical Association*, 1959, *54*, 123-163.

Gustafsson, J. *Testing and obtaining fit of data to the Rasch model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Hambleton, R., & Cook, L. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement*, 1977, *14*, 75-96.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1-47.

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

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alken, & W. James Popham (Eds.), *Problems in criterion-referenced measurement* (Center for the Study of Evaluation Monograph No. 3). Los Angeles: 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*, 285-304.

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

Huang, W. Bayes approach to a problem in partitioning k normal populations. *Bulletin of the Institute of Mathematics Academia Sinica*, 1975, *3*, 87-97.

Huynh, H. Statistical consideration of mastery scores. *Psychometrika*, 1976, *41*, 65-78.

Huynh, H. Statistical inference for false positive and false negative error rates in mastery testing. *Psychometrika*, 1980, *45*, 107-120.

IBM Application Program, System/360. *Scientific subroutines package (360-CM-03X). Version III. Programmer's manual.* White Plains, NY: IBM Corporation Technical Publications Department, 1971.

Katz, L. Unified treatment of a broad class of discrete probability distributions. In G. P. Patil (Ed.), *Classical and contagious discrete distributions*. New York: Pergamon Press, 1963.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. *Psychometrika*, 1962, *27*, 59-72.

Kendall, M. G., & Stuart, A. *The advanced theory of statistics* (Vol. 2). New York: Hafner, 1973.

Lam, K., & Chiu, W. K. On the probability of correctly selecting the best of several normal populations. *Biometrika*, 1976, *63*, 410-411.

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

Livingston, S. A., & Wingersky, M. S. Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, 1979, *16*, 247-260.

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

Lord, F. M. Estimating true-score distributions in psychological testing (An empirical Bayes estimation problem). *Psychometrika*, 1969, *34*, 259-299.

Lord, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2). San Francisco, CA: Freeman, 1974.

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.

Maritz, J. S. *Empirical Bayes methods*. London: Methuen, 1970.

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

McHugh, R. B. Efficient estimation and local identification in latent class analysis. *Psychometrika*, 1956, *21*, 331-347.

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

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

Molenaar, W. On Bayesian formula scores for random guessing in multiple choice tests. *British Journal of Mathematical and Statistical Psychology*, 1977, *30*, 70-89.

Morgan, G. *A criterion-referenced measurement model with corrections for guessing and carelessness* (Occasional Paper No. 13). Victoria: The Australian Council for Educational Research Limited, 1979.

Murray, G. D. A note on the estimation of probability density functions. *Biometrika*, 1977, *64*, 150-151.

Novick, M. R. High school attainment: An example of a computer-assisted Bayesian approach to data analysis. *International Statistical Review*, 1973, *41*, 264-271.

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in *m* groups. *Psychometrika*, 1973, *38*, 19-46.

Reulecke, W. A. A statistical analysis of deterministic theories. In H. Spada & F. Kempf (Eds.), *Structural models of thinking and learning*. Bern: Huber, 1977.

Rustagi, J. S. *Variational methods of statistics*. New York: Academic Press, 1976.

Skibinsky, M. Sharp upper bounds for probability on an interval when the first three moments are known. *The Annals of Statistics*, 1976, *4*, 187-213.

Skibinsky, M. The maximum probability on an interval when the mean and variance are known. *Sankhya, Series A*, 1977, *39*, 144-159.

Springer, M. D. *The algebra of random variables*. New York: Wiley, 1979.

Tarone, R. E. Testing the goodness of fit of the binomial distribution. *Biometrika*, 1979, *66*, 585-590.

Tong, Y. L., & Wetzell, D. E. On the behaviour of the probability function for selecting the best normal population. *Biometrika*, 1979, *66*, 174-176.

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

van der Linden, W. Forgetting, guessing, and mastery: The Macready and Dayton models revisited and compared with a latent trait approach. *Journal of Educational Statistics*, 1978, *3*, 305-317.

von Mises, R. *A mathematical theory of probability and statistics*. New York: Academic Press, 1964.

Weitzman, R. A. Ideal multiple-choice items. *Journal of the American Statistical Association*, 1970, *65*, 71-89.

Wilcox, R. R. Estimating the likelihood of a false-positive or false-negative decision with a mastery test: An empirical Bayes approach. *Journal of Educational Statistics*, 1977, *2*, 289-307.

Wilcox, R. R. Applying ranking and selection techniques to determine the length of a mastery test. *Educational and Psychological Measurement*, 1979, *39*, 13-22. (a)

Wilcox, R. R. Comparing examinees to a control. *Psychometrika*, 1979, *44*, 55-68. (b)

Wilcox, R. R. On false-positive and false-negative decisions with a mastery test. *Journal of Educational Statistics*, 1979, *4*, 59-73. (c)

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

Wilcox, R. R. Estimating the parameters of the beta-binomial distribution. *Educational and Psychological Measurement*, 1979, *39*, 527-535. (e)

Wilcox, R. R. Some results and comments on using latent structure models to measure achievement. *Educational and Psychological Measurement*, 1980, in press. (a)

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

Wood, R. Trait measurement and item banks. In D. de Gruijter & L. van der Kamp (Eds.), *Advances in Psychological and Educational Measurement*. New York: Wiley, 1976.

Author's Address

Rand R. Wilcox, Center for The Study of Evaluation, 145 Moore Hall, University of California, Los Angeles, CA 90024.