DOCUMENT RESUME

ED 235 206 TM 830 620

AUTHOR	Jarjoura, David
TITLE	Confidence and Tolerance Intervals for True Scores. ACT Technical Bulletin, Number 42.
INSTITUTION	American Coll. Testing Program, Iowa City, Iowa. Research and Development Div.
PUB DATE	Jul 83
NOTE	63p.
AVAILABLE FROM	Research and Development Division, The American College Testing Program, P.O. Box 168, Iowa City, IA 52243.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Educational Testing; Measurement Techniques; *Models; *Scores; Statistical Analysis; Test Interpretation
IDENTIFIERS	*Confidence Intervals (Statistics); *Tolerance Intervals (Statistics)

ABSTRACT

Issues regarding confidence and tolerance intervals are discussed within the context of educational measurement. Conceptual distinctions are drawn between these two types of intervals; and examples, under various error and true score models, are used to compare such intervals. It is shown that there tend to be only small differences in tolerance intervals under different true score models. It is also demonstrated that confidence and tolerance intervals are not only quite distinct conceptually, but also can be very different numerically. Points are raised about the usefulness of tolerance intervals when the focus is on a particular observed score rather than a particular examinee. (Author)

ACT Technical Bulletin Number 42

Confidence and Tolerance Intervals for True Scores

David Jarjoura
The American College Testing Program
July 1983

A principal purpose of the ACT Technical Bulletin Series is to provide timely reports of the results of measurement research at ACT. Comments concerning technical bulletins are solicited from ACT staff, ACT's clients, and the professional community at large. A technical bulletin should not be quoted without permission of the author(s). Each technical bulletin is automatically superseded upon formal publication of its contents.

Research and Development Division
The American College Testing Program
P.O. Box 168
Iowa City, Iowa 52243

Table of Contents

Page

List of Tables

Abstract

Issues regarding confidence and tolerance intervals are discussed within the context of educational measurement. Conceptual distinctions are drawn between these two types of intervals; and examples, under various error and true score models, are used to compare such intervals. It is shown that there tends to be only small differences in tolerance intervals under different true score models. It is also demonstrated that confidence and tolerance intervals are not only quite distinct conceptually, but also can be very different numerically. Points are raised about the usefulness of tolerance intervals when the focus is on a particular observed score rather than a particular examinee.

Introduction

Through the use of confidence intervals for true scores, one can discourage interpretations of observed test scores that are too literal. Such an interval also provides a gauge for the potential error associated with a measurement procedure. This paper discusses confidence intervals within the context of educational measurement, and contrasts them, conceptually and through numerical examples, against tolerance intervals. A major portion of the paper compares tolerance intervals that are based on various true score models.

Some fundamental issues regarding true score confidence intervals are discussed here so that distinctions can be drawn between various interpretations of these intervals, and so that clear contrasts can be made with true score tolerance intervals. Tolerance intervals, as such, have not been previously suggested for true scores, although intervals with the same or similar form have appeared in both the early and recent literature. For example, intervals around the familiar regressed score estimates can be viewed as tolerance intervals under certain assumptions. Also, true score tolerance intervals can resemble Bayesian credibility intervals; but because true score tolerance intervals fall within the framework of the classical regression model, the two approaches are quite distinct conceptually.

Generally, confidence interval procedures are designed to cover, with a chosen probability, the value of a parameter. It is often emphasized that a realized interval, i.e., one that is based on a particular set of observations or realized sample, either does or does not cover the value of a parameter, and the interpretation of a realized confidence interval must be in terms of the procedure on which it is based. An interpretation that is often suggested is that a confidence interval procedure will, over repeated applications, cover a parameter a chosen proportion of the time.

In a measurement context, a realized confidence interval for a particular examinee is often based on the observed score obtained by that examinee and a standard error of measurement that is estimated from a large sample of examinees. Typically, more than one observed test score is not available for a particular examinee, but we can interpret a confidence interval procedure for that examinee in terms of his/her hypothetical distribution of observable scores. The mean or expected value of this distribution is the parameter of interest; i.e., his/her true score is the parameter to be covered by a confidence interval procedure.

The assumption that the standard error of measurement--the standard deviation of the hypothetical distribution--is the same for all examinees justifies the use of a single estimate of this standard error for constructing confidence intervals across examinees. But a weaker claim could be made about the overall confidence interval procedure which does not depend on this assumption. Instead of claiming that a confidence interval procedure covers a particular examinee's true score with a chosen probability, it might be claimed that "on average" such a procedure covers the true scores of a population of examinees a chosen proportion of the time. This average probability is taken over the examinee population and allows for the possibility that a confidence interval procedure for a particular examinee does not have a coverage probability equal to the average probability across examinees. The average coverage claim is explored in this paper in order to determine the conditions that make it accurate.

The issue of average coverage of a confidence interval procedure raises other issues regarding interval estimation of true scores. In a measurement situation in which potentially many intervals are reported, it seems natural to describe the statistical properties of the overall procedure of setting intervals for some population of examinees, rather than restrict attention to the properties for an isolated examinee. Consider the typical situation in which all examinees with the same observed score receive the same interval. What seems of special interest is the probability of coverage of true scores for an interval based on a particular observed score. More precisely, we can ask: What is the proportion of the true score distribution, conditional on a particular observed score, that is covered by an interval based on that observed score? This is to be distinguished from the interpretation of a realized confidence interval based on a particular observed score, which must be in terms of the confidence interval procedure rather than the realized interval. In a measurement context, there is a distribution of true scores associated with a population of examinees. For this reason, we can interpret an interval based on a particular observed score in terms of the conditional (on that score) distribution of true scores rather than in terms of a particular examinee's true score. Thus, we can design an interval to cover some proportion of the conditional true score distribution.

An interval designed to cover some proportion of a distribution is usually referred to as a tolerance interval. Such intervals are the major focus here. Because tolerance intervals for conditional true score distributions require a "strong" true score model (an explicit specification of the joint distribution of observed and true scores), four such models are used for comparing the intervals they produce. The comparisons, which comprise a major portion of the paper, are based on a variety of test characteristics adapted from standardized tests. Similarly, confidence intervals from three error models are compared, and are then contrasted with tolerance intervals.

Confidence Intervals for True Scores

Considered in isolation, the process of making an inference about a particular examinee's true score suggests that a confidence statement can be a useful part of the process. When we focus on examinee a, we are interested in a parameter τ_a--the true score of that examinee. With τ_a defined as the mean of observable scores, X_a, for that examinee, it seems natural to attempt to acquire information about the distribution of X_a. And, a confidence interval procedure seems to be a succinct method for expressing such information. For example, the confidence statement $P[L(X_a) \leq \tau_a \leq U(X_a)] = 1 - \alpha$; where L and U are variables dependent on the random variable X_a (and possibly other random variables) and $1 - \alpha$ is the confidence coefficient or the probability that potential intervals cover τ_a, can provide information about the distribution of X_a and the accuracy with which X_a measures τ_a.

Obtaining enough information to feel comfortable in making such statements might require several observations on examinee a. But in a measurement context, certain factors usually preclude such an approach. Because of the difficulty in obtaining several observations on full test forms, and potential problems with practice, fatigue, motivation, etc., more than two observations on any examinee are rarely obtained. Instead, properties of the overall measurement procedure, based on a population of examinees, are used to estimate conficence intervals for each examinee. Strong assumptions could be used to justify a conficence interval procedure for a particular examinee. For example, the error variable for examinee a, $e_a = X_a - \tau_a$, could be assumed normal with the same variance, σ_e^2, for all examinees. An accurate estimate of σ_e^2 could then be obtained, say, through the administration of two parallel forms to a large sample of examinees. This would allow a confidence statement of the form $P(X_a - c\hat{\sigma}_e \leq \tau_a \leq X_a + c\hat{\sigma}_e) = 1 - \alpha$ to be used for examinee a. The c could be determined from a z or t table depending on the sample size for $\hat{\sigma}_e$.

Still stronger assumptions might be used so that even estimation of σ_e^2 is avoided. For example, with number correct scoring, the binomial error model (Lord & Novick, 1968, chaps. 11 & 23) is sometimes viewed as appropriate. If this model holds for an examinee, we can simply use the examinee's observed score and the number of items in a test to enter a table of confidence intervals for a binomial parameter. This would provide a confidence interval for the examinee's proportion correct true score.

A Weak Claim About Confidence Intervals

Such strong assumptions allow the strong claim that is made about a confidence interval procedure for a particular examinee. However, if such assumptions are unwarranted, one could still make a weaker claim about the confidence interval procedure. For example, it might be claimed that, on average, intervals of the general form $X \pm z_{\alpha/2}\sigma_e$ cover examinees' true scores with probability $1 - \alpha$; where the average is taken over the population of examinees for which such intervals are reported, and where $z_{\alpha/2}$ refers to the $1 - \alpha/2$ cumulative percentage point of the standard normal distribution. In this case, the confidence interval procedure for a particular examinee can be associated with a coverage probability that is greater or less than $1 - \alpha$, but the average across examinees is $1 - \alpha$.

This average coverage is expressed as

$$E[P(|X_a - \tau_a| \le z_{\alpha/2}\sigma_e)] = 1 - \alpha \quad , \tag{1}$$

where E is the expectation operator over examinees, the probability statement is the coverage probability for examinee a of $X_a \pm z_{\alpha/2}\sigma_e$, and σ_e^2 is the average measurement error variance for the population of examinees. One way of writing this in integral form is

$$N^{-1}\sum_a^N \int_{-z_{\alpha/2}\sigma_e}^{z_{\alpha/2}\sigma_e} dF_a(e) = 1 - \alpha \quad , \tag{2}$$

where $F_a(e)$ is the cumulative distribution function of e_a and the integral is in Stieltjes form to allow for discrete e_a (both end points are included in the integration). The summation is over the N examinees in the population for which the average coverage claim is made. We can switch the order of integration in Equation 2, so that

$$\int_{-z_{\alpha/2}\sigma_e}^{z_{\alpha/2}\sigma_e} N^{-1} \sum_{a}^{N} dF_a(e) = 1 - \alpha \; . \tag{3}$$

Now the limits ± $z_{\alpha/2}\,\sigma_e$ can be viewed as points on a mixture of the distributions of the error variables, or the marginal distribution of error. Thus, the average coverage claim simply states that the area in the marginal distribution between the two points ± $z_{\alpha/2}\,\sigma_e$ is $1 - \alpha$.

As defined, the mean of e_a for each examinee is zero, so these two points are equidistant from the mean. Of course, in order for this average coverage claim to hold at every value of $z_{\alpha/2}$, normality of the marginal error distribution is necessary. However, if we only make the following claim, "$X \pm 1.96\,\sigma_e$ has an average coverage of .95, approximately," many distribution shapes will do.

If we just assume that the marginal distribution of error has one mode, we can use the Camp-Meidell inequality (Rao, 1973, p. 145) which states that

$$P(|X - \mu| \geq \lambda\sigma) \leq \frac{4(1 + s^2)}{9(\lambda - s)^2} \; ,$$

where μ is the mean, σ is the standard deviation, s is the absolute value of the number of standard deviation units that μ is from the mode, and $\lambda > s$. Say, for example, $s = .2$, then average coverage of $X \pm 1.96\,\sigma_e$ is greater than .85 for any uni-modal distribution. Thus, with an accurate estimate of $\bar{\sigma}^2_e$ (the average of individual error variances) and $z_{\alpha/2}$ around 1.96, one might feel comfortable in making an average coverage claim around .9.

Clearly, an average coverage claim is fairly weak. It describes a property of the overall interval estimation procedure in a measurement context, but does little to describe the limitation of the information from X_a about examinee a's true score. Nor does it make any claims about what to expect at different score points. If evidence is available indicating that error variance differs along the score scale, then an average coverage claim seems especially uninformative. However, consideration of such differences when constructing intervals could allow a claim of average coverage for different ranges of observed scores.

Coverage of True Scores Conditional on Observed Score: Tolerance Intervals

That an average coverage claim over different parameters (i.e., true scores) is sensible in a measurement context raises the question of whether we are always interested in an interval estimate for a particular examinee. Under circumstances in which a particular examinee's score is being interpreted by a career counselor or classroom teacher, a proper confidence interval seems quite useful in combination with other information about that examinee. In contrast, the process of score reporting, in which large numbers of examinees are given the same score and interval, is not intimately concerned with a particular examinee. Rather, there is a distribution of true scores that is referenced by a particular observed score. Thus, in a measurement context we can interpret an observed score in terms of the conditional distribution of true scores associated with it.

In a typical situation in which the group of examinees with the same observed score receives the same interval estimate, it seems natural to inquire about the proportion of the distribution of true scores given an observed score that is covered by the interval. In other words, for an observed score x (realized value of the X variable that represents observable scores for the population of examinees), there is a proportion of the conditional distribution of true scores that is covered by an interval like $x \pm c\,\sigma_e$. This proportion, which is a conditional (on x) probability, is conceptually distinct from a confidence coefficient. If we condition on an observed score, then a confidence interval does or does not cover a particular examinee's true score; i.e., the conditional probability is not in reference to any particular examinee. Later, intervals of the form $x \pm c\,\sigma_e$ are evaluated in terms of the conditional distribution of true scores.

Tolerance Intervals

Probability statements about conditional true score distributions require strong assumptions or data that are usually not available. In particular, the joint distribution of observed and true scores is needed. For expository purposes, we assume that the distribution of error conditional on true score is normal with mean zero and a variance that is constant across true scores. In addition, true score is assumed normal. Thus, X is the sum of two independent and normal variables τ and e, where τ is normal(μ, σ_τ^2), and e is normal(0, σ_e^2) .

Under this model it is well-known that the conditional distribution of τ given X = x is normal($\rho^2 x + (1 - \rho^2)\mu$, $\rho^2\sigma_e^2$), where $\rho^2 \equiv \text{CORR}(X, \tau)^2 = 1 - \sigma_e^2/\sigma_X^2$. We will refer to the mean of the conditional distribution as

$$\tau(x) \;=\; \rho^2 x + (1 - \rho^2)\mu \;, \tag{4}$$

which is the familiar regressed score estimate of true score (see e.g., Lord & Novick, 1968, pp. 64-69).

Since the conditional distribution of τ given x is normal,

$$\tau(x) \pm z_{\alpha/2}\rho\sigma_e \tag{5}$$

is an interval which covers the central 100(1 - α)% of the conditional true score distribution associated with x (this holds for all values of X). It is referred to as central because both tails of the conditional distribution, not covered by the interval, contain 100*α/2% of the true scores. A central interval, in the case of the normal, is also the shortest interval that covers 100(1 - α)% of the conditional distribution.

Such intervals are quite distinct from confidence intervals. As noted, a confidence interval procedure is designed to cover, with a chosen probability, some parameter of a distribution. In contrast, the above interval is designed to cover a chosen proportion of the distribution of a random variable. Intervals of this type are referred to as tolerance intervals. Proschan (1953) provides some basic comparisons between tolerance and confidence intervals.

Before discussing some issues regarding the estimation of tolerance intervals, some comparisons will be made between confidence and tolerance intervals in the context of measurement. Stanley (1971, pp. 379-382), among others, discusses an interval similar to that of Equation 5 and appropriately refers to it as a "confidence interval in a loose sense."[1] Also, a few of the following points have been touched on in the measurement literature, though from a different perspective.

[1] The difference between Equation 5 and his Equation 19 is his use of the t distribution instead of the z and his use of estimates of the parameters μ, ρ^2, and σ_e^2. Estimation issues for tolerance intervals are complex and have not been solved for the case where the variable of interest is unobservable. Thus, use of a t distribution just provides wider intervals than the z.

The justification given above for using a confidence interval in a measurement context is that a score interpretation situation calls for isolated interest in a particular examinee's true score. A mistaken interpretation of a reported or realized confidence interval based on a given score might then be that it provides a range of probable values for the true score of that examinee. Instead of considering a reported confidence interval as an indication of the accuracy with which an examinee's true score is estimated, its meaning is distorted to include consideration of the likely values of true scores in the population of examinees for a given score x.[2] Within the context of classical confidence interval estimation, such an interpretation makes little or no sense because, again, the value of a single parameter is of interest. Within a measurement context, however, there is a distribution of true scores of interest, so that such an interpretation may be desirable. But confidence intervals are not designed to provide such interpretations, and they would lead, at the least, to inaccuracies.

Consider again the model with normal and independent error and true scores (and no two examinees have the same true score). A confidence interval of the general form $X \pm z_{\alpha/2}\,\sigma_e$ would, for every examinee, have a confidence coefficient of $1 - \alpha$; i.e., the probability that an interval of this form covers an examinee's true score is $1 - \alpha$. In contrast, a reported confidence interval based on a realized value of X, $x \pm z_{\alpha/2}\,\sigma_e$, either covers an examinee's true score or does not. Now, by considering the population of examinees, this same

[2] It is true that Bayesian approaches allow isolated interest in an examinee's true score and a statement about probable values of that true score. However, the nature of probability changes, and, in any case, a classical confidence interval procedure would not be used, typically.

reported interval will, typically, cover more or less than 1 − α of the true scores that can be associated with x. The interval $x \pm z_{\alpha/2}\, \sigma_e$ covers somewhat more than 1 − α of the true scores associated with x, when x is close to μ, and less when it is far away. It is easily shown that the average proportion covered, taken across the variable X, is in fact 1 − α.

In order to determine, under this model, the proportion of the conditional true score distribution covered by a realized interval of the form $x \pm z_{\alpha/2}\, \sigma_e$, we need only specify the reliability (ρ^2) and the number of σ_X units x is from μ. Let us take $\rho^2 = .8$, and for simplification $z_{\alpha/2} = 1$ (i.e., $1 - \alpha = .68$). When $x = \mu$, the realized interval $x \pm \sigma_e$ covers the central 74% of the distribution of true scores associated with x. When $x = \mu + \sigma_x$ or $x = \mu - \sigma_x$, $x \pm \sigma_e$ covers 68% of the conditional true score distribution, but not the central 68%, i.e., the areas in the tails of the distribution n t covered by the interval are unequal. When $x = \mu + 2\sigma_x$ or $x = \mu - 2\sigma_x$, $x \pm \sigma_e$ covers only 53% of the conditional true score distribution--again not the central 53%. To see this, consider that, under the model, the area between $x - \sigma_e$ and $x + \sigma_e$ is being evaluated for the distribution of τ given x which has mean $\rho^2 x + (1 - \rho^2)\mu$ and variance $\rho^2 \sigma^2_e$. Thus, except when $x = \mu$, realized confidence intervals will be centered farther from μ than the mean (center) of the conditional true score distribution. In contrast, tolerance intervals are centered on this mean. Even though on average the proportion covered is 68% for this example, one-third of the confidence intervals will cover less than 68% of the conditional true score distributions. Thus for this example, at least, realized confidence intervals would be very misleading if interpreted as tolerance intervals.

Another related criticism against interpreting the interval $x \pm z_{\alpha/2}\, \sigma_e$ as a tolerance interval for x is that it can be considered more appropriate for observed scores that are farther from μ (more extreme) than x. This is because this interval covers a greater proportion of the conditional distributions of true scores for scores more extreme than x than it does for x itself. Again, this interval is not centered on the mean of the conditional true score distribution. Specifically, there is an observed score x* such that $P(x - z_{\alpha/2}\ \sigma_e \leq \tau \leq x + z_{\alpha/2}\ \sigma_e | X = x^*)$ takes on the largest value, and this x* is more distant from μ than is x. The value of this probability is also larger for all values between x* and x than it is for x, and the same holds for more extreme scores between 2x* - x and x*. The value of x* is $\mu + (x - \mu)/\rho^2$. To understand why the probability is largest under x*, consider that the conditional mean of τ given x* is x; i.e., x* makes the interval $x \pm z_{\alpha/2}\ \sigma_e$ centered around τ(x*). And, the values between 2x* - x and x are also associated with larger probabilities than x simply because their conditional means are closer to τ(x*) than is τ(x) .

These comparisons have been made within the context of the normal error-normal true score model. However, similar, though perhaps not as strong, statements could be made for other models. We can expect, for instance, that for most reasonable true score models, x will always be further from μ than the mean of τ given x.[3]

Two Perspectives on Tolerance Interval Estimation

When parameters of a distribution are not known precisely, estimation of tolerance intervals have been found to be fairly simple or quite complex depending on, among other things, the properties required of the estimator. There are two

[3] Consider the binomial error model and a true score distribution that is assumed uniform between 0 and 1. The mode of the conditional distribution of τ given x is then x/n (Novick & Jackson, 1974, p. 114). Since the highest density region converges on the mode, this provides a contrast to comments above. However, the uniform is an interesting prior but is unrealistic as an empirical distribution for true scores. Further, central tolerance intervals converge on the median rather than the mode.

alternative properties that are discussed. One is that the interval estimator cover on average the desired proportion of the distribution. The desired proportion is then referred to as the expected coverage. For the normal univariate case, Proschan (1953) provides such an estimator. The expected coverage requirement of tolerance intervals for the conditional distribution of true scores (given x) can be written as

$$E\left[\int_{L(x)}^{U(x)} f(\tau|x)d\tau\right] = 1 - \alpha \quad ;$$

where U(x) and L(x) represent upper and lower limits of the tolerance interval for given x. In the discussion above, $U(x) = \tau(x) + z_{\alpha/2}\,\sigma_e$ and $L(x) = \tau(x) - z_{\alpha/2}\,\sigma_e$. But without knowledge of μ, σ^2, and σ_e^2, U(x) and L(x) are random variables that depend on estimates of these three parameters. Thus, the expectation is over U(x) and L(x).

The other alternative property places a confidence statement on the proportion of the distribution covered by an estimator. It places a probability on the event that a tolerance interval estimator covers at least the desired proportion of the distribution. In terms of the conditional distribution of true score (given x) this can be written as

$$P\left[\int_{L(x)}^{U(x)} f(\tau|x)d\tau \geq 1 - \alpha\right] = \lambda \quad ;$$

where λ is the confidence coefficient. When the parameters of the conditional distribution are assumed known, $\lambda = 1$. Otherwise the probability depends on U(x) and L(x). As an example, U(x) and L(x), as estimators of tolerance limits, might be chosen so that the probability is .95 that the limit estimators cover at least 68% of the true score distribution associated with x.

Probability of coverage estimators receive more attention than expected coverage estimators, mainly because they provide a more informative statement about the behavior of an estimator. Some even define tolerance intervals only in terms of probability of coverage. Also, probability of coverage is more useful in a major application of tolerance intervals, namely quality control problems. It does, however, create greater complexities, and typically λ is close to 1 which produces wider intervals than expected coverage intervals. Wald and Wolfowitz (1946) first provided an approximation under normality for a probability of coverage estimator. Wallis (1951) solved the estimation problem for the linear regression model, which has some relevance to our problem. More current work has focused on simplifying methods and extensions to simultaneous intervals for the regression case (see, e.g., Lieberman & Miller, 1963).

Tolerance intervals are rarely discussed in statistical methods texts (Dixon & Massey, 1962, p. 199; and Graybill, 1976, pp. 270-275, are two exceptions). Instead, the related issue of prediction intervals is often discussed (see e.g., Graybill, 1976, pp. 267-270 for prediction intervals in the linear regression model). Such an interval is used to predict a range of probable values for some future observation or linear function of several observations. Note that a $1 - \alpha$ prediction interval for a single observation is the same as a tolerance interval with expected coverage of $1 - \alpha$ (Proschan, 1953). The key to the identity is that the distribution of a single future observation is the distribution for which a tolerance interval is desired.

Because none of the research on the estimation of tolerance intervals considers the case in which the variable of interest is unobservable, none of it is directly relevant to the problem at hand. Even prediction intervals for the linear regression model would not serve as an expected coverage interval for the normal error and true score model because the basic assumptions are quite different under the two models.

Comparison of Tolerance Intervals Under Four True Score Models

The focus of this section is on the comparison of tolerance intervals calculated under different measurement model assumptions. Since tolerance interval estimators have not been derived for these true score models, comparisons are made under the presumption that accurate estimates of model parameters are available. Essentially, all that is presumed is that large enough samples are available to accurately estimate the mean and variance of the observed scores. This is because two of the models need only these two parameters for calculating tolerance intervals; and the other two models need only one additional parameter that does not seem to play a substantial role in the intervals. It seems important to focus attention on a comparison of true score models before tolerance interval estimators are derived because of the strong and sometimes unwarranted assumptions associated with each. The effects of differences in assumptions on differences in tolerance intervals can facilitate not only an informed choice of a model for calculating intervals with large samples but also a choice for the derivation of estimators.

First, the four true score models are described. Equations for calculating intervals under these models are then provided. This is followed by detailed comparisons among tolerance intervals based on test characteristics that were adapted from standardized tests.

Description of the Models

For the comparison of tolerance intervals, the following measurement models were used: (1) the normal model discussed above in which the conditional distribution of observed score (given true score) is normal and the distribution of true score is normal (NORM); (2) the conditional distribution of observed score is binomial and the distribution of true score is beta (BETA); (3) the conditional distribution of observed score is binomial but an angular (variance stabilizing) transformation provides approximate normality, and yields a normal true score distribution (BINORM); and (4) the conditional distribution of observed score is compound-binomial but an angular transformation provides approximate normality, and yields a normal true score distribution (CONORM).

All four models have been discussed previously in the literature and except for the NORM model, they were developed for number correct scoring. Lord and Novick (1968, chap. 22) provide a discussion of the NORM model--especially normal and independent error. Although this model was not designed especially for number correct scoring, it is included because of its convenience, historical popularity, relation to the BINORM and CONORM models, and as a contrast with the other three models.

The BETA model is discussed in detail by Keats and Lord (1962) and subsequently in work concerned with mastery testing (see, e.g., Huynh, 1976). Although the beta-binomial combination is a mathematical convenience, and the resulting model depends on just two unknown population parameters, the fit to number correct observed score distributions is often impressive (see, e.g., Keats & Lord, 1962). Wilcox (1981) reviews competitors to this model and concludes that it frequently gives satisfactory results and that choosing a more complex model involving additional free parameters can be quite difficult. Robustness of a methodology based on this model has also been shown (Gross & Shulman, 1980).

The BINORM model was adopted from Bayesian treatments of estimating true scores from observed number correct scores (Jackson, 1972; Novick, Lewis, & Jackson, 1973; and Lewis, Wang, & Novick, 1975; also, Hambleton, Swaminathan, Algina, & Coulson, 1978, provide a convenient summary). In this treatment the conditional distribution of observed score given true score is assumed to be binomial and an angular transformation provides, approximately, a normal error variable with stable variance $(4n + 2)^{-1}$ across the true score range, where n is the number of items in a test. It is also assumed that the angular transformation yields, approximately, a normal true score variable (or prior in their

Bayesian treatments). This transformation results in an expansion of the true score scale at the extremes which makes the assumption of normality (unbounded tails) much less of a problem than under the proportion correct scale. In addition, the transformation can account for the skewness that often occurs with observed score distributions associated with a mean (proportion correct) that is not close to .5.

The BINORM model is similar to the BETA in that both begin with the binomial for the conditional distribution of observed score. The contrast in the assumptions about the distribution of true score enables examining the sensitivity of tolerance intervals to such assumptions.

The BINORM and CONORM tolerance intervals provide a comparison of a different nature. As in the BINORM model, the distribution of transformed true score is assumed normal for the CONORM. But under the CONORM model, the conditional distribution of observed scores is assumed compound binomial rather than binomial. The two-term approximation to the compound-binomial suggested by Lord (1965) simplifies considerations in the model. Noting that the conditional variance under the binomial is $n\tau(1 - \tau)$, the conditional variance under the two-term approximation is $(n - 2k)\tau(1 - \tau)$ where k is a parameter to be defined.[4] Thus, with $k > 0$, shorter intervals can be expected under this model, all other things being equal.

[4] From here on, τ can be interpreted as a particular true score or the random variable for true score, depending on the context.

The appeal of this approximation in our case is that it provides an alternative conditional distribution for bounded observed scores and that the overall error variance (across examinees) can be made to correspond (with an appropriate choice of k) to an estimate of average error variance obtained under weaker assumptions. Lord (1965) emphasizes the fact that k can be chosen so that average error variance corresponds to that which would be obtained by using a KR20 estimate of reliability.

The use of an angular transformation with Lord's two-term approximation was pre·iously suggested by Wilcox (1978). Because the conditional variance is $(n - 2k)\tau(1 - \tau)$, a variance stabilizing transformation that is appropriate for the binomial is applicable here also.

Calculation of Tolerance Intervals

As noted above, we presume that accurate estimates of population parameters are available. In other words, we take the liberty of providing details about calculating intervals given the parameters, while, at the same time, providing the estimation equations used for the example tests that follow. Under the BETA and BINORM models, estimation simply involves calculating a mean and variance of observed scores. Estimation for the NORM and CONORM models additionally involves the calculation of the variance of item difficulties. This holds for these two models because interest is restricted here to a KR20 estimate of reliability for the examples provided. More generally, other estimates of reliability could be used for these two models.

Note that all the intervals that are calculated are for a proportion correct true score scale. The observed number correct score is still referred to as x for a particular score and X for the random variable. Thus, a true score for an examinee is defined as the expected number correct score divided by the number of items (n).

The expression for tolerance intervals under the NORM model has been given in Equation 5. However, there are slight changes in the expression because of the change in the true score scale. The conditional distribution of τ given $X = x$ is normal$[\tau(x), \rho^2_{\tau X}\ \sigma^2_e]$, where $\tau(x) = \mu + \rho^2_{\tau X}(x/n - \mu)$, $\mu = E\ X/n$, $\rho^2_{\tau X} = \mathrm{CORR}(\tau, X)^2$, and $\sigma^2_e = (1 - \rho^2_{\tau X})\sigma^2_{X/n}$. Thus, the lower limit of a tolerance interval on the proportion correct scale (the 100*α/2% point of the conditional true score distribution) is

$$\tau(x) \quad - \quad z_{\alpha/2}\ \rho_{\tau X}\ \sigma_e \quad ; \tag{6}$$

and for the upper limit (100*[1 − α/2]% point) a plus replaces the minus.

For the example data used here, it seemed appropriate to estimate $\rho^2_{\tau X}$ by KR20. Note that this means error variance is a function of n, as is the case for the other three models. It also means that the tolerance intervals can be expressed in terms of just three population parameters: $E\ X/n$, $\sigma^2_{X/n}$, and a parameter for the variance of item difficulty. This last parameter will be discussed later under the CONORM model.

Under the BETA, the conditional distribution of X given τ is binomial, and the distribution of τ is beta with population parameters a and b. This makes the conditional distribution of τ given x beta$(a + x, b + n - x)$. Using the usual notation for the cumulative distribution function of a beta, the lower limit is calculated by solving

$$I_L(a + x,\ b + n - x) \quad = \quad \alpha/2 \tag{7}$$

for L, where L is the point below which $100*\alpha/2\%$ of a beta $(a + x, b + n - x)$ falls. Similarly, the upper limit can be determined by solving

$$I_U(a + x, b + n - x) = 1 - \alpha/2 \tag{8}$$

for U. The inverse beta function subroutine MDBETI of IMSL (1979) was used for the calculations.

Note that this choice of L and U provide central tolerance intervals. Because the BETA is asymmetric when its parameters are unequal, shorter $1 - \alpha$ intervals could be found than central intervals. However, interpretations of L and U would then vary from one x to the next; i.e., the tail areas beyond each limit would change.

Under the BETA model, convenient estimates of the true score distribution parameters are:

$$\hat{a} = n(1/KR21 - 1)\hat{\mu} \quad , \tag{9}$$

$$\hat{b} = n(1/KR21 - 1) - \hat{a} \quad , \tag{10}$$

where $\hat{\mu}$ is a mean (proportion correct) observed score (see Lord & Novick, 1968, pp. 516-517 and pp. 520-521, and note that their "b" differs from ours by n - 1). Since KR21 is a function of n and the mean and variance of observed scores, just these two statistics are necessary for calculating approximate intervals under the BETA. (Recall that it is assumed that sample sizes are large enough to provide accurate estimates of the observed score mean and variance.)

For the BINORM model, the angular transformation suggested by Freeman and Tukey (1950) is used:

$$z = \frac{1}{2}\left[\text{SIN}^{-1}\sqrt{\frac{x}{n + 1}} + \text{SIN}^{-1}\sqrt{\frac{x + 1}{n + 1}}\right] \quad . \tag{11}$$

Under the binomial, this transformation is considered to provide the most stability in variance among the angular transformations suggested for this distribution (Mosteller & Tukey, 1968), and is the transformation used in most of the Bayesian references given above. The conditional distribution of G (variable for g) given τ is, approximately, normal with mean $\gamma = \text{SIN}^{-1}\sqrt{\tau}$ and variance $(4n + 2)^{-1}$, and the γ variable is assumed normal. Thus, tolerance limits can be calculated under the transformation by using the fact that the conditional distribution of γ given g is, approximately, normal $[\gamma(g), \rho^2_{\gamma G}/(4n + 2)]$, where

$$\gamma(g) = E\,\gamma + \rho^2_{\gamma G}(g - E\,G),$$

and

$$\rho^2_{\gamma G} = 1 - [\sigma^2_G(4n + 2)]^{-1} .$$

Thus, a central $1 - \alpha$ tolerance interval on the γ scale is

$$\gamma(g) \pm z_{\alpha/2}\,\rho_{\gamma G}(4n + 2)^{-\frac{1}{2}} . \qquad (12)$$

The inverse transformation $(\text{SIN})^2$ is then applied to the limits to return to the original true score scale.

The needed mean and variance $(E\,G, \sigma^2_G)$ can be estimated in a number of ways. A simple approach is to apply the Freeman-Tukey transformation to the observed scores and to calculate their mean and variance as estimates of $E\,G$ and σ^2_G . Since $E\,\gamma$ is approximately equal to $E\,G$, the mean of transformed observed scores can be used here also.

For the typical situation in which the mean and variance of observed proportion correct scores are already available, a more convenient approach to estimation employs a Taylor series approximation (see, e.g., Johnson & Kotz, 1969, pp. 28-29) for E G and σ^2_G. Under the Freeman-Tukey transformation,

$$E\ G \cong \frac{1}{2}\left(\mathrm{SIN}^{-1}\sqrt{p} + \mathrm{SIN}^{-1}\sqrt{q}\right) + \frac{\sigma^2_X}{16(n+1)^2}\left[\frac{1-2p}{[p(1-p)]^{3/2}} + \frac{1-2q}{[q(1-q)]^{3/2}}\right], \tag{13}$$

and,

$$\sigma^2_G \cong \frac{\sigma^2_X}{16(n+1)^2}\left[[p(1-p)]^{-\frac{1}{2}} + [q(1-q)]^{-\frac{1}{2}}\right]^2, \tag{14}$$

where $p = \mu n/(n + 1)$ and $q = (\mu n + 1)/(n + 1)$. Thus, accurate estimates of $\mu = E\ X/n$ and $\sigma^2_X = n^2\sigma^2_{X/n}$ are all that are needed for the BINORM intervals.[5]

Calculations for the CONORM model parallel those for the BINORM; i.e., tolerance limits are calculated under the Freeman-Tukey transformation to normal error and true score and then transformed back to the proportion correct scale. For this model we refer to true score under the transformation as $\eta = \mathrm{SIN}^{-1}\sqrt{\tau}$ -- in contrast to γ above. The distinction is made because of a difference in variances. Under the CONORM, the conditional variance of G given τ is $(n - 2k)/(4n^2 + 2n)$, approximately. This implies that with $k > 0$, $\sigma^2_{\eta G} > \sigma^2_{\gamma G}$.

[5] Again, this estimate of E G can be used for E γ since the two parameters are approximately equal. However, the following Taybr series approximation can also be used:

$$E\ \gamma \cong \mathrm{SIN}^{-1}\sqrt{\mu} + \sigma^2_\tau\ (1-2\mu)\ /\ [8(\mu-\mu^2)^{3/2}],$$

where σ^2_τ can be estimated from μ and σ^2_X. For the examples, the tolerance limits reported in two decimal places do not differ under the two approaches.

Tolerance intervals under the η scale are expressed as

$$\eta(g) \pm z_{\alpha/2}\, \rho_{\eta G}[(n - 2k)/(4n^2 + 2n)]^{\frac{1}{2}} \quad , \qquad (15)$$

where

$$\eta(g) = E\,\eta + \rho^2_{\eta G}(g - E\,G) \quad ,$$

and

$$\rho^2_{\eta G} = 1 - (n - 2k)/[\sigma^2_G(4n^2 + 2n)] \quad .$$

The $(SIN)^2$ transformation provides limits on the τ scale. Comparison of Equations 12 and 15 indicate that with $k > 0$, CONORM intervals will tend to be shorter and less regressed to the mean than BINORM intervals.

In addition to estimating $E\,G$ and σ^2_G for the CONORM, an estimate of k is also needed. As suggested by Lord (1965, p. 266),

$$\hat{k} = \frac{S^2_i(n - 1)}{2[\hat{\mu}(1 - \hat{\mu}) - \hat{\sigma}^2_{X/n} - S^2_i/n]} \approx 2n\, S^2_i \quad , \qquad (16)$$

where

$$S^2_i = \sum_{j=1}^{n} i^2_j/n - \hat{\mu}^2 \quad ;$$

i_j being the item difficulty (proportion correct) of the j-th item, and $\hat{\mu}$ being the m proportion correct for the sample on which the difficulties are based. As noted above, this estimate of k makes the average error variance on the observed score scale the same as would be obtained through a KR20 (Tucker, 1949, expresses KR20 in terms of n, $S^2_{X/n}$, and S^2_i).

Examples

Tables 1 through 4 provide selected tolerance limits for four different but realistic test characteristics. Test characteristics refer to the four parameters sufficient for calculating tolerance limits for all four models; namely, n, E X/n, $S^2_{X/n}$, and S^2_i (or KR20). The characteristics used are realistic because they were taken, with one exception, from established standardized tests, and are different because they allow contrasts between long and short tests and symmetric and skewed distributions.

Insert Table 1 about here

Table 1, with n = 35 and E X/n = .5, provides tolerance limits for observed number correct scores of 7, 14, 21, 28, and 35. Columns headed x/n and N.H. Dens. provide proportion correct scores and the corresponding negative hypergeometric densities that are associated with the BETA (Lord & Novick, 1968, pp. 515-520). The three tolerance coefficients, 50%, 68%, and 95%, were chosen to provide indications of differences in the conditional distributions at different percentage points. Also, these three coefficients have had historical popularity (50% for setting "probable error" intervals and for the interquartile range, 68% for one standard error intervals). Note also that the 95% intervals in Table 1 are, *approximately*, three times wider than the 50% intervals and twice as wide as the 68% intervals.

The similarity of the limits of the four models is striking. With the exception of the extreme observed score, 35 correct, and excluding the limits of the NORM model, the limits differ by no more than .01.

Considering the NORM model, the largest differences with the other models are at the extreme scores. Recall that under the NORM model, intervals for every observed score are the same width. Because error variances of the other three models decrease as true score moves away from .5 (the mean in this example), narrower intervals are found at the extremes. In Table 1 this can be seen at the score of 35, and to a lesser extent at scores of ?8 and 7. Around the mean, all four models provide limits that are quite close. Notice also that limits under the NORM model can be below zero or above one.

Recall that error variance under the CONORM is smaller than under the BINORM to a degree that is dependent on S_i^2 (or equivalently, the difference between KR21 and KR20). The effect of this on the limits in Table 1 appears to be slight but predictable. The intervals for the CONORM model are shorter by .01, approximately. Also, the differences are primarily reflected in the upper limits when observed scores are below the mean, and the lower limits when observed scores are above the mean; i.e., these limits, under the CONORM model, are .01 more distant from the mean (less regression) than under the BINORM.

The error variances for the BETA and BINORM models are the same but the shape of the true score distributions are different. From Table 1, the effect of this differ ce appears to be mainly on the extreme score of 35. The lower limits under the BETA are closer to the mean than under the BINORM and CONORM. A slight but opposite trend is found at scores of 7 and 28; i.e., the intervals under the BINORM are slightly closer to the mean than under the BETA.

These detailed descriptions of differences in Table 1 seem insignificant, but they do reflect some general patterns of differences that are discussed later.

The example in Table 1 does not provide a sufficient comparison of the models, because under all four models the true score distribution is symmetric (E X/n = .5). In Table 2, skewed true score distributions are introduced with E X/n = .75. Under the BETA model, the left skewness introduced by this mean is reflected in the negative hypergeometric densities. Of course, under the NORM model there is no skewness. Under the BINORM and CONORM models, skewness is allowed for through the expansion, above the mean, of the transformed true score scale.

Insert Table 2 about here

The effects of skewness on differences between the NORM and the other three models is quite noticeable. At observed scores below the mean (7, 14, 21), the upper limits of the NORM are further from the mean than those of the other models. Also, the intervals for these same scores are narrower for the NORM than for the others, but for scores above the mean (28, 35), the intervals are wider for the NORM. Comparing this with results from Table 1 (same n), the left-skewness appears to affect the width of the intervals at scores below the mean, making them wider than for a symmetric distribution. Also, scores above the mean have narrower intervals than those of Table 1. This result can be intuitively understood by considering the density of the true score distribution below and above the mean and its effect on the conditional distribution on which the limits are based.

The clear pattern of differences among the BETA, BINORM, and CONORM models that were found for the symmetric distributions associated with Table 1 are not as apparent in Table 2. Still, as expected, the CONORM intervals are typically shorter and sometimes further from the mean. Also, differences among limits for these three models are again small--just 12 out of 90 possible differences are greater than .01, and 10 of these are .02.

Table 3 contains intervals that are to be compared with those of Table 1. The test characteristics for Table 3, rather than being calculated from an existing test, were derived from those in Table 1. Note that the mean and S_i^2 are the same in both tables, but that n is 25 in Table 3. The observed score variance in Table 3 was derived by keeping the KR20 estimate of true score variance the same in both tables and increasing error variance by the multiple 35/25.

Insert Table 3 about here

The increase in widths of the intervals due to the decrease in n can be expressed algebraically for the NORM. So, under the NORM, the differences in widths between Tables 1 and 3 follow a simple pattern. For the 50% intervals, the differences are .014, for the 68%, they are .02, and for the 95%, they are .04. Considering the 95% interval widths for the 35-item test, a 13% decrease in width is obtained from a 40% increase in n. Also, there is less regression toward the mean that comes with the higher reliability of the 35-item test. For the other three models, similar differences are found, but there is less consistency in their pattern.

Similar to Table 1, when the limits for the NORM and those for a perfect score of 25 are excluded, differences among limits for the three other models are mainly .00 or .01. There are, however, six differences equal to .02, indicating that intervals tend to differ more with smaller n.

Table 4 provides intervals for a 100-item test (E X/n = .75). Note that for this long test the reliability is below .9. The test characteristics reflect those of a certification examination that has a small true score variance. The intervals here are much smaller than in the other tables. For the example in Table 1 (n = 35), the reliability is similar to the 100-item test, but it has 95% intervals that are almost twice the length of those in Table 4. Clearly, the number of items plays the primary role in the width of intervals for all the models considered here.[6]

Insert Table 4 about here

With the exception of the perfect score of 100, all four models have very similar limits. This is in contrast to the limits of Table 2 in which skewness introduced by E X/n = .75 made the limits for the NORM quite distinct from those of the other models. Actually, the coefficient of skewness under the BETA is smaller for the example in Table 4 than for Table 2, but the effect of skewness on the intervals is still noticeable. For example, under the BETA, scores below the mean are associated with wider intervals than those above the mean, while the NORM intervals are a constant width.

[6] For all four models the error variances depend primarily on n, but recall that this need not be the case for the CONORM and NORM models in which one has an option of using an estimate of overall error variance different from that obtained from a KR20.

Some General Comparisons

To obtain a more general idea of differences among the models, mean-absolute-differences were calculated using all the limits from the example tests in Tables 1 through 4 as well as those from three other examples. Table 5 contains these means which were calculated by contrasting limits for the six possible pairs of the four models. As an example, consider the first test depicted in Table 5 (n = 25, E X/n = .5). Here we find the mean-absolute-difference between the lower 95% limits of the BETA and the BINORM models is .006. This mean appears consistent with Table 3, in which most differences in limits between these two models are either .00 or .01.

Insert Table 5 about here

Each mean-absolute-difference was calculated by contrasting limits for a pair of models at each observed score, and by weighting each absolute difference by the negative hypergeometric density associated with the BETA model. This weighting was especially valuable for the longer tests. Consider the 100-item test with E X/n = .75. From empirical data and according to the density function, there are very few, if any, examinees who score below 20 on this test. So, it seems clear that some function is necessary that avoids weighting differences at scores below 20 in the same way as differences around the mean observed score Otherwise, mean-absolute-differences could fail to reflect the nature of the differences that occur in practice.

An obvious trend in Table 5 is the decrease in mean-absolute-differences with an increase in number of items. Of course, the intervals are also shorter for the longer tests. However, from other calculations, it was found that the percentage decrease in widths for longer tests is less than the percentage decrease in mean-absolute-differences; i.e., the decrease in mean-absolute-differences is not simply a result of a decrease in the widths of intervals.

Another result from Table 5 i. that the largest means are found for the contrast of the NORM with the other three models. This is consistent with the selected limits of Tables 1 through 4. Of these differences, some of the largest occur with contrasts of the upper limits for the examples with E $X/n > .5$. Recall that the other three models have left-skewed true score distributions when E $X/n > .5$. In effect, these results are an indirect indicator of the fact that for observed scores below the mean the upper limits for the NORM model are farther from the mean than those of the other three models. A more direct indicator is the mean difference (with sign) of the upper limits for observed scores below the mean. For the examples in which E $X/n > .5$, mean differences of upper limits (NORM-others) are all positive. For the two cases in which E $X/n = .5$, the means are also positive but much smaller. Table 6 provides means for three examples.

Insert Table 6 about here

Returning to Table 5, the mean-absolute-differences between the BINORM and CONORM models (different error variances) do not seem any larger than differences between the BETA and BINORM models (differences in shape of the true score distributions) Recall that the larger the value of S_i^2 the greater the difference in error variances between the CONORM model and both the BINORM and BETA models. This makes the CONORM intervals shorter and slightly less regressed to the mean. Apparently, the values of S_i^2 are not large enough to cause important differences in limits. Since the values of S_i^2 used here seem typical of standardized tests, the small effect of this parameter on the limits can be considered general.

There is a convenient contrast of S_i^2 in the examples. Consider the two tests with n = 35. One has an S_i^2 that is 50% larger than the other. Table 7 provides the mean widths of the intervals for these two examples (the negative hypergeometric density is used for these means also). Notice that the differences in mean widths between the BINORM and CONORM models for the test with $S_i^2 = .027$ are about 50% larger than the differences in mean widths for the test with $S_i^2 = .018$. Still, the differences for the larger S_i^2 represent just 8% of the mean widths of CONORM intervals, and differences for the smaller S_i^2 are only 4% of the mean widths.

Insert Table 7 about here

Table 7 also contains the mean widths of intervals for the example with n = 100. These are provided as an indication of the decrease ~~in~~ width that comes with an increase in n.

Plots of interval widths against observed scores were made for all the examples. The plots are not included here, but their general nature can be described. From Equation 6, the interval widths for the NORM model are constant across the observed scores. For the other three models, the plots are similar and depend on the mean. With E X/n = .5, the interval widths increase from zero to the mean, and then decrease symmetrically from the mean to 1.0. Around the mean, the widths for the three models are larger than for the NORM and smaller otherwise. For the examples with E X/n = .75, the intervals are approximately the same width up to the mean, and then they decrease from the mean to 1.0. They decrease at a faster rate past the mean than when E X/n = .5 and end up (at 1.0) with a smaller width.

Differences between the BETA and both the BINORM and CONORM models that were noted in Table 1 were found more generally for all examples. Plots of differences in limits against observed scores reveal that the largest differences between the BETA and both the BINORM and CONORM models occur at very extreme observed scores. For very low scores, the upper limits of the BETA are closer to the mean than those of the BINORM and the CONORM. Similarly, for very high scores, the lower limits of the BETA are closer to the mean.

The plots of the differences also revealed that an opposite but slighter trend occurs for scores that are not extreme. That is, for such scores that are below the mean, the upper limits under the BETA are slightly further from the mean than under the BINORM, and for such scores that are above the mean, the lower limits under the BETA are further from the mean than under the BINORM. A similar change in trend occurs for the BINORM-CONORM contrast. These results are most apparent for the symmetric true score distributions, and they were noted in Table 1. For a skewed distribution, the trend is mitigated.

Summary and Conclusions for Tolerance Comparisons

The detailed differences in the tolerance intervals for our examples appear to follow a pattern, and many of the differences reflect what was expected from differences in the models. In this sense, the difference can be considered generalizable to other realistic test characteristics.

The NORM model seems inappropriate for number correct scoring. The bounded nature of a proportion correct score scale is an apparent problem, and the assumption of independence of error and true score (without a transformation) seems unwarranted (Lord, 1960). These issues are reflected in differences between the NORM intervals and the other three models, especially for shorter tests. But, recall that for longer tests the intervals are quite similar for all four models.

The differences among the BETA, BINORM, and CONORM intervals seem unimportant. Because the BETA model has been frequently discussed in the literature, appears useful for a variety of applications, and does not involve approximations, one might feel satisfied in calculating intervals under the BETA and ignoring the other two models. However, the intervals that were calculated under the CONORM were based on KR20, whereas a different estimate of reliability could be incorporated. In other words, the small differences found between CONORM and BETA intervals were based on typical values of S_i^2, and might have been larger if reliabilities were estimated in a different manner.

From the results, there seems to be little reason to choose the BINORM over the BETA model for calculating intervals. However, it could serve as a substitute for the BETA, especially since the BINORM model has some mathematical conveniences that might prove useful for the problem of estimating tolerance intervals with small sample sizes.

The NORM model is quite distinct from the others, yet the tolerance intervals for scores not at the extreme were similar to the other models. Since average error variances were similar for all four models, the comparisons can be considered to be among differences in the sizes of error variance at different score levels and in the shapes of true score distributions. One can conclude that the small differences in intervals, at other than extreme scores, indicate that tolerance intervals are not very sensitive to differences in shapes of true score distribution or in assumptions about the variability of error variance along the true score scale. However, all four models do have regularly-shaped distributions and differences among them in error variances, at other than extreme scores, are not that large.

Comparison of Confidence Intervals

Three error models are used for calculating confidence intervals in the examples below: normal error with equal variance for all examinees, binomial error, and compound-binomial error.

The normal intervals are of the form

$$x/n \pm z_{\alpha/2}\,\sigma_e \qquad (17)$$

Note that σ_e is calculated through a KR20 for the examples below, and the same values of σ_e were used for the NORM model tolerance intervals in Tables 1 through 4.

For the binomial error model, there are many published tables specifically developed for confidence intervals on a binomial parameter. See Kendall and Stuart (1979, p. 129), and Johnson and Kotz (1969, p. 59) for references. However, none of the available tables provide confidence intervals for the 50% and 68% coefficients, so these calculations had to be performed for this paper. The calculations are straightforward enough to be generally useful. Some details about the calculation of these intervals are reported below to allow an analytic comparison with tolerance intervals under the BETA model.

Most of the published tables on binomial confidence intervals were generated by solving the following equations for the lower (L) and upper (U) limits of the intervals:

$$\sum_{j=x}^{n} \binom{n}{j} L^{j} (1 - L)^{n-j} = \alpha/2 \quad , \qquad (18)$$

$$\sum_{j=0}^{x} \binom{n}{j} U^j (1 - U)^{n-j} = \alpha/2 \quad . \tag{19}$$

Here, x is the observed number of successes (correct) in n trials (items).

Because of the discrete nature of the binomial, it is not possible, in general, to construct intervals with a particular coefficient. Intervals constructed from Equations 18 and 19 do have a coverage probability greater than or equal to $1 - \alpha$; i.e.,

$$P(L \leq \pi \leq U) \geq 1 - \alpha \quad , \tag{20}$$

where π is the binomial parameter and L and U are now considered random variables that are functions of X rather than x. Kendall and Stuart (1979, pp. 113-116 and pp. 129-131) provide a discussion about the issue of inexact intervals for the binomial. And, Wilks (1962, p. 368) provides a general theorem for setting confidence intervals for discrete variables.

Intervals constructed from Equations 18 and 19 are referred to as central intervals. This is because, in addition to the claim made in Equation 20, $P(L \leq \pi) \geq 1 - \alpha/2$ and $P(U \geq \pi) \geq 1 - \alpha/2$. These two additional statements seem to be a desirable feature of confidence intervals, and most tables are set up this way. However, by relinquishing these two claims, i.e., only requiring Equation 20 to hold, shorter noncentral intervals can be calculated. Crow (1956), among others, provides such intervals.

Equations 18 and 19 can be expressed in terms of the cumulative distribution function of a beta. Equation 18 can be written as

$$I_L(x, n - x + 1) = \alpha/2 \quad . \tag{21}$$

Thus, one can enter a beta table to find the L that corresponds to $\alpha/2$, or as was done for the tables below, use a computing routine for finding the inverse of a beta [IMSL (1979) subroutine MDBETI]. Similarly, for Equation 19, the upper limits can be determined by solving

$$I_{(1 - U)} (n - x, x + 1) = \alpha/2 \tag{22}$$

for U. (The F distribution can also be used; see Johnson & Kotz, 1969, p. 59.)

Recall Equation 7 for the lower limit of a tolerance interval under the BETA. Note that if a = 0 and b = 1 in that equation, it would equal Equation 21; making equal the lower limits of the binomial confidence interval and the BETA tolerance interval. Equation 8 for the upper tolerance limits can be reexpressed as $I_{(1 - U)} (b + n - x, a + x) = \alpha/2$. Note that a = 0 and b = 1 do not make this equal to Equation 22. Clearly, it is not possible to choose the a and b parameters of the true score distribution such that the confidence and tolerance limits are the same. This is not surprising given the different nature of the intervals. Consider also that under the binomial we can only make inequality statements because the coverage probability is a function of the discrete variable X . Under the BETA model, we make exact coverage probability statements because the variable τ given x is continuous.

For compound-binomial error, the two-term approximation which was discussed under the CONORM model was also used for error variance here. Recall that error variance under the approximation is $(n - 2k)/(4n^2 + 2n)$, where k was chosen to make average error variance the same as that calculated from a KR20. Also, recall that this made average error variance the same for the CONORM and NORM models.

Intervals for the compound-binomial are only approximations. The Freeman-Tukey transformation was used to yield approximate normality with constant variance. Intervals were then calculated $\left[g \pm z_{\alpha/2}[(n - 2k)/(4n^2 + 2n)]^{\frac{1}{2}}\right]$, a continuity correction was added, and a transformation back to the proportion correct scale was applied.

Two Examples

Two tables are provided for comparison of confidence intervals under the three error models. Table 8 contains intervals for a test with n = 35. The error variance, σ_e^2, for the normal error model corresponds to error variance under the NORM model for Table 1. Similarly, the same value of k was used in Tables 1 and 8. Table 9 has n = 100 and corresponds to parameters used in Table 4.

Insert Tables 8 and 9 about here

From Tables 8 and 9, confidence intervals under the three models are similar except at extreme scores. At the extreme score of 35, for example, all three error models have quite different limits. Typically, the normal error intervals extend beyond 1.0 and are much wider than intervals for the other two models. The compound-binomial intervals at 1.0 appear quite short relative to the binomial. This is not true at other observed scores, and seems to reflect problems with the properties of the transformation or the approximations at this extreme score.

At other observed scores, the binomial intervals are, for the most part, longer by .01 or the same as the compound-binomial intervals. This reflects the difference in error variance under the two models. Recall that k depends on S_i^2 and that k for Table 8 is associated with the largest S_i^2 in the examples. Also, k for Table 9 is the largest among all the examples.

Error distribution shapes affect the intervals in Table 8. Under the normal error distribution, the intervals are symmetric about the proportion correct score. In contrast, under the other two models, the distributions are skewed toward .5. For these two models, the lower limits are more distant from the observed score than the upper limits when the observed score is above the mean. The reverse holds for scores below the mean. This is not as noticeable for n = 100 in Table 9.

Comments on the Binomial Error Model

Under special circumstances, the binomial error model can be said to hold by definition (Lord & Novick, 1968, chap. 11, & chap. 23, p. 524; Lord, 1957). If test forms are constructed by random sampling of items and the proportion correct true score of interest is defined by the domain from which items are sampled (rather than for a particular sample of items), the binomial error model holds for any particular examinee as long as item responses are independent from one item to the next for that examinee (independence of responses is violated by context and other similar effects). Gross and Schulman (1980) provide a succinct justification of the binomial under such circumstances, and contradict some statements made by van der Linden (1979) in his claim of deterministic assumptions underlying the binomial.

The binomial error model is often criticized because items are not the same difficulty. It is true that the binomial distribution cannot be used for the joint distribution of error of examinees that are administered the same set of items. Errors are correlated across examinees. But when we isolate interest to a particular examinee under the circumstances above (random sampling of items, etc.), the distribution of observable scores for that examinee is binomial and it follows that confidence intervals based on the binomial are appropriate. Of course, this does not consider the nature of errors made in providing such confidence intervals for the set of examinees administered the same test form.

In any case, tests are not typically constructed by random sampling. For example, items are frequently sampled from fixed categories (Jarjoura & Brennan, 1982, provide a model for such circumstances). Also, test form difficulty and other adjustments are typical of standardized testing. It is usually judged that these factors make average error smaller than under the binomial, and binomial intervals are often viewed as conservative. Still, violation of other assumptions, like independence of item responses for an examinee, can make error larger than under the binomial. Binomial intervals can be considered a useful approximation as long as average error variance, estimated without resorting to binomial assumptions, agrees with that estimated under the binomial ($[1 - \text{KR21}]\hat{\sigma}^2_{X/n}$), and as long as there is no evidence that error variances at different points along the score scale are larger than under the binomial.

Comparison of Confidence and Tolerance Intervals

A comparison of Tables 1 and 8 provide an idea of differences between confidence and tolerance intervals under the same test characteristics. For the binomial, the fact that n = 35 is enough to allow comparisons across the tables. Recall that average error variance from the KR20 of Table 1 was used in determining error variance for the normal and compound-binomial intervals.

Contrasts between the BETA tolerance intervals and binomial confidence intervals reveal, as expected, that tolerance intervals are typically narrower and shifted from the observed score toward the mean. Differences in limits are most apparent at extreme scores. Note that the contrast in interval widths reverses at the extreme score of 1.0. Similar differences are found for contrasts between the normal and NORM intervals and between the compound-binomial and CONORM intervals. The BETA intervals of Table 2 can also be compared directly with the binomial intervals of Table 8. Here, we find some large differences at the low scores that are distant from the mean.

Direct comparisons can also be made between Tables 4 and 9. Recall that with n = 100, tolerance intervals for all four true score models are quite similar. In contrast, differences between confidence and tolerance intervals are large at scores that are distant from the mean. Consider, for example, the observed score of 20. There, 50% confidence and tolerance intervals do not even overlap, and for 68% intervals, the upper limits of the confidence intervals are the same or close to the lower limits of the tolerance intervals. The major reason for such a difference is that the observed score (20) is approximately 3.7

standard deviations below the mean (75). This difference might be considered unimportant because examinees do not score that low (empirically no one has scored this low on current forms of this example test). But the contrast does dramatize points made earlier. When we are conditionally interested in examinee a, then, from the perspective taken here, we are isolating interest in that examinee's distribution of observed scores, not in the distribution of scores of other examinees. This is not to say that information about other examinees cannot be used in interpreting a confidence interval. The point is that if we want a confidence interval for a particular examinee, then that interval is not designed to take the performance of other examinees into consideration. In contrast, when we condition on observed score, we are formally interested in observations from the population of examinees; i.e., in the associated distribution of true scores. Information that an observed score is very unlikely is obviously important and affects the nature of the tolerance interval.

Discussion

Such strong assumptions as used for setting tolerance or confidence intervals need to be checked. Some methods for checking are discussed below. Also, Bayesian credibility intervals and confidence and tolerance intervals are contrasted.

Checking Assumptions

All four true score models are specific about what to expect for observed score distributions. Thus, the usual chi-square test of fit could be calculated and differences between observed and expected frequencies examined. None of these models are likely to closely fit observations. However, consider the possibility that the BETA fits but the BINORM does not. Under such circumstances, one would prefer the BETA tolerance intervals, but, from the results above, they would not differ substantially from those of the BINORM.

If one assumes that an approximate compound-binomial error model is appropriate, then procedures developed in Lord (1969) and implemented in a computer program by Wingersky, Lees, Lennon, and Lord (1969) can be used to estimate a "smooth" true score distribution without specifying its form. This could be compared to a beta or the other true score distributions assumed in the models above in order to determine if there are large discrepancies. For example, the estimated distribution might be noticeably bi-modal or might be truncated at some point above zero. Clearly, this could cause problems in tolerance intervals. Lord and Stocking (1976) derive a procedure for setting simultaneous confidence intervals around the conditional means for true scores at every observed score. They assume the binomial error model but do not specify the true score distribution. These intervals could be compared with the conditional means that are specified by each of the four true score models. Also, Wilcox (1981) reviews procedures for checking the beta-binomial assumptions.

For the BETA, BINCRM, and CONORM models, the true score distribution is bounded by zero and one. The possibility of guessing correctly in multiple choice tests is often considered to imply that true scores do not extend down to zero.[7] Also, evidence of this effect has been found by Lord (1965) in using a four parameter beta distribution for true scores (two of the parameters are end-points). For a true score distribution that ends, say, at .15, tolerance intervals of number correct scores near zero would obviously be affected. This is rather unimportant if few examinees score near or below a guessing level (as is the case in most of the example tests above). Otherwise, a nonzero end-point should be considered in setting tolerance or confidence intervals.

Perhaps the most important checking is with regard to measurement error variance. Both confidence and tolerance interval widths are, for the most part, determined by error variance. And, under the above models, assumptions about error variance are quite strong. These assumptions could be checked, if deemed appropriate, by obtaining realized values of the error variable in a parallel forms study. A simple check on the binomial or the approximation to the compound-binomial error variances would involve transforming the observed scores (Freeman-Tukey), estimating error variance for appropriate ranges of observed scores, and comparing these with the constant values specified by the two models. If the estimated error variances are fairly constant but different from that specified under either model, this constant could be used for estimating k differently from that given in Equation 16.

Bayesian Credibility Intervals

With a Bayesian approach, we can isolate interest in a particular examinee's true score and still interpret an interval set up for that true score as covering a proportion of a distribution (posterior) of that true score for a given observed score or scores. This is because we start with a

[7] Note that true score is defined here as the expected proportion correct, not the expected proportion an examinee knows without guessing.

distribution (prior) for the true score. This is in contrast with a confidence interval that does not consider a distribution for the true score. In a sense, a Bayesian approach appears to provide a more informed statement or inference because it uses information besides an examinee's observed score in determining an interval for that examinee. As argued above, a confidence interval seems useful in the situation in which a career counselor or classroom teacher is interpreting a particular examinee's score. How a confidence interval ends up being interpreted will likely depend on all the other information a counselor or teacher has about that examinee, and perhaps information about the performance of other examinees. In this sense, a confidence interval can be considered a less formal method of inference as compared to a credibility interval.

Although a conceptual distinction exists between tolerance and credibility intervals, they can be made to coincide numerically. Consider that tolerance intervals under the BETA model are the same as central credibility intervals in the case in which every examinee is given the same prior (beta[a, b], where a and b are population parameters for the true score distribution) and the conditional distribution of observed scores is assumed binomial. It is not clear that they could be made the same when estimation issues are considered for tolerance intervals.

Conclusions

In consideration of issues regarding intervals for true scores, confidence intervals seem useful when score interpretation is intimately concerned with a particular examinee. In contrast, a tolerance interval is quite informative for interpreting a particular observed score with respect to a population of examinees. Also, knowledge that examinees who obtain a particular observed score likely have true scores within a 95% tolerance interval is a useful adjunct to a confidence interval for a particular examinee.

The claim that a confidence interval procedure covers, on average, the true scores of a population of examinees with some chosen probability depends on weak assumptions. It is not a very informative claim with respect to a particular examinee. If such a claim is the basis for interpreting confidence intervals, their usefulness for a particular examinee is diminished. Further, when a population of examinees are considered in the interpretation of an observed score, tolerance intervals are to be preferred.

Tolerance intervals provide simultaneously information about the discrimination afforded by a measurement procedure for some population of examinees and information about the precision of measurement. Consider the possibility of narrow tolerance intervals relative to the proportion correct scale (high precision) combined with few, if any, nonoverlapping tolerance intervals in the probable range of observed scores (low discrimination). This possibility can be translated simply to low reliability and small error variance (relative to the proportion correct scale), but it does much to clarify the meaning of such a statement. Confidence intervals are lacking in this regard.

Because tolerance intervals require the specification of the true score distribution conditional on observed score, it was necessary to address the issue of sensitivity of the intervals to differing strong assumptions about the joint distribution of observed and true scores. For realistic standardized test characteristics, tolerance intervals are, for the most part, insensitive to differences in the shapes of true score distributions and to small differences in error variances and reliabilities. In contrast, it is clear that confidence and tolerance intervals are quite distinct, especially for scores not close to the mean.

References

Crow, E. L. Confiden~e intervals for a proportion. Biometrika, 1956, 43, 423.

Dixon, W. J., & Massey, F. J., Jr. Introduction to statistical analysis. New York: McGraw Hill, 1969.

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

Graybill, F. A. Theory and application of the linear model. North Scituate, Mass.: Duxbury Press, 1976.

Gross, A. L., & Shulman, V. The applicability of the beta binomial model for criterion-referenced testing. Journal of Educational Measurement, 1980, 17, 195-200.

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.

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

International Mathematical and Statistical Libraries. IMSL Libraries (7th ed.). Houston: Author, 1979.

Jackson, P. H. Some simple approximations in the estimation of many parameters. British Journal of Mathematical and Statistical Psychology, 1972, 25, 213-228.

Jarjoura, D., & Brennan, R. L. A variance components model for measurement procedures associated with a table of specifications. Applied Psychological Measurement, 1982, 6, 161-171.

Johnson, N. L., & Kotz, S. Distributions in statistics: Discrete distributions. Boston: Houghton Mifflin, 1969.

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

Kendall, M., & Stuart, A. The advanced theory of statistics (4th ed., Vol. 2). New York: MacMillan, 1979.

Lewis, C., Wang, M., & Novick, M. R. Marginal distributions for the estimation of proportions in m groups. Psychometrika, 1975, 40, 63-75.

Lieberman, G. J., & Miller, R. G. Simultaneous tolerance intervals in regression. Biometrika, 1963, 50, 155-168.

Lord, F. M. Do tests of the same length have the same standard error of measurement? Educational and Psychological Measurement, 1957, 17, 510-521.

Lord, F. M. An empirical study of the normality and independence of errors of measurement in test scores. Psychometrika, 1960, 25, 91-104.

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., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Lord, F. M., & Stocking, M. An interval estimate for making statistical inferences about true scores. Psychometrika, 1976, 41, 79-87.

Mosteller, F., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronsen (Eds.), The handbook of social psychology. Reading, Mass.: Addison-Wesley, 1968.

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

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.

Proschan, F. Confidence and tolerance intervals for the normal distribution. Journal of the American Statistical Association, 1953, 48, 550-564.

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

Stanley, J. C. Reliability in R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington: American Council on Education, 1971, 356-442.

Tucker, L. R. A note on the estimation of test reliability by the Kuder-Richardson formula (20). Psychometrika, 1949, 14, 117-120.

van der Linden, W. J. Binomial test models and item difficulty. Applied Psychological Measurement, 1979, 3, 401-411.

Wald, A., & Wolfowitz, J. Tolerance limits for a normal distribution. Annals of Mathematical Statistics, 1946, 17, 208-215.

Wallis, W. A. Tolerance intervals for linear regression in J. Neyman (Ed.), Second Berkeley Symposium on Mathematical Statistics and Probability. Berkeley: University of California Press, 1951, 43-51.

Wilcox, R. R. Estimating true score in the compound binomial error model. Psychometrika, 1978, 43, 245-258.

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

Wilks, S. S. Mathematical statistics. New York: Wiley, 1962.

Wingersky, M. S., Lees, D. M., Lennon, V., & Lord, F. M. A computer program for estimating true-score distributions and graduating observed score distributions (ETS Research Bulletin 69-4). Princeton, N.J.: Educational Testing Service, 1969.

TABLE 1

Tolerance Intervals for n = 35, E X/n = .5, $\sigma^2_{X/n}$ = .0423, and S^2_i = .027

Obs.	$\frac{x}{n}$	N.H.		BETA		BINORM		CONORM		NORM	
Score		Dens.	Coeff.	L	U	L	U	L	U	L	U
			50%	20	29	21	30	20	29	19	28
7	.2	.024	68%	18	31	19	32	19	31	17	31
			95%	13	38	13	39	14	38	10	37
			50%	36	47	37	47	37	47	37	46
14	.4	.046	68%	34	49	34	50	35	49	34	48
			95%	27	57	28	57	28	56	28	55
			50%	53	64	53	63	54	63	54	63
21	.6	.046	68%	51	66	51	66	51	65	52	66
			95%	43	73	43	72	44	72	45	72
			50%	71	80	70	79	71	80	72	81
28	.8	.024	68%	69	82	68	81	69	81	69	83
			95%	62	87	61	87	62	87	63	90
			50%	91	96	94	98	95	98	89	98
35	1.0	.001	68%	90	97	93	99	94	99	87	101
			95%	83	99	88	100	90	100	80	107

Note. KR20 = .87, KR21 = .86, k = 2.2, beta a = 2.953, and beta b = 2.953. Decimal points on limits are omitted.

TABLE 2

Tolerance Intervals for n = 35, E X/n = .75, $\sigma^2_{X/n}$ = .0227, and S^2_i = .018

Obs.	$\frac{x}{n}$	N.H.		BETA		BINORM		CONORM		NORM	
Score		Dens.	Coeff.	L	U	L	U	L	U	L	U
			50%	27	36	28	38	27	36	26	34
7	.2	.001	68%	25	39	26	40	25	38	25	36
			95%	19	46	20	47	19	45	19	42
			50%	42	53	44	54	43	53	42	51
14	.4	.008	68%	40	55	42	56	41	55	41	53
			95%	33	62	35	63	34	62	35	58
			50%	58	68	59	69	59	68	59	67
21	.6	.038	68%	56	70	57	71	56	70	57	69
			95%	49	77	50	77	50	76	51	74
			50%	75	83	74	83	75	83	75	83
28	.8	.075	68%	73	85	72	84	73	84	73	85
			95%	66	90	66	89	66	89	68	91
			50%	93	97	95	98	95	99	91	100
35	1.0	.018	68%	92	98	94	99	94	99	89	101
			95%	87	99	90	100	91	100	84	107

Note. KR20 = .81, KR21 = .79, k = 1.9, beta a = 7.109, and beta b = 2.370. Decimal points on limits are omitted.

TABLE 3

Tolerance Intervals for $n = 25$; $E\ X/n = .5$, $\sigma^2_{X/n} = .0444$, and $S^2_i = .027$

Obs. Score	$\frac{x}{n}$	N.H. Dens.	Coeff.	BETA L	BETA U	BINORM L	BINORM U	CONORM L	CONORM U	NORM L	NORM U
5	.2	.034	50%	20	31	22	32	21	31	20	30
			68%	18	34	19	35	19	34	17	33
			95%	12	42	13	43	13	41	10	40
10	.4	.063	50%	36	48	37	48	37	48	36	47
			68%	33	51	34	51	34	51	34	50
			95%	25	59	26	60	27	58	26	57
15	.6	.063	50%	52	64	52	63	52	63	53	64
			68%	49	67	49	66	50	66	51	66
			95%	41	75	41	74	42	73	43	74
20	.8	.034	50%	69	80	68	78	69	79	70	80
			68%	66	82	65	81	66	81	67	83
			95%	58	88	58	88	57	87	60	90
25	1.0	.003	50%	87	94	91	97	92	97	86	97
			68%	85	95	89	98	91	98	84	100
			95%	78	98	84	100	86	100	76	107

Note. KR20 = .83, KR21 = .81, k = 1.6, beta a = 2.985, and beta b = 2.985. Decimal points on limits are omitted.

TABLE 4

Tolerance Intervals for n = 100, E X/n = .75, $\sigma^2_{X/n}$ = .0119, and S^2_i = .020

Obs.	$\frac{x}{n}$	N.H.		BETA		BINORM		CONORM		NORM	
Score		Dens.	Coeff.	L	U	L	U	L	U	L	U
20	.2	.000	50%	25	31	25	31	25	30	25	30
			68%	24	32	24	32	24	31	24	31
			95%	21	37	21	37	20	35	20	35
40	.4	.001	50%	42	48	43	49	42	48	43	47
			68%	41	50	41	50	41	50	41	48
			95%	36	54	37	55	37	54	37	52
60	.6	.013	50%	59	65	60	66	59	65	60	65
			68%	58	67	58	67	58	67	58	66
			95%	53	71	54	71	54	70	55	69
80	.8	.036	50%	77	82	77	82	77	82	77	82
			68%	76	83	75	83	76	83	76	83
			95%	72	86	71	86	72	86	72	87
100	1.0	.000	50%	95	98	98	99	98	99	94	99
			68%	95	98	97	99	98	100	92	100
			95%	92	99	96	100	96	100	90	104

Note. KR20 = .87, KR21 = .85, k = 6.2, beta a = 13.137, and beta b = 4.379. Decimal points on limits are omitted.

TABLE 5

Mean-Absolute-Differences of Tolerance Limits for Seven Test Characteristics[a]

Test Characteristics	Coeff.	BETA-BINORM L	BETA-BINORM U	BETA-CONORM L	BETA-CONORM U	BINORM-CONORM L	BINORM-CONORM U	BETA-NORM L	BETA-NORM U	BINORM-NORM L	BINORM-NORM U	CONORM-NORM L	CONORM-NORM U
n=25 EX/n=.50	50%	9	9	6	6	5	5	6	6	12	12	8	8
σ^2(X/n)=.044	68%	8	8	6	6	6	6	9	9	14	14	9	9
S^2(i)=.027	95%	6	6	11	11	8	8	18	18	21	21	15	15
n=25 EX/n=.75 [b]	50%	8	8	5	7	4	4	4	10	7	14	5	11
σ^2(X/n)=.024	68%	8	7	4	7	5	4	7	13	8	18	7	14
S^2(i)=.018	95%	7	4	9	5	6	5	18	27	17	30	15	28
n=35 EX/n=.50	50%	7	7	5	5	4	4	5	5	9	9	6	6
σ^2(X/n)=.042	68%	6	6	5	5	5	5	7	7	11	11	8	8
S^2(i)=.027	95%	5	5	9	9	8	8	16	6	18	18	13	13
n=35 EX/n=.75	50%	6	6	4	5	3	3	4	9	6	12	4	9
σ^2(X/n)=.023	68%	6	5	4	5	4	3	7	12	7	15	6	13
S^2(i)=.018	95%	5	3	8	5	6	5	17	24	15	26	13	24
n=50 EX/n=.60	50%	4	4	3	3	2	2	3	3	3	5	2	4
σ^2(X/n)=.029	68%	4	4	3	4	3	3	4	4	4	7	3	6
S^2(i)=.020	95%	3	3	7	5	5	5	7	10	7	13	5	12
n=75 EX/n=.60	50%	3	3	2	2	2	2	2	4	3	6	2	4
σ^2(X/n)=.023	68%	3	3	2	3	3	3	4	5	4	7	3	5
S^2(i)=.022	95%	2	2	6	5	5	5	8	10	8	12	5	9
n=100 EX/n=.75	50%	2	2	1	2	2	2	3	3	2	4	2	2
σ^2(X/n)=.012	68%	2	2	2	3	2	2	4	4	3	5	2	2
S^2(i)=.020	95%	2	1	6	4	5	4	8	9	7	10	5	4

[a] Means are in thousandths; i.e., 5=.005.

[b] Derived from example test with n=35, EX/n=.75.

TABLE 6

Mean Differences of Upper Limits For Observed Scores Below the Mean

		E X/n = .5			E X/n = .75		
		BETA-NORM	BINORM-NORM	CONORM-NORM	BETA-NORM	BINORM-NORM	CONORM-NORM
	50%	.005	.010	.004	.017	.019	.014
n=35	68%	.007	.012	.005	.020	.023	.018
	95%	.017	.019	.008	.026	.029	.023
	50%	.007	.013	.006	.022	.024	.018
n=25	68%	.009	.016	.007	.025	.027	.021
	95%	.020	.022	.009	.029	.033	.026

TABLE 7

Mean Widths of Intervals

Test Characteristics	Coeff.	BETA	BINORM	CONORM	NORM
$n=35$ $EX/n=.5$	50%	.098	.097	.091	.092
$\sigma^2(X/n)=.042$	68%	.144	.143	.135	.137
$S^2(i)=.027$	95%	.281	.277	.262	.267
KR20=.87					
$n=35$ $EX/n=.75$	50%	.083	.082	.078	.080
$\sigma^2(X/n)=.023$	68%	.121	.121	.115	.118
$S^2(i)=.018$	95%	.237	.234	.224	.232
KR20=.83					
$n=100$ $EX/n=.75$	50%	.052	.052	.049	.053
$\sigma^2(X/n)=.012$	68%	.077	.077	.072	.078
$S^2(i)=.020$	95%	.151	.150	.141	.153
KR20=.87					

TABLE 8

Confidence Intervals for n = 35

Obs.	$\frac{x}{n}$		Binomial		Normal		Comp.-Bin.	
Score		Coeff.	L	U	L	U	L	U
		50%	15	27	15	25	15	26
7	.2	68%	13	29	13	27	13	28
		95%	8	37	6	34	8	35
		50%	33	47	35	45	33	47
14	.4	68%	31	50	33	47	31	50
		95%	24	58	26	54	24	57
		50%	53	67	55	65	53	67
21	.6	68%	50	69	53	67	51	69
		95%	42	76	46	74	43	76
		50%	73	85	75	85	74	85
28	.8	68%	71	87	73	87	72	87
		95%	63	92	66	94	65	92
		50%	96	100	95	105	99	100
35	1.0	68%	95	100	93	107	98	100
		95%	90	100	86	114	95	100

Note: Decimal points on limits are omitted.

TABLE 9

Confidence Intervals for n = 100

Obs.	$\frac{x}{n}$		Binomial		Normal		Comp.-Bin.	
Score		Coeff.	L	U	L	U	L	U
		50%	17	23	17	23	17	23
20	.2	68%	16	25	16	24	16	24
		95%	13	29	13	28	13	28
		50%	36	44	37	43	36	44
40	.4	68%	35	46	36	44	35	45
		95%	30	50	32	48	31	50
		50%	56	64	57	63	56	64
60	.6	68%	55	65	56	64	55	65
		95%	50	70	52	68	50	69
		50%	77	83	77	83	77	83
80	.8	68%	75	84	76	84	76	84
		95%	71	87	72	88	72	87
		50%	99	100	97	103	100	100
100	1.0	68%	98	100	96	104	99	100
		95%	96	100	92	108	98	100

Note: Decimal points on limits are omitted.