DOCUMENT RESUME

ED 229 387 SP 022 333

AUTHOR Feldt, Leonard S.
TITLE A Theory-based Comparison of the Reliabilities of Fixed-length and Trials-to-criterion Scoring of Physical Education Skills Tests.
PUB DATE Apr 83
NOTE 12p.; Paper presented at the National Convention of the American Alliance for Health, Physical Education, Recreation and Dance (Minneapolis, MN, April 7-11, 1983).
PUB TYPE Statistical Data (110) -- Speeches/Conference Papers (150) -- Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Methods; *Mathematical Formulas; Mathematical Models; *Measurement Techniques; Measures (Individuals); Observation; Physical Education; Psychomotor Skills; Statistical Analysis; Statistical Data; *Test Reliability
IDENTIFIERS *Fixed Length Testing; *Trials to Criterion Testing

ABSTRACT

This paper considers, from a theoretical point of view, two measurement approaches used in measuring success and failure in skills tests in physical education. The first, "fixed length" (FL) testing, entails counting the number of successful performances in a fixed number of trials. The second, "trials-to-criterion" (TTC) testing, involves counting the number of trials required to achieve a specified number of successes. TTC measurement results in high measurement error variance for individuals with low probabilities of success on a single trial. Error variance declines as the probability rises. If there are many more people with low probabilities than there are with high probabilities, which is the case for a positively skewed distribution, the TTC approach will result in less reliable measurement than will the FL approach. Under the latter, error variance is largest for people with a probability of .5. Individuals lower and higher will have smaller error variances. Two generalizations based on these results can be made with regard to skills testing: (1) If the skills test task is one on which most untrained individuals perform poorly, FL testing would be the better choice; and (2) If the test scores tend to be negatively skewed, then TTC testing would be more efficient and reliable for the same total testing time. Two formulas are presented for estimating the reliability of TTC measures. (JM)

A Theory-based Comparison of the Reliabilities of Fixed-length and Trials-to-criterion Scoring of Physical Education Skills Tests

Leonard S. Feldt

University of Iowa

Paper presented at the 1983 AAHPERD National Convention

Whenever a skills test involves an action that may be classified unambiguously as successful or unsuccessful, such as shooting a free throw in basketball, the tester has a choice between two measurement approaches. The first entails counting the number of successful performances in a fixed number of trials. The second involves counting the number of trials required to achieve a specified number of successes. The first of these is by far the more common, but there are situations in which the second may have clear advantages. In this paper the first approach is called "fixed length" or FL testing. On such a test the higher the score is, the better the performance. The second approach will be referred to as "trials-to-criterion" or TTC testing. On this type of measurement the lower the score is, the better the performance. The purpose of this paper is to consider -- from a theoretical point of view -- how these approaches compare in reliability.

In order to lay the foundation for the comparison, it will be necessary to present some theoretical results for the two types of measurement. Under either type, each examinee is assumed to have a personal probability of succeeding on any trial. In the literature, this unknown parameter of subject $\underline{i}$ is symbolized by ϕ_i, and it is assumed that during testing ϕ_i remains constant. Under fixed length testing, with $\underline{k}$ trials for everyone, person $\underline{i}$'s true score equals $k(\phi_i)$. Under trials-to-criterion testing, with $\underline{R}$ successes required for the test to end, the true score of person $\underline{i}$ equals R/ϕ_i. True score is defined here as the long-run average, or expected value, of the person's observed score if the individual could be measured many times in the one way or the other. The variance of observed

scores for person i under repeated measurement--a concept commonly called measurement error variance--will be $k(\phi_i)(1-\phi_i)$ under fixed length measurement and $R(1-\phi_i)/\phi_i^2$ under trials-to-criterion measurement.

These expressions for true score and error variance are deducible from well know statistical distributions: the binomial distribution and the negative binomial or Pascal distribution. Their application to measurement situations is quite direct and the theories are well established.[1]

In order to compare the reliabilities of FL and TTC measurement, some ground rules must be adopted which render FL and TTC tests comparable in length. Any test A can be shown to be potentially more reliable than another test B if test A can be made much longer than test B. This demand for equity in length is made somewhat complicated by the fact that TTC measurement doesn't have a fixed stopping point. The number of trials is certain to vary from one examinee to another. However, if one postulates one or another population distribution of ϕ values, as we shall do later in this paper, one can use the theory to deduce the population average of the number of trials needed per examinee. In our comparison of the two types of measurements we took k, the number of trials for the FL measurement, equal to the theory-deduced average number under TTC measurement. This seemed a reasonable basis for comparison. It also turns out to have an unexpected, unanticipated virtue. Under this definition of comparable length, the value of the criterion (R) for TTC measurement does not influence the decision as to which form of testing is more reliable. A value of R equal to 5 will lead to the same conclusion as R equal to 10 or any other required number of successes.

The final bit of background theory that is needed is the variance definition of reliability, that is, $\rho_{rel} = \sigma_T^2 / \sigma_X^2 = \sigma_T^2 / \left(\sigma_T^2 + \sigma_E^2\right)$. Reliability equals the ratio of true score variance to observed score variance, and observed score variance equals the sum of true score variance plus error variance. Under both types of measurement, error variance is not the same for all examinees. It varies from person to person, depending upon ϕ_i. When this is the case, the variance of observed scores equals true score variance pl ɔ the average error score variance, σ_E^2 being averaged over the entire population of examinees.

With this foundation, it is possible to get on with the comparisons. No one ever knows how examinees distribute themselves with respect to ϕ. Therefore, six different possibilities were considered. In each hypothetical case ϕ_i ranges from .2 to .8. Some examinees are postulated to be rather inept, some are assumed very proficient, and most fall somewhere in between. Figure 1 shows these distributions in graphical form. It can be seen that the distributions include a crude normal distribution, two degrees of both positive and negative skewness, and a symmetrical distribution that is rather flat (platykurtic). For purposes of TTC measurement a value of 5 was adopted for R. However, the adoption of five successes was not material. The decision regarding which type of measurement would be more reliable in each case would have been the same regardless of the value chosen for R.

Table 1 summarizes the crucial statistics for the two types of measurements under each postulated distribution. To illustrate the meaning of the values: under a normal distribution of ϕ_i, TTC measurement would result in an average of 11+ trials per subject. This is the meaning of μ_T. Consistent with this value, the value of k for FL testing

was taken as 11. The variance of true scores under TTC was 16.229; under FL measurement true score variance equaled 2.468. Observed score variances were about 33 and 5, respectively. In a numerical sense, subjects spread out much more under TTC measurement than under FL measurement, in which everyone is allotted exactly 11 trials. But these quantities are not the primary facts of interest here.

The most important facts are the reliability coefficients in the last row of the upper and lower halves of the table. These values indicate which type of measurement is superior, in terms of reliability, for each population. As one may see, in some cases the advantage lies with TTC testing and in other cases with FL testing. The trends may be summarized as follows:

1) When the values of ϕ are close to being normally distributed around $\phi = .5$, the approaches are about equal in reliability. (A normal distribution centering around $\phi = .6$, and $\phi = .7$, gave practically the same results.)

2) When the bulk of the distribution is below $\phi = .5$, and only a light tail extends upward toward $\phi = .8$, FL is the better approach. The stronger the degree of positive skewness, the more marked the FL superiority.

3) When the bulk of the ϕ distribution is above $\phi = .5$, and only a long tail extends downward toward $\phi = .2$, TTC is the better approach. The stronger the degree of negative skewness, the more marked is the TTC superiority.

4) Platykurtosis, accompanied by symmetry, results in an advantage for FL measurement. Heaviness in the upper range of ϕ values does not compensate for similar heaviness in the lower range of ϕ values. Thus,

as the symmetry of the normal distribution shifts toward the symmetry of a platykurtic distribution, the equal reliability situation changes to an advantage for FL measurement.

To summarize the trends briefly, TTC measurement results in high measurement error variance for individuals with low probabilities of success on a single trial. Error variance declines as the probability rises. If there are many more people with low probabilities than there are with high probabilities, which is the case for a positively skewed distribution, the TTC approach will result in less reliable measurement than the FL approach. Under the latter, error variance is largest for people with a probability of .5. Individuals lower and higher will have smaller error variances.

What implications do these results have for skills testing? A few tentative generalizations may be offered. If the skills test task is one on which most untrained individuals perform poorly (say, $\phi < .5$), FL testing would be the better choice. Such might be the case with pre-instruction tests, placement tests, or any measurement likely to yield scores that are positively skewed. If the test scores tend to be negatively skewed, then TTC testing would be more efficient and reliable for the same total testing time. This is more likely to be true of post-instruction scores than pre-instruction scores, although it could be true of both. Symmetrical distributions, particularly those that are "flatter-than-normal," call for the use of the FL approach.

These recommendations are predicated on the use of a value of k reasonably close to the expected value of R/ϕ. If TTC is not used, there is no reason to specify R, nor would the examiner ever know the

expected value of R/¢. However, if the choices of k and R were made equitably, the foregoing recommendations would apply. The recommendations are valid in the sense of getting the highest reliability out of the total number of trials by all examinees.

To conclude this paper, two formulas are presented for estimating the reliability of TTC measures. These formulas have been derived by Dr. Judy Spray and her students. The first formula bears a striking resemblance to the familiar KR#21. The second is the general form of Cronbach's coefficient alpha. For this application of coefficient alpha, the TTC test is perceived as having R parts. The first part ends with the first successful trial, the second part ends with the second successful trial, and so on. The score of subject i on part j is the number of additional trials required by the subject to achieve the j th success after achieving the (j-1)st success. These formulas can be shown to be algebraically identical when population parameters are substituted in each. They are not necessarily equal when sample statistics are used. Investigations are underway to compare these formulas with respect to bias and sampling error.

$$\rho_{XX'} = \frac{R}{R+1}\left\{\frac{\sigma_X^2 - \mu_X(\mu_X - R)/R}{\sigma_X^2}\right\}$$

$$\rho_{XX'} = \frac{R}{R-1}\left\{\frac{\sigma_X^2 - \sum_{}^{R}\sigma_{Y_j}^2}{\sigma_X^2}\right\}$$

Y_j = the number of trials needed to achieve success j after achieving success (j-1)

Reference Notes

1. The application of binomial theory to tasks that may be scored as unsuccessful or successful (0 or 1 scoring) was first discussed by Lord, F. in "Estimating Test Reliability." Educational and Psychological Measurement, Winter, 1955, pp. 325-336.

   The application of the negative binomial distribution to measures defined as the number of trials required for a specified number of successes is discussed in Hays, W. Statistics (3rd Edition) New York: Holt, Rinehart, & Winston, 1981.

Figure 1. Six Population Distributions of ϕ, the Probability of Success for a Single Trial on a Hypothetical Skills Test

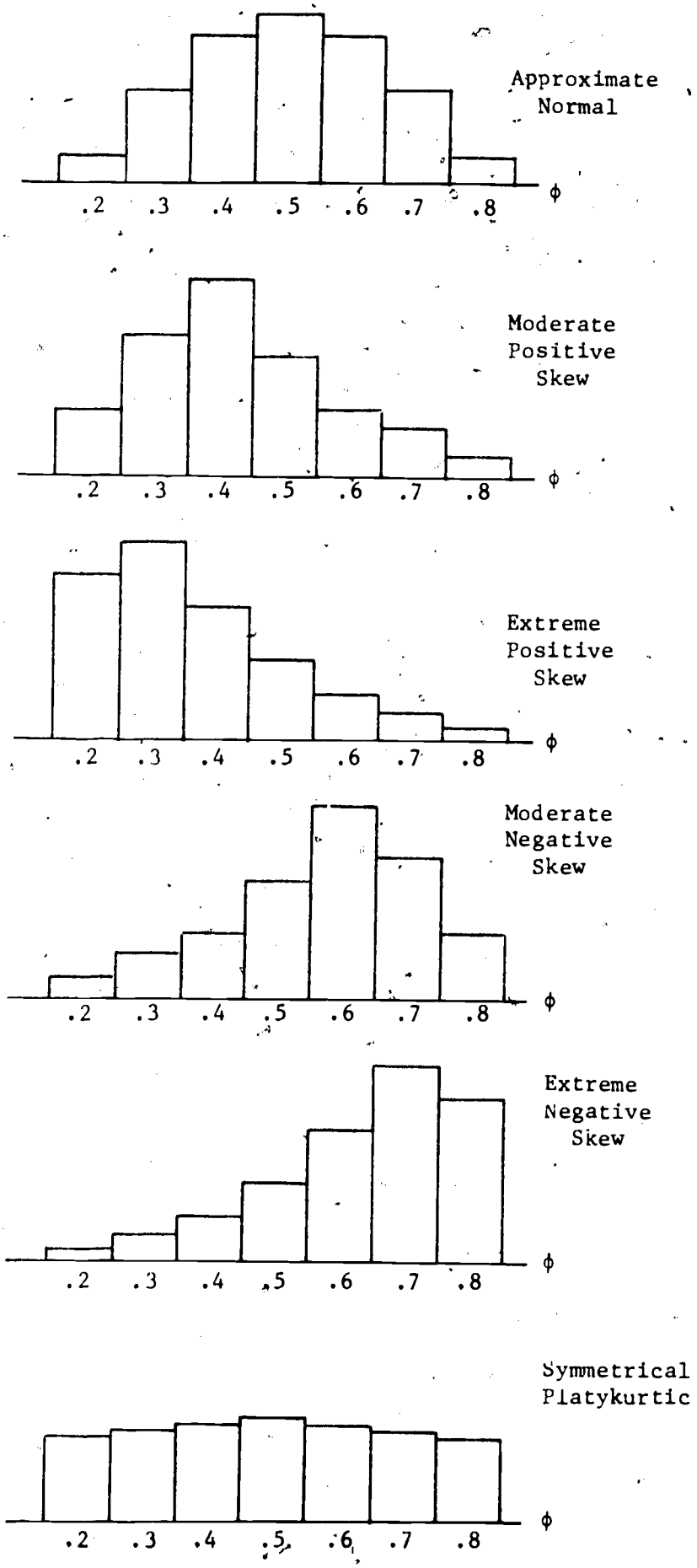

Table 1

Mean True Score, True Score Variance, Error Score Variance, Observed Score Variance and Reliability for Six Populations on a Hypothetical Skills Test

	Normal	Moderate Pos. Skew	Extreme Pos. Skew	Moderate Neg Skew	Extreme Neg. Skew	Flat Sym.
ITC Measurement (R=5)						
μ_T	11.052	13.238	15.944	9.663	8.614	12.121
σ_T^2	16.229	24.943	36.304	14.409	11.314	35.300
$\bar{\sigma}_e^2$	16.624	26.797	42.159	11.894	8.488	24.322
σ_X^2	32.853	51.740	78.463	26.303	19.802	59.622
$\rho_{XX'}$	.494	.482	.463	.548	.571	.592
FL Measurement						
k	11	13	16	10	9	12
μ_T	5.500	5.577	5.856	5.710	5.706	6.000
σ_T^2	2.468	3.762	5.798	2.226	1.835	5.414
$\bar{\sigma}_e^2$	2.526	2.895	3.350	2.227	1.885	2.549
σ_X^2	4.994	6.657	9.148	4.453	3.720	7.963
$\rho_{XX'}$	.494	.565	.634	.500	.493	.680

Formula Sheet for FL and TTC Measurement

	FL	TTC
1. True score for person i	$k\phi_i$	R/ϕ_i
2. Population mean true score	$k\overline{\phi}$	$R\left(\overline{\frac{1}{\phi}}\right)$
3. True score variance	$\sigma^2_{(k\phi)} = k^2\sigma^2_\phi$	$\sigma^2_{(R/\phi)} = R^2\sigma^2_{(1/\phi)}$
4. Error score variance of person i	$k\phi_i(1-\phi_i)$	$R\left\{\frac{(1-\phi_i)}{\phi_i^2}\right\}$
5. Population mean error variance	$k\overline{\phi(1-\phi)}$	$R\left\{\overline{\frac{(1-\phi)}{\phi^2}}\right\}$
6. Observed score variance	$k^2\sigma^2_\phi + k\overline{\phi(1-\phi)}$	$R^2\sigma^2_{(1/\phi)} + R\overline{(1-\phi)/\phi^2}$
7. Reliability (theoretical)	$\frac{k^2\sigma^2_\phi}{k^2\sigma^2_\phi + k\overline{\phi(1-\phi)}}$	$\frac{R^2\sigma^2_{(1/\phi)}}{R^2\sigma^2_{(1/\phi)} + R\overline{(1-\phi)/\phi^2}}$
	$\frac{\sigma^2_\phi}{\sigma^2_\phi + \overline{\phi(1-\phi)}/k}$	$\frac{\sigma^2_{(1/\phi)}}{\sigma^2_{(1/\phi)} + \overline{(1-\phi)/R\phi^2}}$
8. Reliability estimation formulas	$\frac{k}{k-1}\left\{\frac{\tilde{\sigma}^2_x - M_x(k-M_x)/k}{\tilde{\sigma}^2_x}\right\}$	$\frac{R}{R+1}\left\{\frac{\tilde{\sigma}^2_x - M_x(M_x-R)/R}{\tilde{\sigma}^2_x}\right\}$
	$\frac{ms_s - ms_{ws}}{ms_s}$	$\frac{R}{R-1}\left\{\frac{\tilde{\sigma}^2_x - \sum^R \tilde{\sigma}^2_{Y_j}}{\tilde{\sigma}^2_x}\right\}$

Y_j = number of trials needed to achieve success j after achieving success (j-1)