Systematic Errors in Approximations to the Standard Error of Measurement and Reliability

David J. Kleinke
Syracuse University

Lord's approximation to the standard error of measurement of a test uses only n, the number of items. Millman's is based on n and $\bar{p}$, the mean difficulty. Saupe has used Lord's approximation to derive an approximation to the reliability. Through an empirical demonstration involving 200 classroom tests, all three approximations are shown to be biased. The Lord and Millman approximations overestimate $s_X\sqrt{1\text{-}KR20}$, and thus Saupe's underestimates $r_{xx'}$ for these tests. The unweighted mean of the tests' mean item difficulties was .68, supporting Lord's original warning that his approximation be used cautiously with tests that are either very difficult or very easy. Still, the approximations did correlate very highly with their criteria, supporting their continued limited use.

For some time, Lord's (1959) approximation to the standard error of measurement of a binarily scored test has been considered to yield useful values. The approximation is

$$SE_{Meas} \hat{=} \frac{3}{7}\sqrt{n} \quad , \qquad [1]$$

where n is the number of items in the test. Lord based Approximation 1 on examination of the results of 58 tests to which he had access and also on earlier theoretical work (Lord, 1957). Lord specifically stated that Approximation 1 is appropriate only for tests with relative mean scores (mean item difficulties) between .35 and .65 (1959, p. 238). However, Lord's approximation was supported conceptually by Swineford (1959) and empirically by McMorris (1972), who used the results from 85 classroom tests. Further, Saupe (1961) used Approximation 1 to derive approximations to $r_{xx'}$, the reliability of a test.

A slightly different approach to approximating the standard error of measurement (SE_{Meas}) was more recently mentioned by Millman (1973). He used

$$SE_{Meas} \hat{=} \sqrt{n\bar{p}\bar{q}} \qquad [2]$$

where $\bar{p}$ is the mean item difficulty and $\bar{q} = 1 - \bar{p}$. The right-hand side of Approximation 2 is the standard deviation of a binomial distribution with possible scores of zero through n and a mean of $n\bar{p}$. Lord (1957, Equation 6) presented Approximation 2, as an equation, as the standard deviation of observed scores for one examinee. The standard error of measurement being considered in the present paper, however, is the average across examinees of all of the individual examinees' standard errors.

Note that neither Approximation 1 nor Approximation 2 uses $r_{xx'}$. Millman's approximation is the standard error of measurement of a

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 3, No. 2 Spring 1979 pp. 161–164

test composed of uncorrelated items. Lord's approximation is equal to Millman's when either $\bar{p}$ or $\bar{q}$ is equal to .76. When $.24 < \bar{p} < .76$, Approximation 1 is greater than Approximation 2. This implies that the quality of the items, their interdependence, has no effect on the measurement error of a test. The total item covariance, as expressed through a test's internal consistency reliability, would have no effect on either of these approximations.

Nonetheless, Saupe (1961) used Lord's approximation to derive an approximation to $r_{xx'}$. Saupe began with the customary equation for the standard error of measurement,

$$SE_{Meas} = s_X\sqrt{1 - r_{XX'}} \quad , \qquad [3]$$

set it equal to the right-hand side of Approximation 1, and solved for $r_{xx'}$. This yielded

$$r_{XX'} \hat{=} 1 - \frac{9n}{49s_X^2} \quad . \qquad [4]$$

Saupe also suggested an unbiased approximation,

$$r_{XX'} \hat{=} \frac{n}{n-1}\left(1 - \frac{.2n}{s_X^2}\right) \quad . \qquad [5]$$

Using the same approach as Saupe, the right-hand side of Approximation 2 could be set approximately equal to the right-hand side of Equation 3. This would yield

$$r_{XX'} \hat{=} \frac{s_X^2 - n\bar{p}\bar{q}}{s_X^2} \quad . \qquad [6]$$

Approximation 6 is $(n-1)/n$ times the value of the Kuder-Richardson (1937) Formula 21 (KR21). As such, it is essentially useless as an approximation. Not only is it consistently smaller than (positive) KR21, but also its computation is only one step short of that KR21. This is also true for the $n\bar{p}\bar{q}$—Kuder-Richardson Formula 20 (KR20) version of Approximation 6.

Empirical Demonstration

Data from 200 classroom tests were used. Initially, there were 208 tests that had been submitted to a university's test scoring service during a 2-month period, but 8 were not used for the present study because they had zero or negative KR20s. Test means, standard deviations, KR20s, and associated estimates of the standard errors of measurement (SEM20) were computed with a locally developed test analysis program. All subsequent computations were performed with the assistance of *Statistical Package for the Social Sciences* (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975).

Three approximations to SE_{Meas} were computed and compared with two criterion values. The approximations were

LORDSEM	Approximation 1
MILSEM20	Approximation 2, using $\bar{p}\bar{q}$
MILSEM21	Approximation 2, using $\bar{p}\bar{q}$

The criterion values, SEM20 and SEM21, were respectively equal to $s_X\sqrt{1-KR20}$ and $s_X\sqrt{1-KR21}$.[1] Approximations 4 and 5 were also computed and designated SAUPE1 and SAUPE2, respectively.

Descriptive statistics for the tests are summarized in Table 1. The correlations among the various approximations are presented in Table 2.

Discussion

Before discussion of the approximations, some comments on the tests themselves are

[1]An anonymous reviewer has pointed out that as KR20 and KR21 are themselves estimates of the reliability, $s_X\sqrt{1-KR20}$ and $s_X\sqrt{1-KR21}$ are also estimates. The conclusions of this paper, then, are to be understood regarding these values and not the (usually unobtainable) reliability and its associated standard error of measurement.

Table 1
Descriptive statistics for 200 tests

Statistics	Mean	Standard Deviation	Lowest	Highest
Number of items (n)	44.8	29.8	8	140
Number of examinees	90.8	104.9	6	883
Mean	30.4	20.5	4.0	102.4
Mean item difficulty ($\bar{p}$)	.68	.10	.41	.89
Mean item variance ($\bar{p}\bar{q}$)	.17	.03	.09	.23
Standard deviation	5.4	3.4	1.3	17.4
KR20	.663	.209	.025	.945
KR21	.565	.286	.539	.933
SEM20	2.5	.9	1.0	5.0
SEM21	2.8	1.0	1.1	5.6

appropriate. The requirement that KR20 be greater than zero eliminated 8 of the 208 tests originally screened. Even with this restriction, the tests varied widely in length, number of examinees, and KR20. As a group, they tended to be "easy," at least by the standards of classical measurement theory, which describes a "moderately difficult" item as one with $.40 < p < .60$.

Turning to the approximations, first note the correlations between the approximations and the criteria. Those for $r_{xx'}$ are very high, .93 to two decimal places. The correlations for SE_{Meas} are extremely high, especially those involving the Millman approximations.

Examination of the means, however, reveals the systematic errors of the approximations. All three approximations to SE_{Meas} average higher than the mean SEM20, and the means of the reliability approximations are both smaller than that of KR20. Since Approximation 1 tends to be an even greater overapproximation than Approximation 2, the Saupe approximations tend to be smaller than KR20.

The approximations involving KR21 pose some problems. It should be recalled that the original form of Approximation 2 is the only one of the three approximations which is directly related to KR21. It should also be recalled that KR21 is less than KR20 whenever the items on a test have unequal difficulties. From that point of view, KR21 is an approximation to KR20 and, for the tests in the present study, a relatively poor one.

To an extent, the findings of previous investigators have been similar. They (especially Lord, 1959, and Swineford, 1959) have tended to use tests with $\bar{p}$ close to .5; and they have reported high correlations and close, but biased, approximations. The difference is that the present study is focused on that systematic bias. This bias increases as $\bar{p}$ departs from .5. An inaccuracy with that characteristic is most serious in many measures of typical behavior, criterion-referenced educational tests, mastery tests, and other scales which tend to have high values of $\bar{p}$ for appropriate examinees.

Conclusions

Approximations serve three purposes. Because of their computational simplicity, they may take the place of more accurate, but more computationally involved, statistics. Also, approximations may be used as rough guesses to statistics, either before scoring a test or upon receiving a computer-generated test analysis. Finally, the approximation may give insight into the nature of the statistic.

For the first two purposes, the Lord (1959), Millman (1973), and Saupe (1961) approxima-

Table 2
Correlations among Approximations and Actual Values of SE_{Meas} and $r_{XX'}$

		Std. Error of Measurement					Reliability			
		Approximations			Criteria		Approx.		Criteria	
		1	2	3	4	5	6	7	8	9
1	LORDSEM	--	.974	.978	.976	.976	.648	.619	.682	.657
2	MILSEM20	.974	--	.995	.998	.990	.720	.693	.693	.675
3	MILSEM21	.978	.995	--	.996	.999	.692	.665	.669	.640
4	SEM20	.976	.998	.996	--	.994	.691	.664	.661	.643
5	SEM21	.976	.990	.999	.994	--	.664	.636	.639	.609
6	SAUPE1	.648	.720	.692	.691	.664	--	.999	.929	.927
7	SAUPE2	.619	.693	.665	.664	.636	.999	--	.925	.926
8	KR20	.682	.693	.669	.661	.639	.929	.925	--	.982
9	KR21	.657	.675	.640	.643	.609	.927	.926	.982	--
Mean		2.71	.260	2.89	2.51	2.81	.590	.570	.663	.565
Std. dev.		.94	.95	1.06	.92	1.03	.270	.298	.209	.286

tions appear to be useful. They are close to their respective criterion statistics. (The mean error was about 5%.) The approximations ranked the tests in the present study in almost exactly the same order that the criteria did. If one observes Lord's (1959) precautions limiting his, and thus Saupe's, application to tests with mean difficulties within .15 of .5, then these approximations are excellent.

It is the interpretative value of these approximations that is most seriously open to question. Examination of Approximations 1 and 2 would lead to the conclusion that only the number of items in a test, and not their quality, matters. If Approximation 1 were always accurate, then only length would matter. If Approximation 2 were always accurate, only test length and item difficulty would be sources of improving the precision of a test. However, there is also obviously an "item quality," or at least an "item interdependence," component. The influence of this component should not be overlooked by test constructors or test users.

References

Kuder, G. F., & Richardson, N. W. The theory of estimation of test reliability. *Psychometrika*, 1937, *2*, 151–160.

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

Lord, F. M. Tests of the same length do have the same standard error of measurement. *Educational and Psychological Measurement*, 1959, *19*, 233–239.

McMorris, R. F. Evidence on the quality of several approximations for commonly used measurement statistics. *Journal of Educational Measurement*, 1972, *9*, 113–122.

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

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *Statistical package for the social sciences* (2nd ed.). New York: McGraw-Hill, 1975.

Saupe, J. L. Some useful estimates of the Kuder-Richardson Formula Number 20 reliability coefficient. *Educational and Psychological Measurement*, 1961, *21*, 63–71.

Swineford, F. Note on "Tests of the same length do have the same standard error of measurement." *Educational and Psychological Measurement*, 1959, *19*, 241–242.

Author's Address

Send requests for reprints or further information to David J. Kleinke, Skytop Office Building, Syracuse University, Syracuse, NY 13210.