Evaluation of Small-Sample Statistics That Test Whether Variables Measure the Same Trait

Jeffrey Lee Rasmussen
Indiana University—Purdue University at Indianapolis

This study investigated the performance of five small-sample statistics (Lord's, Kristof's, McNemar's, Forsyth and Feldt's, and Braden's) that test whether two variables measure the same trait except for measurement error. The conservative Type I error rates of the Lord and Kristof procedures and the liberal error rates of the McNemar, Forsyth and Feldt, and Braden procedures were corrected by determining appropriate critical values. Power comparisons were then made at the fixed α levels. In general, the McNemar statistic was shown to be the most powerful. Finally, the effects of non-normality were investigated, and it was demonstrated that the Braden technique became very liberal, whereas the other statistics tended to be somewhat liberal at the .01 significance level and reasonably robust at the .05 level. *Index terms: Disattenuated measures, Measurement error, Monte carlo simulation, Non-normality, Small-sample statistics, Type I error and power rates.*

It is often important to test whether two variables measure the same trait. For example, a researcher might want to assess whether a "creativity" test and a "fluidity of thought" test measure the same trait. Due to measurement error, the correlation between the two variables would not be expected to be perfect even if they do measure the same trait. The joint reliability of the measures puts an upper bound on the magnitude of the correlation between the measures (Ghiselli, Campbell, & Zedeck, 1981; Nunnally, 1978).

Measurement error is said to attenuate the correlation between the two scores. A statistic, called the disattenuated correlation, corrects for this attenuation. Its formula is

$$r_d = \frac{r_{xy}}{(r_{xx}r_{yy})^{1/2}} \quad , \tag{1}$$

where r_{xy} is the raw score correlation,
r_{xx} is the reliability of the X scores, and
r_{yy} is the reliability of the Y scores.

The disattenuated correlation represents the correlation between two variables from which measurement error has been eliminated; that is, it measures the correlation of the true scores.

Unfortunately, the disattenuated correlation is not distributed as a Pearson product-moment correlation coefficient. For example, values greater than 1.0 are possible. Hence the disattenuated correlation cannot

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 12, No. 2, June 1988, pp. 177–187
© Copyright 1988 Applied Psychological Measurement Inc.
0146-6216/88/020177-11$1.80

be tested for significance in the same fashion as a Pearson correlation (Cureton, 1936; DuBois, 1957, 1965; Kelley, 1947; Shen, 1924). There are a number of formulas that are appropriate for testing whether two variables measure the same trait which are useful for large samples (Bobko & Rieck, 1980; Forsyth & Feldt, 1969; Jöreskog, 1971; Lord, 1957).

For smaller sample sizes, Rae (1982) evaluated four statistics (Forsyth & Feldt, 1970; Kristof, 1973; Lord, 1973; McNemar, 1958) that test whether two variables measure the same attribute. Rae used monte carlo methods to compare these four statistics in terms of Type I error rates and power. He found that the Lord and the Kristof procedures tended to be conservative, whereas the McNemar and the Forsyth and Feldt procedures were often overly liberal. In terms of the power of the tests, McNemar's tended to be the most powerful, followed by Forsyth and Feldt's, Kristof's, and Lord's. The power comparisons, however, were of limited value as they were not made at a fixed α level. That is, it is difficult to interpret the fact that Lord's statistic shows a power of, say, 15.7% when α is at the 3.4% level, relative to McNemar's statistic showing a power of 26.5% when α is at the 6.0% level. Power comparisons only make sense when all other relevant factors (e.g., sample size, falsity of the null hypothesis, α level) are held constant (see Levine & Dunlap, 1983).

The Statistics

Lord's (1973) statistic, L, is based upon work by Villegas (1964). Lord's statistic is determined by first calculating

$$\mathbb{Q} = \mathbb{A} - F_p \mathbb{W} \quad , \tag{2}$$

where $\mathbb{A}$ denotes a 2 × 2 sum of squares cross-products (SSCP) between matrix,
$\mathbb{W}$ denotes a 2 × 2 SSCP within matrix, and
F_p denotes the F critical value at N and N degrees of freedom for the appropriate $p = .01/.05$ significance level.

If the determinant and the two major diagonal elements of the resulting matrix, $\mathbb{Q}$, are positive, then the statistic is declared significant at probability p. In addition to a bivariate normality assumption, Lord's procedure assumes that there are no practice effects. That is, if the reliability is calculated using a test-retest procedure, there cannot be any population mean difference between the test and retest scores.

Kristof (1973) presented a statistic that takes different forms depending upon which of three cases of assumptions the user is willing to make. To be consistent with Rae's (1982) study, the present study investigated Kristof's Case II statistic with data where there is not a practice effect. Kristof's statistic, K, is set as the smallest eigenvalue of the matrix $\mathbb{J}$, where

$$\mathbb{J} = \mathbb{V}_t \mathbb{V}_d^{-1} \quad , \tag{3}$$

$\mathbb{V}_t$ is the variance-covariance matrix of the total scores ($X_1 + X_2$, $Y_1 + Y_2$), and $\mathbb{V}_d^{-1}$ is the inverse of the variance-covariance matrix of the difference scores ($X_1 - X_2$, $Y_1 - Y_2$). The (X_1,Y_1) and (X_2,Y_2) scores are the test-retest or parallel-halves scores for X and Y. Kristof's K is evaluated for significance by comparing it with the appropriate critical values of the F distribution with $N - 1$ degrees of freedom.

McNemar's (1958) statistic, M, takes the form of a mean square residual expressed in terms of correlations and reliabilities,

$$M = \frac{2 + r_{xx} + r_{yy} - r_{x1y1} - r_{x1y2} - r_{x2y1} - r_{x2y2}}{2 - r_{xx} - r_{yy}} \quad . \tag{4}$$

The statistic is declared significant if it exceeds the appropriate .01/.05 critical values of the F distribution with ($N - 1$) and 2($N - 1$) degrees of freedom. Essentially, McNemar's procedure assesses whether there is an examinee × test interaction, that is, whether examinees perform differently depending upon which of two tests they are administered. If the two tests have identical true scores, then there will be an insignificant interaction. McNemar's statistic assumes that the X and Y variables have equal reliability. In the case of unequal reliabilities it is overly liberal.

Forsyth and Feldt's (1970) statistic, FF, is equal to McNemar's statistic when the latter is divided by a value

$$C = 1 + \frac{2[\bar{r} - (r_{xx} r_{yy})^{1/2}]}{1 - \bar{r}} \quad , \qquad (5)$$

where $\bar{r} = (r_{xx} + r_{yy})/2$. They made this modification to correct for the difference between the arithmetic and the geometric means of the reliabilities; note that when the reliabilities are equal, then the C term reduces to 1. Their results indicated that the modified statistic performed better than the McNemar version. It was, however, still overly liberal. Additionally, as Forsyth and Feldt (1970) pointed out, because M and FF do not have the same α level when the reliabilities are unequal, it is not possible to reasonably compare the two statistics except in the limited case of equal reliabilities.

Braden's (1986) statistic B is calculated by comparing the observed correlation with the maximum value it can achieve, which is the square root of the product of the population joint reliability, $\rho_{xxyy} = [(\rho_{xx})(\rho_{yy})]^{1/2}$. The closer the observed value is to the maximum, the more likely the two variables measure the same trait. The observed and maximum values are calculated and evaluated for significance using Fisher's z test. That is,

$$B = (Z_{xy} - Z_{xxyy})/\mathrm{SE} \quad , \qquad (6)$$

where $Z_{xy} = .5 \log [(1 + r_{xy})/(1 - r_{xy})]$,
$Z_{xxyy} = .5 \log [(1 + \rho_{xxyy})/(1 - \rho_{xxyy})]$, and
$\mathrm{SE} = 1/(N - 3)^{1/2}$.

Braden's statistic assumes that the population reliabilities, and therefore the maximum value, are known from previous normative data.

The present study expanded upon Rae's (1982) study. First, it included the additional statistic proposed by Braden (1986). Second, it determined the appropriate critical values that hold α constant. This allows for power comparisons to be made at fixed α levels. Finally, it investigated the effects of non-normality on the five statistics. Studies that have investigated the effect of non-normality on the Type I error rate for the regular Pearson correlation coefficient have found it to be relatively insensitive to violation of the normality assumption. For example, Edgell and Noon (1984) and Havlicek and Peterson (1977) found that the actual Type I error rates of tests of significance of r are reasonably close to their nominal levels. As not all variables are normally distributed, the present study also investigated the effect of non-normality on these statistics.

Method

The present study was implemented in four phases. In Phase 1 a FORTRAN program was written to estimate the actual Type I error rates of the five statistics using their authors' critical values. This was done in order to compare the present study with Rae's (1982) results and to evaluate the Type I error rate of Braden's statistic. In Phase 2 the program was modified to estimate the appropriate critical values that would maintain the statistics at the .01 and .05 significance levels. In Phase 3 the critical values derived in Phase 2 were used to compute Type I error and power rates of the five statistics. Phase 4 investigated the effect of sampling from two non-normal distributions on the Type I error rates.

Type I Error Rates Using Authors' Critical Values

In Phase 1 a FORTRAN program was written to estimate the Type I error rates of Lord's L, McNemar's M, Forsyth and Feldt's FF, Kristof's K, and Braden's B statistics. Sample sizes of $N = 25$ and 50, significance levels of $p = .05$ and .01, and reliability combinations of $r_{xx}/r_{yy} = .8/.8$, .6/.6, .8/.9, .8/.7, .7/.9, .6/.8, .6/.9, and .5/.9 were simulated. These N, p, and r_{xx}/r_{yy} values were crossed to yield a 2 × 2 × 8 design with 32 conditions.

For each of the 32 conditions, the program generated five scores: T, E_{11}, E_{12}, E_{21}, and E_{22}. From these scores, four linear composites were created:

$$X_1 = T + E_{11},$$
$$X_2 = T + E_{12},$$
$$Y_1 = T + E_{21},$$
$$Y_2 = T + E_{22}.$$

The scores were generated such that the population correlations of the T and E scores (e.g., T and E_{11}, E_{11} and E_{12}) were 0.

The scores were generated from a normal population with mean = 0 and standard deviation = 1 using the Ahren and Dieter normal deviate pseudorandom number algorithm (cited in Lehman, 1977). The reliabilities between X_1 and X_2, and between Y_1 and Y_2, were generated using an algorithm by Knapp and Swoyer (1967).

For each of the sample size × reliability combinations, 1,000 experiments were carried out. The program calculated the five statistics and evaluated the obtained statistics for significance. The Type I error rate for each of the significance levels was calculated as the number of the test statistics declared significant divided by 1,000.

Determination of Appropriate Critical Values

The results from Phase 1, which are presented more fully in the Results section, indicated that none of the tests had accurate Type I error rates when they were calculated according to their authors' recommendations. Consequently, Phase 2 was carried out to determine the appropriate critical values. Due to the nature of the statistics, the method for estimating the appropriate critical values was different for the Lord statistic than for the other four statistics.

Evaluation of the significance of Lord's statistic is not done by comparison of the obtained statistic to some critical value; rather, the critical value F_p is incorporated into the calculation formula, and significance is declared if the resulting matrix is positive definite. It was therefore necessary to estimate the appropriate values of F_p using an iterative approach. These values were estimated by bracketing the .01 and .05 significance levels. First, the F_p values that Lord recommended were used and evaluated for significance by carrying out 10,000 replications at each of the sample size and reliability combinations. The values of F_p were decremented by .01 and 10,000 more replications were carried out. In all, 20 decrements were carried out for a total of 3.2 million replications.

The obtained Type I error rates for the 20 decrements were plotted and the appropriate .01 and .05 critical values were interpolated from the plot. As the relationship between the decrements and the Type I error rates was approximately linear, the critical values were estimated by linear (as opposed to curvilinear) interpolation. Visual inspection indicated that the critical values were independent of the reliability combinations (i.e., there was not a discernible positive or negative covariation of the critical values with the reliability conditions), and therefore they were averaged over the r_{xx}/r_{yy} condition. (Because Phase 3 would serve as a check on the appropriateness of using the averaged values, the visual inspection was deemed a sufficient check.)

For the other statistics, the Type I error rate was estimated by generating 2,000 replications of the 16 sample size × reliability conditions. The test statistics were calculated for each replication. Each of the test statistics was rank ordered and the 95th and the 99th percentile values were obtained. This procedure was carried out five times and the 99th and 95th percentile values were averaged over the five repetitions. Hence the critical values for the 32 conditions of the Kristof, Forsyth and Feldt, Braden, and McNemar statistics were based on 10,000 replications.

The critical values for the Braden and Kristof procedures were apparently unrelated to the reliability condition, and were therefore averaged across this condition. The critical values of the McNemar and the Forsyth and Feldt statistics were positively related to the increasing disparity in the population reliabilities. Thus, in the subsequent phases, the ''averaged across reliabilities'' values were used for the Lord, Braden, and Kristof procedures, and the unaveraged values were used for the Forsyth and Feldt test and the McNemar test.

Type I Error Rates and Power Using Appropriate Critical Values

In Phase 3 the appropriate critical values determined in the previous phase were used to calculate the Type I error rate and power of the statistics. The calculation of Type I error rate represented a check on the accuracy of the critical values determined in Phase 2. The Type I error rates were calculated in the same fashion as in Phase 1, with the exception of the use of the new critical values. The power of the statistics was evaluated by generating 1,000 replications in which the disattenuated correlation between X and Y was less than 1.0. Disattenuated correlations of .95, .90, and .85 were simulated using the Knapp and Swoyer (1967) algorithm.

Type I Error Rates Under Assumption Violation

Finally, in Phase 4 of the experiment, non-normal data were simulated to evaluate the effect of assumption violation on statistics. Non-normal data were simulated by sampling from an exponential distribution and a lognormal distribution, which are commonly investigated distributions (e.g., Blair & Higgins, 1985). The exponential distribution was generated by taking the logs of random uniform deviates for the T and E scores. The lognormal distribution was generated by taking the exponents of normal deviates.

Results

Type I Error Rates Using Authors' Critical Values

The Phase 1 results are given in Table 1. The Type I error rates for the Lord and Kristof procedures are conservative, whereas they are generally liberal for the McNemar and the Forsyth and Feldt procedures; these results are similar to those of Rae (1982). The Braden statistic is very liberal, with actual .01 Type I error rates in the .21 to .49 range and actual .05 rates in the .45 to .75 range.

Determination of Appropriate Critical Values

The critical values derived from the monte carlo techniques of Phase 2 are given in Tables 2 and 3. The values for the Lord and Kristof procedure are, as expected, less than the critical values recommended

Table 1
Type I Error Rates for the Lord (*L*), Kristof (*K*), McNemar (*M*), Forsyth and Feldt (*FF*), and Braden (*B*) Procedures as a Function of the Reliabilities of the Two Measures, for Two Sample Sizes (Decimal Points Omitted)

Relia-bilities	*p*=.01					*p*=.05				
	L	*K*	*M*	*FF*	*B*	*L*	*K*	*M*	*FF*	*B*
N=25										
.8 .8	006	009	011	010	218	031	037	055	052	458
.6 .6	006	009	012	011	211	031	037	056	051	461
.8 .9	005	008	013	012	221	032	039	061	056	459
.8 .7	006	008	012	011	212	031	038	057	051	449
.7 .9	007	009	021	016	219	034	041	080	065	461
.6 .8	006	008	015	013	211	032	039	071	058	457
.6 .9	006	009	030	020	216	031	039	103	074	451
.5 .9	008	011	045	026	211	037	043	144	086	453
N=50										
.8 .8	008	009	011	011	488	036	043	053	050	745
.6 .6	008	009	013	012	490	035	040	053	048	742
.8 .9	006	007	014	012	487	034	041	061	054	739
.8 .7	007	008	012	010	491	037	041	058	051	748
.7 .9	007	008	022	016	491	037	043	087	062	746
.6 .8	005	006	016	011	489	031	037	070	053	735
.6 .9	007	008	039	022	489	038	043	129	072	744
.5 .9	007	008	063	024	487	035	040	186	078	741

by the respective authors. The values for the Braden statistic are much larger than the *z*-score critical values he recommended. For the McNemar and the Forsyth and Feldt statistics, as Table 3 demonstrates, the monte carlo values increase with an increasing disparity in the reliabilities. This should serve to counteract the liberal drift of these statistics displayed in Table 1 using the authors' critical values.

Type I Error Rates and Power Using Appropriate Critical Values

The results for Phase 3 are given in Tables 4 and 5. Table 4 presents the Type I error rates using the monte carlo values derived in Phase 3. The Type I error rates are greatly improved from those in Table 1. The standard errors were calculated as SE $= [(a - a^2)/n]^{1/2}$, where a is the α level (.01 or .05) and $n = 10{,}000$, the number of replications (see Zwick, 1986, p. 176). Values outside a range of ±2 standard errors are flagged in Table 4. For example, the 2 standard error range was .0456 to .0544 for the .05 significance level (the values in Table 4 are rounded to the third significant figure; the flagging, however, was carried out on the values rounded to the fourth significant figure). Of the 160 Type I error rates, 12 are outside the ±2 SE boundary. Because the 160 values are not strictly independent (i.e., the .01 and .05 values for a given comparison are related), and because there does not appear to be a

Table 2
Monte Carlo Critical Values for Averaged Statistics

Statistic	*p*=.01		*p*=.05	
	N=25	*N*=50	*N*=25	*N*=50
Lord	2.60	1.94	1.96	1.60
Kristof	2.60	1.96	1.98	1.61
Braden	3.90	4.67	3.17	3.97

Table 3
Monte Carlo Critical Values for Unaveraged Statistics as a Function of the Reliabilities of the Two Measures, for Two Sample Sizes

Relia-	p=.01		p=.05	
bilities	N=25	N=50	N=25	N=50
McNemar				
.8 .8	2.22	1.75	1.78	1.49
.6 .6	2.21	1.77	1.74	1.49
.8 .9	2.29	1.79	1.80	1.51
.8 .7	2.29	1.79	1.79	1.51
.7 .9	2.41	1.86	1.89	1.59
.6 .8	2.39	1.83	1.84	1.56
.6 .9	2.54	1.98	2.01	1.68
.5 .9	2.73	2.13	2.12	1.79
Forsyth and Feldt				
.8 .8	2.21	1.75	1.76	1.48
.6 .6	2.19	1.77	1.72	1.48
.8 .9	2.27	1.77	1.78	1.48
.8 .7	2.27	1.77	1.76	1.49
.7 .9	2.37	1.81	1.84	1.52
.6 .8	2.36	1.79	1.78	1.51
.6 .9	2.43	1.87	1.87	1.56
.5 .9	2.54	1.93	1.92	1.59

discernible pattern of values falling outside the boundaries, it seems reasonable to conclude that the deviations of the obtained Type I error rates from their nominal levels are due to sampling variation.

Table 5 presents the power of the statistics at the fixed α levels. To conserve space, only the results for the N = 50 condition are given; the N = 25 condition showed the same pattern of results. In general the results indicate that the McNemar M and the Forsyth and Feldt FF procedures show the greatest power, followed by the Lord and Kristof procedures. The power of the Braden procedure tends to fall far short of the other four approaches.

The average power for the McNemar statistic is 34.7% at the .01 level and 54.1% at the .05 level. For the Forsyth and Feldt procedure, the average power rates are 32.4% and 51.8% respectively. The slight power advantage is mainly attributable to the more disparate reliability conditions. (Although the tabled Type I error rates are higher for the McNemar test than for the Forsyth and Feldt test on the N = 50, reliability = .6/.9 and .5/.9 conditions, this is probably not the cause of the superior power; the McNemar test still shows a power superiority over the Forsyth and Feldt test on the N = 25, reliability = .6/.9 and .5/.9 conditions even though the tabled Type I error rates for the McNemar test are less than the Forsyth and Feldt test here.)

The results in Table 5 indicate that the Lord and Kristof procedures have very similar power curves. The average absolute difference for the procedures is .005 for the α = .01 and .004 for the α = .05 conditions.

Type I Error Rates Under Assumption Violation

The results from Phase 4 are presented in Table 6. Once again to conserve space, only the results from the N = 50 condition are given. The results indicate that the Braden test is highly liberal on both the exponential and lognormal distributions, with actual error rates in the .04 to .24 range for the nominal

Table 4
Type I Error Rates Using Monte Carlo Critical Values
for the Lord (*L*), Kristof (*K*), McNemar (*M*),
Forsyth and Feldt (*FF*), and Braden (*B*) Procedures
as a Function of the Reliabilities of the Two Measures,
for Two Sample Sizes (Decimal Points Omitted)

Relia-	*p*=.01					*p*=.05				
bilities	*L*	*K*	*M*	*FF*	*B*	*L*	*K*	*M*	*FF*	*B*
N=25										
.8 .8	010	010	010	010	013+	048	047	048	049	054
.6 .6	010	011	012	012	009	046	045–	050	052	053
.8 .9	012	011	010	010	009	052	050	050	049	056+
.8 .7	010	010	009	009	010	046	047	047	047	050
.7 .9	011	011	011	010	011	049	050	052	049	051
.6 .8	010	010	008	008	010	050	050	054	052	050
.6 .9	010	011	010	010	009	048	050	052	054	052
.5 .9	012	013+	012	012	009	054	056+	056+	057+	050
N=50										
.8 .8	010	010	010	010	010	050	051	049	050	053
.6 .6	010	011	010	009	010	047	047	050	048	048
.8 .9	010	009	009	009	011	048	050	052	054	050
.8 .7	009	010	008	008	010	048	049	047	047	050
.7 .9	010	011	010	009	010	051	051	050	050	052
.6 .8	007–	008–	009	009	009	044–	046	046	045–	047
.6 .9	010	010	011	010	008–	050	050	050	047	046
.5 .9	009	010	010	010	011	047	048	049	045	051

Note. "–" indicates that the rejection rate is 2 or more standard errors below the Type I error rate; "+" indicates that the rejection rate is 2 or more standard errors above the Type I rate.

.01 level and actual error rates in the .11 to .33 range for the nominal .05 level. The error rates for the remaining tests tend to be overly liberal for the .01 significance level, especially on the disparate reliability conditions. For the .05 level the remaining tests are fairly close to the nominal level, although the Lord and Kristof procedures tend to be more conservative than the McNemar and the Forsyth and Feldt procedures.

Discussion

The present study demonstrated a number of aspects of the five statistics studied. First, the results indicated that the Braden procedure, like the other procedures, did not have stable and acceptable .01 and .05 Type I error rates. The fact that the Braden procedure performed so poorly can probably be traced to Braden's questionable use of Fisher's test to assess the difference between the observed correlation and the joint reliability. Given that there is no apparent mathematical justification for assuming that the joint reliabilities are distributed as Pearson correlations, the erratic performance of Fisher's test is not unexpected.

Second, it was demonstrated that accurate critical values could be estimated by monte carlo procedures, and that these critical values could be used to obtain comparative power rates. It was shown that when the reliabilities of the statistics are known, the McNemar test was the most powerful statistic for detecting that two variables measured the same trait.

Third, the Lord and Kristof procedures were shown to have very similar power curves. From the description of the two statistics, their formulas, and their evaluation of significance, they appear to be

Table 5
Power of Statistics Using Monte Carlo Critical Values for the Lord (*L*), Kristof (*K*), McNemar (*M*), Forsyth and Feldt (*FF*), and Braden (*B*) Procedures as a Function of the Reliabilities of the Two Measures and the Correlation (*r*) of *x* and *y* (Decimal Points Omitted)

Relia-		*p*=.01					*p*=.05				
bilities	*r*	*L*	*K*	*M*	*FF*	*B*	*L*	*K*	*M*	*FF*	*B*
.8 .8	.95	116	120	182	172	053	314	321	381	385	194
	.90	363	360	505	491	132	622	622	744	747	331
	.85	646	649	814	808	292	853	853	932	931	570
.6 .6	.95	031	033	039	032	021	127	123	154	151	085
	.90	084	079	103	095	031	233	232	272	268	133
	.85	123	125	189	173	072	334	342	438	433	223
.8 .9	.95	204	210	307	302	097	459	455	575	571	220
	.90	597	602	761	746	240	827	823	903	896	496
	.85	870	858	942	939	478	959	959	988	989	749
.8 .7	.95	070	080	103	106	035	229	229	276	274	109
	.90	227	228	319	312	110	473	465	595	588	283
	.85	462	473	625	608	203	713	703	829	820	449
.7 .9	.95	090	094	170	145	042	288	287	369	347	169
	.90	304	308	460	413	151	578	569	714	680	361
	.85	589	591	769	706	291	813	813	916	891	563
.6 .8	.95	054	052	068	066	019	169	173	226	210	102
	.90	138	144	212	197	067	342	332	439	416	207
	.85	255	263	421	382	116	524	519	655	616	330
.6 .9	.95	058	063	083	071	038	167	174	223	200	132
	.90	188	193	263	224	104	408	413	533	459	261
	.85	346	350	490	421	172	595	588	721	646	401
.5 .9	.95	041	042	051	044	024	138	141	165	140	103
	.90	115	111	150	123	056	284	293	379	314	188
	.85	180	179	303	216	106	429	424	566	477	287

very different statistics. Moreover, Rae's (1982) findings that they had different Type I error rates and power also indicated that they were different. The present study indicates that the differences are probably superficial—at least for the Case II situation.

Finally, the present study showed that, with the exception of the Braden test, these statistics tended to be reasonably robust to assumption violation at the .05 level and somewhat liberal at the .01 level. In general, the tests became more liberal with increasing disparity in the variabilities. (Although some pilot comparisons of the power functions of the tests were carried out, they are not presented because the α levels were different for the different tests. The pilot study, however, did indicate that the McNemar and the Forsyth and Feldt procedures were the most powerful, followed by the Lord and Kristof tests and then the Braden test. It was not possible to assess the impact of non-normality—relative to normality—on the tests due to the method of data generation; see Levine & Dunlap, 1982, 1983.)

There are some areas of further study that may prove fruitful:

1. The present study indicated that it was necessary to have different critical values for the different reliability combinations with the McNemar and the Forsyth and Feldt procedures. This would present a distinct problem for a researcher who wanted to use these statistics. One solution would be to incorporate the reliabilities into the formula, much as Forsyth and Feldt attempted to do by calculating the *C* value. The present study, as well as that of Rae (1982), showed that this value did not correct enough for the disparity in the reliabilities. Further refinement of the correction value would be useful. The corrected formula should be tested under a wider range of reliability combinations—for example, $r_{xx} = .4 r_{yy} = .9$.

Table 6
Type I Error Rates Under Non-normality for the Lord (*L*), Kristof (*K*), McNemar (*M*), Forsyth and Feldt (*FF*), and Braden (*B*) Procedures, as a Function of the Reliabilities of the Two Measures
(Decimal Points Omitted)

Relia-	p=.01					p=.05				
bilities	*L*	*K*	*M*	*FF*	*B*	*L*	*K*	*M*	*FF*	*B*
Exponential Distribution										
.8 .8	009	010	014	013	064	052	053	053	053	153
.6 .6	009	010	010	009	040	041	041	050	046	113
.8 .9	011	012	014	013	073	052	052	055	056	159
.8 .7	008	009	012	011	060	048	047	051	050	141
.7 .9	012	011	016	014	065	052	051	054	055	152
.6 .8	012	012	017	015	053	053	052	054	052	132
.6 .9	015	016	118	017	060	056	055	058	055	140
.5 .9	016	017	019	017	045	056	057	059	057	121
Lognormal Distribution										
.8 .8	010	010	018	016	228	039	039	049	046	326
.6 .6	008	008	015	013	155	035	035	050	042	261
.8 .9	012	013	018	018	235	042	043	053	050	330
.8 .7	011	012	019	017	209	040	040	052	048	311
.7 .9	014	014	023	020	216	044	045	056	051	320
.6 .8	012	013	020	018	193	044	043	053	047	291
.6 .9	016	017	022	020	195	048	048	055	052	292
.5 .9	017	017	021	020	167	049	050	054	049	265

2. It would be helpful to develop more feasible means to test for significance. The critical values given in the present study may be used for the limited number of conditions investigated. For other sample sizes and reliability conditions, however, they are not accurate. Appropriate critical values could be discovered by either (1) identifying the underlying sampling distribution of the statistics in terms of some formula; (2) finding a transformation to a known distribution, such as *F* or *z*; or (3) using a tailor-made monte carlo algorithm that would generate the critical values for a given sample size and reliability condition.
3. The effects of non-normality on the five statistics could be further investigated. Different non-normal distributions could be used, and the effect of non-normality on power could be studied. If non-normality does reduce the power of these tests—as might be predicted on the basis of studies concerning the effect of power on Pearson's *r* (Edgell & Noon, 1984)—it might be interesting to investigate whether it is possible to detect such non-normality on these small sample sizes, or to select a transformation that would result in normality (D'Agostino, 1971; Dixon & Massey, 1969; Games, 1983, 1984; Rasmussen, 1985).
4. A comparison of these small-sample statistics with the techniques recommended for larger samples (Bobko & Rieck, 1980; Forsyth & Feldt, 1969; Jöreskog, 1971; Lord, 1957) might prove useful. Although the extant literature (e.g., Forsyth & Feldt, 1969; Rae, 1982) has categorized the test procedures into small-sample versus large-sample procedures, there has been no empirical justification for such a dichotomization.

References

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples *t* test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, *97*, 119–128.

Bobko, P., & Rieck, A. (1980). Large sample estimators for standard errors of functions of correlation coefficients. *Applied Psychological Measurement, 4,* 385–398.

Braden, J. P. (1986). Testing correlations between similar measures in small samples. *Educational and Psychological Measurement, 46,* 143–148.

Cureton, E. E. (1936). On certain estimated correlation functions and their standard errors. *Journal of Experimental Education, 4,* 252–264.

D'Agostino, R. B. (1971). An omnibus test of normality for moderate and large size samples. *Biometrika, 58,* 341–347.

Dixon, W. J., & Massey, F. J. (1969). *Introduction to statistical analysis* (3rd ed.). New York: McGraw-Hill.

DuBois, P. H. (1957). *Multivariate correlational analysis.* New York: Harper & Brothers.

DuBois, P. H. (1965). *An introduction to psychological statistics.* New York: Harper & Row.

Edgell, S. E., & Noon, S. M. (1984). Effect of violation of normality on the *t* test of the correlation coefficient. *Psychological Bulletin, 95,* 567–583.

Forsyth, R. A., & Feldt, L. S. (1969). An investigation of empirical sampling distributions of correlation coefficients corrected for attenuation. *Educational and Psychological Measurement, 29,* 61–71.

Forsyth, R. A., & Feldt, L. S. (1970). Some theoretical and empirical results related to McNemar's test that the population correlation coefficient corrected for attenuation equals 1.0. *American Educational Research Journal, 7,* 197–207.

Games, P. A. (1983). Curvilinear transformations of the dependent variable. *Psychological Bulletin, 93,* 382–387.

Games, P. A. (1984). Data transformation, power, and skew: A rebuttal to Levine and Dunlap. *Psychological Bulletin, 95,* 345–347.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences.* San Francisco: Freeman.

Havlicek, L. L., & Peterson, N. L. (1977). Effect of the violation of assumptions upon significance levels of the Pearson *r*. *Psychological Bulletin, 84,* 373–377.

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36,* 109–133.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge MA: Harvard University Press.

Knapp, T. H., & Swoyer, V. H. (1967). Some empirical results concerning the power of Bartlett's test of the significance of a correlation matrix. *American Education Research Journal, 4,* 13–17.

Kristof, W. (1973). Testing a linear relation between true scores of two measures. *Psychometrika, 38,* 101–111.

Lehman, R. S. (1977). *Computer simulation and modeling: An introduction.* Hillsdale NJ: Erlbaum.

Levine, D. W., & Dunlap, W. P. (1982). Power of the *F* test with skewed data: Should one transform or not? *Psychological Bulletin, 92,* 272–280.

Levine, D. W., & Dunlap, W. P. (1983). Data transformation, power, and skew: A rejoinder to Games. *Psychological Bulletin, 93,* 596–599.

Lord, F. M. (1957). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. *Psychometrika, 22,* 207–220.

Lord, F. M. (1973). Testing if two measuring procedures measure the same dimension. *Psychological Bulletin, 79,* 71–72.

McNemar, Q. (1958). Attenuation and interaction. *Psychometrika, 23,* 259–266.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Rae, G. (1982). A Monte Carlo comparison of small sample procedures for testing the hypothesis that two variables measure the same trait except for errors of measurement. *British Journal of Mathematical and Statistical Psychology, 35,* 228–232.

Rasmussen, J. L. (1985). Data transformation maximizing homoscedasticity and within-group normality. *Behavior Research Methods, Instruments, and Computers, 17,* 411–412.

Shen, E. (1924). The standard error of certain estimated coefficients of correlation. *Journal of Educational Psychology, 15,* 462–465.

Villegas, G. (1964). Confidence region for a linear relation. *Annals of Mathematical Statistics, 35,* 780–788.

Zwick, R. (1986). Rank and normal scores alternatives to Hotelling's *T. Multivariate Behavioral Research, 21,* 169–186.

Author's Address

Send requests for reprints or further information to Jeffrey Lee Rasmussen, Department of Psychology, Purdue School of Science, Indiana University-Purdue University at Indianapolis, Indianapolis IN 46623, U.S.A.