Psicológica (2003), 24, 133-158.

Bias in Estimation and Hypothesis Testing of Correlation

Donald W. Zimmerman*, Bruno D. Zumbo[']** and Richard H. Williams***

*Carleton University, **University of British Columbia, ***University of Miami

This study examined bias in the sample correlation coefficient, *r*, and its correction by unbiased estimators. Computer simulations revealed that the expected value of correlation coefficients in samples from a normal population is slightly less than the population correlation, ρ, and that the bias is almost eliminated by an estimator suggested by R.A. Fisher and is more completely eliminated by a related estimator recommended by Olkin and Pratt. Transformation of initial scores to ranks and calculation of the Spearman rank correlation, r_S, produces somewhat greater bias. Type I error probabilities of significance tests of zero correlation based on the Student *t* statistic and exact tests based on critical values of r_S obtained from permutations remain fairly close to the significance level for normal and several non-normal distributions. However, significance tests of non-zero values of correlation based on the *r* to *Z* transformation are grossly distorted for distributions that violate bivariate normality. Also, significance tests of non-zero values of r_S based on the *r* to *Z* transformation are distorted even for normal distributions.

This paper examines some unfamiliar properties of the Pearson product-moment correlation that have implications for research in psychology, education, and various social sciences. Some characteristics of the sampling distribution of the correlation coefficient, originally discovered by R.A. Fisher (1915), were largely ignored throughout most of the 20th century, even though correlation is routinely employed in many kinds of research in these disciplines. It is known that the sample correlation coefficient is a biased estimator of the population correlation, but in practice researchers rarely recognize the bias and attempt to correct for it.

[']Send *correspondence* to: Professor Bruno D. Zumbo. University of British Columbia. Scarfe Building, 2125 Main Mall. Department of ECPS. Vancouver, B.C. CANADA V6T 1Z4. e-mail: bruno.zumbo@ubc.ca Phone: (604) 822-1931. Fax: (604) 822-3302.

There are other gaps in the information available to psychologists and others about properties of correlation. Although the so-called *r* to *Z* transformation is frequently used in correlation studies, relatively little is known about the Type I error probabilities and power of significance tests associated with this transformation, especially when bivariate normality is violated. Furthermore, not much is known about how properties of significance tests of correlation, based on the Student *t* test and on the Fisher *r* to *Z* transformation, extend to the Spearman rank-order correlation method. For problems with bias in correlation in the context of tests and measurements, see Muchinsky (1996) and Zimmerman and Williams (1997). The present paper examines these issues and presents results of computer simulations in an attempt to close some of the gaps.

The Sample Correlation Coefficient as a Biased Estimator of the Population Correlation

The sample correlation coefficient, *r*, is a biased estimator of the population correlation coefficient, ρ, for normal populations. It is not widely recognized among researchers that this bias can be as much as .03 or .04 under some realistic conditions and that a simple correction formula is available and easy to use in practice. This discrepancy may not be crucial if one is simply investigating whether or not a correlation exists. However, if one is concerned with an accurate estimate of the magnitude of a non-zero correlation in test and measurement procedures, then the discrepancy may be of concern.

Fisher (1915) proved that the expected value of correlation coefficients based on random sampling from a normal population is approximately $E[r] = \rho - \rho(1-\rho^2)/2n,$ and that a more exact result is given by an infinite series containing terms of smaller magnitude. Solving this equation for ρ provides an approximately unbiased estimator of the population correlation,

$$\hat{\rho} = r\left[1 + \frac{(1-r^2)}{2n}\right], \qquad (1)$$

which we shall call the Fisher approximate unbiased estimator. Further discussion of its properties can be found in Fisher (1915), Kenny and Keeping (1951), and Sawkins (1944). Later, Olkin and Pratt (1958) recommended using $\hat{\rho} = r\left[1 + (1-r^2)/2(n-3)\right]$ as a more nearly unbiased estimator of ρ.

From the above equations, it is clear that the bias, $E[r]-\rho$, decreases as sample size increases and that it is zero when the population correlation is zero. For n = 10 or n = 20, it is of the order .01 or .02 when the correlation is about .20 or .30, and about .03 when the correlation is about .50 or .60. Differentiating $\rho(1-\rho^2)/2n$ with respect to ρ, setting the result equal to zero, and solving for ρ, shows that .577 and −.577 are the values for which the bias is a maximum. The bias depends on n, while the values .577 and −.577 are independent of n.

It should be emphasized that this bias is a property of the mean of sample correlation coefficients and is distinct from the instability in the variance of sample correlations near 1.00 that led Fisher to introduce the so-called *r* to *Z* transformation. Simulations using the high speed computers available today, with hundreds of thousands of iterations, make it possible to investigate this bias with greater precision than formerly, not only for scores but also for ranks assigned by the Spearman rank correlation method.

Transformation of Sample Correlation Coefficients to Stabilize Variance

In order to stabilize the variance of the sampling distribution of correlation coefficients, Fisher also introduced the *r* to *Z* transformation,

$$Z = \frac{1}{2}\ln\left[\frac{1+r}{1-r}\right], \qquad (2)$$

where *ln* denotes the natural logarithm and *r* is the sample correlation. It is often interpreted as a non-linear transformation that normalizes the sampling distribution of *r*. Although sometimes surrounded by an aura of mystery in its applications in psychology, the formula is no more than an elementary transcendental function known as the inverse hyperbolic tangent function.

Apparently, Fisher discovered in serendipitous fashion, without a theoretical basis, that this transformation makes the variability of *r* values which are close to +1.00 or to –1.00 comparable to that of *r* values in the mid-range. At the end of his 1915 paper, Fisher had some doubts about its efficacy and wrote: "In these respects, the function … $\tanh^{-1}\rho$ is not a little attractive, but so far as I have examined it, it does not simplify the analysis, and approaches relative constancy at the expense of the constancy proportionate to the variable, which the expressions in τ exhibit" (p. 521). Later, Fisher (1921) was more optimistic, and he proved that sampling distributions of *Z* are approximately normal. Inspection of graphs of the inverse hyperbolic tangent function in calculus texts makes this result

appear reasonable. With computers available, the extensive r to Z tables in introductory statistics textbooks are unnecessary, because most computer programming languages include this function among their built-in functions. Less frequently studied, and not often included in statistical tables in textbooks, the inverse function, that is, the Z to r transformation, is

$$r = \frac{e^{Z} - e^{-Z}}{e^{Z} + e^{-Z}}, \qquad (3)$$

where e is the base of natural logarithms (for further discussion, see Charter and Larsen, 1983). In calculus, this function is known as the hyperbolic tangent function, and in statistics it is needed for finding confidence intervals for r and for averaging correlation coefficients[1].

Rank Transformations and Correlation

Another transformation of the correlation coefficient, introduced by Spearman (1904), has come to be known as the Spearman rank-order correlation. One applies this transformation, not to the correlation coefficient computed from initial scores, but rather to the scores themselves prior to the computation. It consists simply of replacing the scores of each of two variables, X and Y, by the ranks of the scores. This uncomplicated procedure has been obscured somewhat in the literature by formulas intended to simplify calculations. If scores on each of two variables, X and Y, are separately converted to ranks, and if a Pearson r is calculated from the ranks replacing the scores, the result is given by the familiar formula

$$r_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}, \qquad (4)$$

where $D = X_R - Y_R$ is the difference between the ranks X_R and Y_R corresponding to X and Y, and n is the number of pairs of scores. If there are no ties, the value found by applying this formula to any data is exactly equal to the value found by calculating a Pearson r on ranks replacing the scores.

These relations can be summarized by saying that, if the initial scores are ranks, there is no difference between a Pearson r and a Spearman r_S, except for an algebraic detail of computation. They can also be expressed by saying that r_S is by definition a Pearson correlation coefficient obtained when scores have been converted to ranks before performing the usual calculations. Derivation of a simple equation containing only the sum of

[1] Most programming languages include a command such as Tanh(Z) or Tanh[Z] , which returns the desired value with far greater accuracy and convenience than interpolating in r to Z statistical tables and reading them backwards.

squared difference scores, D^2, and the number of pairs, n, is made possible by the fact that the variances of the two sets of ranks are equal and are given by $(n^2 - 1)/12$.

For present purposes, it is important to note the following properties of the sampling distributions of these statistics. If initial data are ranks, there is only one sampling distribution to be considered—that of r_S. It is conceivable that sampling is not from a population of numerical values, but rather from a population of non-numerical objects that can be compared and ranked. If initial data are scores on continuous variables, X and Y, then r computed from ranks corresponding to X and Y is the same as r_S. However, the sampling distribution of r computed from initial scores is not necessarily the same as that of r_S. That is, transformation of scores to ranks before calculating a statistic can alter the sampling distribution of that statistic.

If initial scores are numerical values rather than ranks, the Spearman r_S and the Pearson ρ are related by the formula

$$E[r_S] = \frac{6}{\pi(n+1)}\left[\sin^{-1}\rho + (n-2)\sin^{-1}\frac{\rho}{2}\right] \qquad (5)$$

(Daniels, 1950, 1951; Durbin and Stuart, 1951; Hoeffding, 1948; Kendall and Gibbons, 1990). This relation indicates the bias introduced by using the Spearman r_S obtained from ranks as an estimate of the population correlation between variables underlying the ranks. In other words, it indicates the bias introduced by transforming scores to ranks before estimating the population correlation. We find $\lim_{n\to\infty} E[r_S] = (6/\pi)\left[\sin^{-1}(\rho/2)\right]$, so that substantial bias exists for large sample sizes. Differentiating $\lim_{n\to\infty} E[r_S] - \rho$ and setting the result equal to 0 indicates that in the limit this bias is a maximum when the absolute value of ρ is .594. Figure 1 plots the theoretical bias of r_S as a function of ρ for n = 10, 20, 40, 80, and ∞ (corresponding to the curves from bottom to top).

Several authors have recommended using the r to Z transformation to test non-null hypotheses about the Spearman rank correlation (David and Mallows, 1961; Fieller, Hartley, and Pearson, 1957; Fieller and Pearson, 1961) in the same way as done for the Pearson correlation. These authors have found, however, that a more precise estimate of the standard deviation of the Z values obtained from ranks is $\sqrt{1.060/(n-3)}$ instead of $\sqrt{1/(n-3)}$ typically used in significance tests with the r to Z transformation. Apparently, there is not much evidence of the advantages or disadvantages of using these formulas in practice.

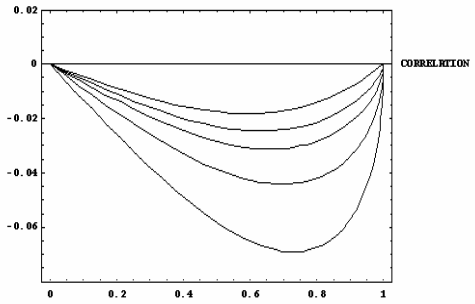

Figure 1. Theoretical bias of sample rank correlation as a function of population correlation for sample sizes. (The vertical axis measures bias and the values of rho are on the horizontal axis; the curves, from bottom to top trace n=10, 20, 40, 80, infinity)

COMPUTER SIMULATION METHOD[2]

In a simulation study, ordered pairs of scores of each of two normally distributed variables, *X* and *Y*, were transformed to have various population distributions and further transformed to have specified correlations. Normal deviates were generated by the method of Box and Muller (1958), where $X = (-2\log U_1)^{1/2}\cos(2\pi U_2)$ and $Y = (-2\log U_1)^{1/2}\sin(2\pi U_2)$, and where U_1 and U_2 are pseudorandom numbers on the interval (0,1). As a check, normal deviates were also generated by the rejection method of Marsaglia and Bray (1964).

---

[2] *Author Notes*: The computer program was written in PowerBASIC, version 3.2, PowerBASIC, Inc., Carmel, CA. A listing of the program can be obtained by writing to Donald W. Zimmerman, 1978 134A Street, South Surrey, B.C., Canada, V4A 6B6, E-mail: zimmerma@direct.ca. Calculation of theoretical values in Table 1 and Figures 1 and 3 was done with MATHEMATICA, version 4.0, Wolfram Research, Champaign, IL.

The exponential component of the non-normal distributions to be described below was obtained by $X = -\log(U) - 1$, where U is uniform on (0,1). The lognormal component was $Y = [\exp(X - 1) - .607]/.796$, where X is N(0,1), and the rectangular component was $X = (U - .5)/.289$, where U is uniform on (0,1), the constants insuring that the distributions have mean 0 and standard deviation 1. For further details concerning simulation of variates see Devroye (1986) and Morgan (1984).

The random number generator used in the present study was devised by Marsaglia, Zaman, and Tsang (1990), and was described in detail by Pashley (1993, pp. 395-415). In addition, random numbers were obtained from the PowerBASIC compiler used in the present study. The algorithms were tested by the above methods of generating random numbers and obtaining normal deviates, and differences among the methods turned out to be insignificant.

Correlations were induced by adding a multiple of a random variable, U, to both X and Y, the multiplicative constant, c, being chosen to produce the desired correlation. The algorithm was $X' = (X + cU)/(1 + c^2)$ and $Y' = (Y + cU)/(1 + c^2)$, where $c = \sqrt{r(1-r)}$. If X and Y are independent with mean 0 and variance 1, then $\rho(X', Y') = r$. The simulations consisted of at least 20,000 iterations and sometimes as many as 100,000 iterations for each condition investigated. In trial runs we attempted to locate the number of iterations that would yield stable calculated values.

Significance tests of the hypothesis of zero correlation employed the formula

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}, \qquad (6)$$

where the Student t statistic was evaluated at $n - 2$ degrees of freedom. Typically, this formula is used for significance testing of both Pearson and Spearman correlation (Glasser and Winter, 1961; Kendall, Kendall, and Smith, 1939). Equation (2) was used for Fisher r to Z transformations. Scores of X and Y were separately converted to ranks, and significance tests were performed on the ranks replacing the scores. Additional significance tests of the Spearman r_S used critical values obtained from permutations, given in tables in Siegel and Castellan (1988). The study employed the .01, .05, and .10 significance levels. In various parts of the study, sample sizes were n = 10, 20, and 40, where n denotes the number of ordered pairs of scores. All tests were two-tailed.

RESULTS

Table 1. Means and standard deviations of sample correlations (r) and Fisher approximate unbiased estimates of population correlation (estimated ρ) obtained from sample scores (Pearson r) and ranks of sample scores (Spearman r_S) for various population correlations (ρ) and sample sizes (n).

			ρ				
n			.10	.30	.50	.70	.90
10	scores	Mean r	.094	.284	.479	.679	.889
		Predicted mean r	.095	.286	.481	.682	.891
		Mean estimated ρ	.097	.293	.494	.694	.897
		SD r	.331	.310	.267	.195	.084
		SD estimated ρ	.342	.319	.272	.194	.080
	ranks	Mean r	.086	.261	.443	.631	.842
		Mean estimated ρ	.089	.270	.456	.647	.853
		SD r	.331	.314	.279	.217	.118
		SD estimated ρ	.343	.324	.284	.217	.112
20	scores	Mean r	.098	.294	.491	.690	.895
		Predicted mean r	.098	.293	.491	.691	.896
		Mean estimated ρ	.100	.300	.499	.699	.899
		SD r	.228	.211	.177	.126	.049
		SD estimated ρ	.233	.215	.179	.125	.048
	ranks	Mean r	.091	.275	.462	.656	.866
		Mean estimated ρ	.093	.281	.470	.664	.871
		SD r	.228	.215	.187	.142	.068
		SD estimated ρ	.233	.219	.188	.141	.066
40	scores	Mean r	.099	.297	.495	.695	.898
		Predicted mean r	.099	.297	.495	.696	.898
		Mean estimated ρ	.100	.300	.499	.699	.900
		SD r	.159	.147	.123	.085	.032
		SD estimated ρ	.160	.148	.124	.084	.032
	ranks	Mean r	.094	.281	.471	.669	.878
		Mean estimated ρ	.095	.284	.476	.673	.881
		SD r	.159	.149	.130	.096	.043
		SD estimated ρ	.161	.151	.131	.096	.042

Bias of the sample correlation coefficient and correction by unbiased estimators.

The simulation results presented in Table 1 reveal the bias of the sample correlation coefficient and confirm the accuracy of the Fisher approximate unbiased estimator. First, it is clear that the mean value of *r*, over 100,000 samples, is consistently below the value of ρ, although the difference is slight. The bias is greater in the mid-range of ρ, especially .50 and .70. and also is greater for small sample sizes.

The second row of each section of the table (predicted mean *r*) gives the expected value of *r* predicted from Fisher's derivation, that is, the value of $\rho - \rho(1-\rho^2)/2n$. It is apparent that in all cases the Fisher estimator (third row of each section), based on this predicted value, almost completely eliminates the bias, although the mean of the Fisher estimates still is very slightly below ρ. This small difference, which occurs consistently, can be attributed to the fact that the Fisher formula is an approximation. The standard deviations of *r* values and estimated ρ values in the table are consistent with the well-known truncation of the sampling distribution in the upper range.

Bias of the Spearman rank correlation.

The table also indicates that the Spearman rank correlation is characterized by a similar bias. In fact, the bias in all cases becomes somewhat larger when scores are converted to ranks. In the mid-range from .50 to .70, for $n = 10$, the mean of correlation coefficients based on ranks is as much as .05 to .07 below the population correlation. Even for the larger sample size, n = 40, the mean of r_S remains considerably less than ρ. The Fisher estimator applied to these rank correlations increases their value slightly, but does not nearly restore them to the population values.

Figure 2 provides a somewhat more detailed picture of the bias and its correction by the Fisher estimator and the Olkin-Pratt estimator and of the result of transforming scores to ranks, for sample sizes of 10 and 20. The bias, defined as the mean of sample *r* values minus the population ρ, is plotted as a function of ρ. Each data point is based on 100,000 iterations. These curves show clearly that the bias is greatest in the mid-range of about .50 to .80 and that the Fisher estimator, over many samples, almost, but not entirely, restores the mean *r* to ρ. They also show that bias is considerably greater for ranks than for scores. The pattern of results is the same for the two sample sizes, and the bias is greater for the smaller sample size.

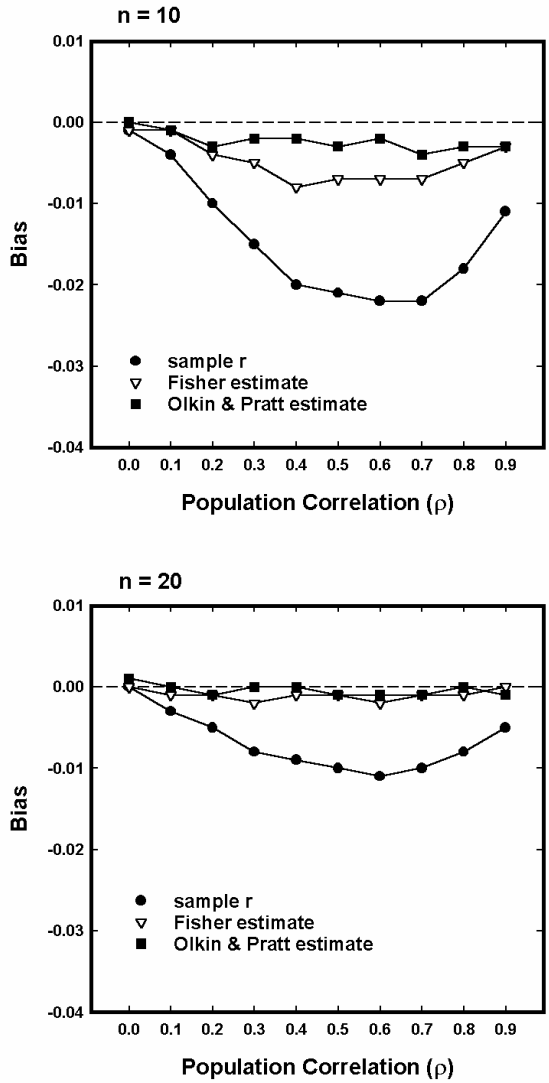

Figure 2. Bias of sample correlation and correction by approximately unbiased estimators as a function of population correlation, for scores and ranks.

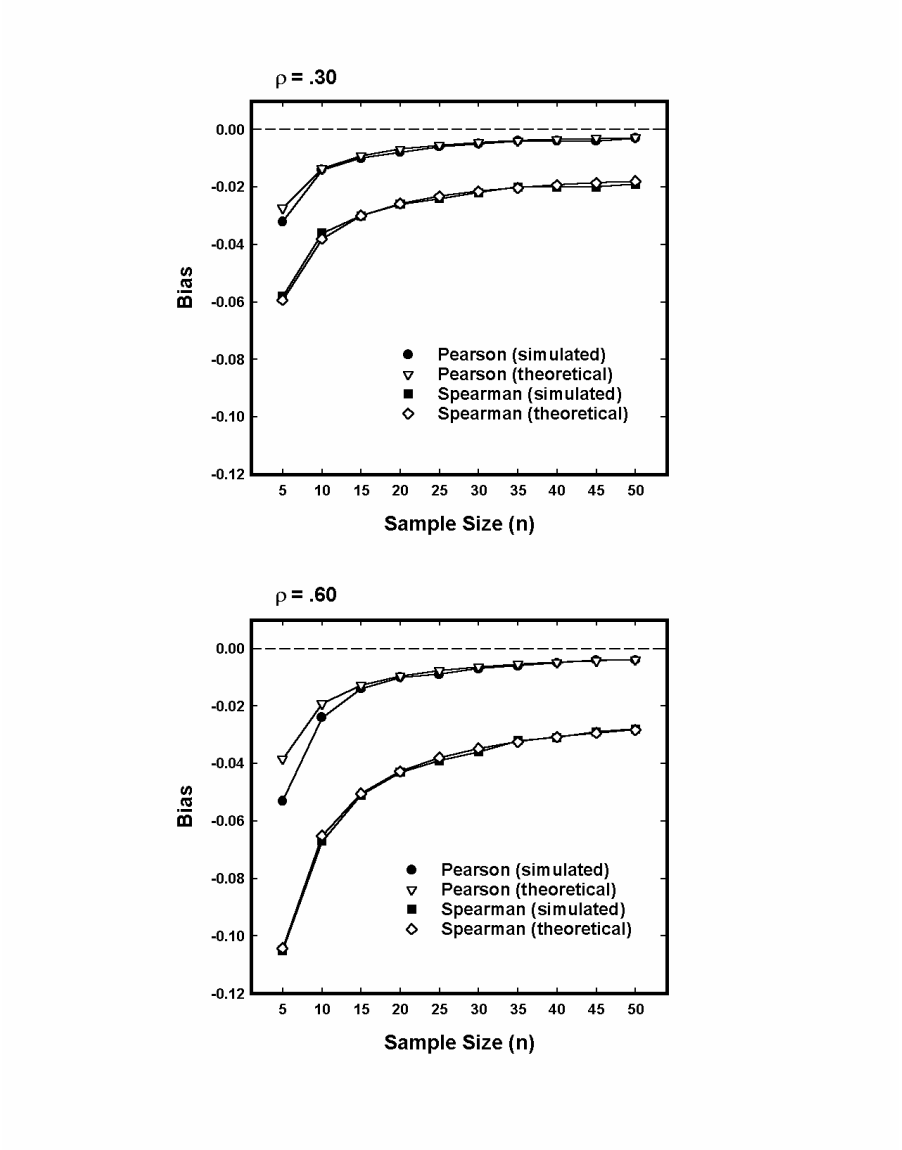

Figure 3. Bias of sample correlation and correction by approximately unbiased estimators as a function of sample size.

Figure 3 plots the bias for both scores and ranks as a function of sample size, varying from 5 to 50 in increments of 5, for ρ = .30 and ρ = .60. The upper curve in each section (labeled theoretical) is the bias predicted from Fisher's equation—the value of $-\rho(1-\rho^2)/2n$. The figure reveals that, even for larger sample sizes, the correlation based on ranks remains biased and apparently does not approach zero asymptotically, as does the correlation based on scores.

Significance tests of hypotheses about correlation.

Table 2 presents Type I error probabilities and power of various significance tests of correlation. The table gives results for three significance levels— .01, .05, and .10 (two-tailed). Tests of H: ρ = 0 performed on both scores and ranks were based on the Student *t*-statistic, using equation (5). These results are presented in the first two columns of each section of the table, labeled *t*-scores and *t*-ranks. As the actual value of ρ ranges from 0 to .80 in increments of .20, the columns indicate Type I error probabilities and power of the tests. In addition, tests of the same hypothesis were performed using critical values of the Spearman rank correlation obtained from permutations, and these results are presented in the third column, labeled r_S.

Tests of H: ρ = .40 and H: ρ = .80 were based on the Fisher *r* to *Z* transformation, For the first hypothesized value, the actual value of ρ ranged from .40 to .90 in increments of .10, and for the second hypothesized value, ρ ranged from .80 to .95 in increments of .05, so that the columns again indicate both Type I error probabilities and power. In the columns labeled scores and ranks, the estimated standard deviation of Z in the test statistic was $\sqrt{1/(n-3)}$, and in the column labeled ranks-1, the estimate was the more accurate value $\sqrt{1.060/(n-3)}$ suggested by Fieller and Pearson (1961).

For ρ = 0, for all three significance levels, and for all sample sizes, the test based on the Student *t*-statistic on scores was somewhat more powerful than the test on ranks, and the t-test on ranks was slightly more powerful than the r_S test based on permutations. For ρ = .40 and ρ = .80 and for all three significance levels, the test based on the *r* to *Z* transformation of correlation obtained from scores was more powerful than the test based on the *r* to *Z* transformation of correlation obtained from ranks. The power difference becomes somewhat larger as sample size increases. Also, the Type I error probability of the test based on ranks is inflated as ρ increases. Figure 4 compares in more detail the Type I error

probabilities of the r to Z transformation in the non-null case both the unmodified and the modified standard deviations of Z.

Table 2. Type I error probability and power of tests of the hypothesis $\rho = 0$, based on the Student *t* test on scores and ranks and on critical values of r_S obtained from permutations, and the hypotheses $\rho = .40$ and $\rho = .80$, based on the Fisher r to Z transformation ($\alpha = .01,.05$, and .10).

H_0: $\rho = 0$, $\alpha = .05$

	$n = 10$			$n = 20$			$n = 40$		
ρ	*t*-scores	*t*-ranks	r_S	*t*-scores	*t*-ranks	r_S	*t*-scores	*t*-ranks	r_S
0	.049	.054	.048	.052	.053	.045	.050	.050	.051
.20	.084	.083	.075	.132	.126	.118	.235	.217	.213
.40	.207	.185	.170	.427	.376	.371	.741	.688	.687
.60	.488	.417	.396	.830	.764	.760	.988	.976	.978
.80	.870	.773	.760	.995	.986	.985	1.000	1.000	1.000

H_0: $\rho = .40$, $\alpha = .05$ (using *r* to *Z*)

	$n = 10$			$n = 20$			$n = 40$		
ρ	scores	ranks	ranks-1	scores	ranks	ranks-1	scores	ranks	ranks-1
.40	.050	.055	.051	.050	.054	.048	.050	.056	.049
.50	.068	.066	.064	.086	.076	.069	.123	.101	.091
.60	.122	.105	.104	.214	.171	.158	.390	.309	.290
.70	.243	.186	.186	.481	.382	.363	.787	.677	.656
.80	.482	.356	.356	.826	.702	.685	.987	.957	.951
.90	.846	.668	.668	.994	.965	.961	1.000	1.00	1.000

H_0:$\rho = .80$, $\alpha = .05$ (using *r* to *Z*)

	$n = 10$			$n = 20$			$n = 40$		
ρ	scores	ranks	ranks-1	scores	ranks	ranks-1	scores	ranks	ranks-1
.80	.049	.076	.055	.051	.072	.066	.050	.081	.073
.85	.081	.091	.056	.114	.085	.079	.173	.119	.109
.90	.196	.156	.099	.373	.227	.216	.649	.449	.431
.95	.556	.357	.248	.885	.650	.637	.995	.949	.944

Table 2 (continued)

H_0: $\rho = 0$, $\alpha = .01$

	$n = 10$			$n = 20$			n = 40		
ρ	*t*-scores	*t*-ranks	r_S	*t*-scores	*t*-ranks	r_S	*t*-scores	*t*-ranks	r_S
0	.010	.013	.007	.011	.012	.008	.010	.010	.010
.20	.019	.023	.012	.040	.039	.032	.088	.078	.076
.40	.064	.061	.042	.200	.171	.159	.506	.443	.435
.60	.226	.188	.134	.623	.533	.513	.950	.916	.916
.80	.655	.518	.427	.979	.944	.934	1.000	1.000	1.000

H_0:$\rho = .40$, $\alpha = .01$ (using *r* to *Z*)

	$n = 10$			$n = 20$			$n = 40$		
ρ	scores	ranks	ranks-1	scores	ranks	ranks-1	scores	ranks	ranks-1
.40	.012	.015	.014	.011	.013	.010	.010	.012	.010
.50	.019	.021	.020	.024	.022	.019	.037	.030	.026
.60	.038	.037	.037	.078	.066	.057	.183	.140	.124
.70	.096	.078	.078	.251	.190	.171	.570	.452	.424
.80	.246	.175	.175	.619	.481	.450	.947	.872	.856
.90	.645	.441	.441	.970	.892	.878	1.000	.999	.998

H_0:$\rho = .80$, $\alpha = .01$ (using *r* to *Z*)

	$n = 10$			$n = 20$			$n = 40$		
ρ	scores	ranks	ranks-1	scores	ranks	ranks-1	scores	ranks	ranks-1
.80	.012	.019	.018	.011	.019	.016	.011	.021	.018
.85	.023	.027	.026	.034	.029	.023	.059	.039	.034
.90	.070	.056	.055	.173	.102	.088	.406	.246	.225
.95	.305	.163	.163	.715	.450	.415	.976	.860	.847

Table 2 (continued)

H_0: $\rho = 0$, $\alpha = .10$

	$n = 10$			$n = 20$			$n = 40$		
ρ	*t*-scores	*t*-ranks	r_S	*t*-scores	*t*-ranks	r_S	*t*-scores	*t*-ranks	r_S
0	.100	.104	.085	.103	.101	.095	.098	.099	.101
.20	.149	.150	.126	.223	.207	.199	.345	.319	.320
.40	.320	.289	.256	.560	.508	.504	.834	.790	.790
.60	.631	.554	.516	.899	.855	.850	.995	.989	.990
.80	.928	.865	.845	.997	.994	.994	1.000	1.000	1.000

H_0: $\rho = .40$, $\alpha = .10$ (using *r* to *Z*)

	$n = 10$			$n = 20$			$n = 40$		
ρ	scores	ranks	ranks-1	scores	ranks	ranks-1	scores	ranks	ranks-1
.40	.097	.105	.092	.100	.106	.096	.100	.108	.098
.50	.122	.120	.105	.152	.137	.124	.203	.168	.155
.60	.201	.178	.158	.319	.259	.241	.518	.423	.405
.70	.356	.290	.263	.609	.504	.482	.868	.776	.763
.80	.616	.490	.457	.894	.800	.784	.994	.978	.976
.90	.910	.784	.760	.997	.983	.981	1.000	1.000	1.000

H_0:$\rho = .80$, $\alpha = .10$ (using *r* to *Z*)

	$n = 10$			$n = 20$			$n = 40$		
ρ	scores	ranks	ranks-1	scores	ranks	ranks-1	scores	ranks	ranks-1
.80	.093	.136	.103	.099	.134	.121	.100	.143	.132
.85	.143	.147	.105	.188	.149	.134	.269	.190	.178
.90	.299	.224	.161	.500	.323	.302	.759	.562	.546
.95	.683	.451	.358	.936	.752	.731	.998	.972	.969

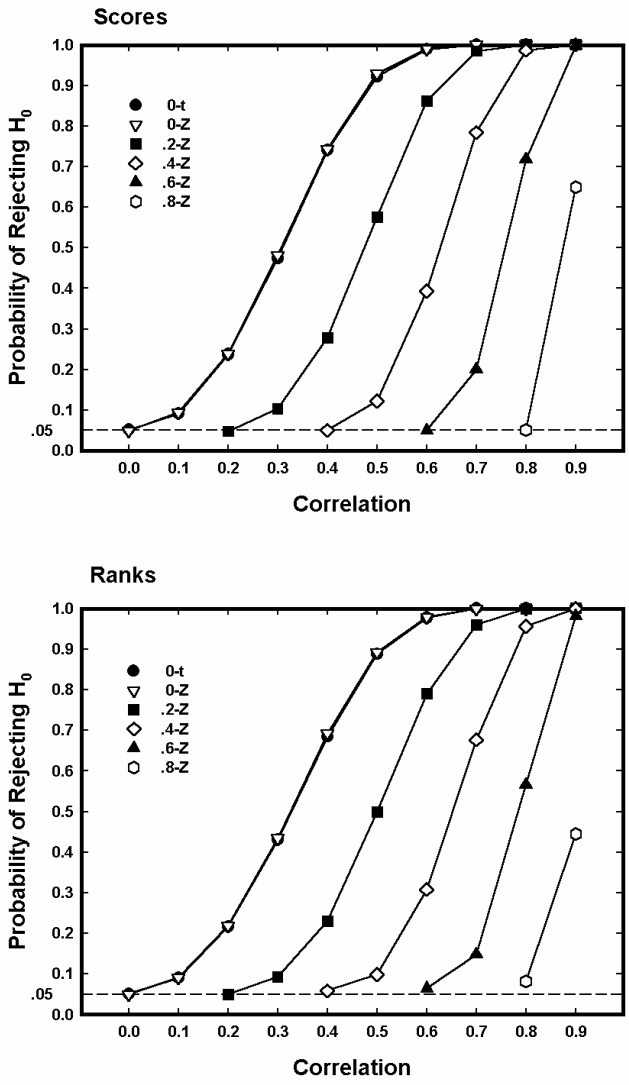

Figure 4. Type I error probability and power of tests of hypotheses about correlation based on the Student *t* statistic and on the *r* to *Z* transformation, for scores and ranks.

Figure 4 presents more detailed power functions of significance tests, based on scores and ranks, for a normal distribution, with $n = 40$ and $\alpha = .05$. These functions represent tests of hypotheses of successive correlations ranging from 0 to .80 in increments of .20, as the population correlation assumes values ranging from the hypothesized value to .90 in increments of .10. Tests of the hypothesis of zero correlation were made by both the *t*-test method and the *r* to *Z* method; all other hypotheses employed the *r* to *Z* method. For values of ρ greater than zero, the curves of the tests on scores dominate the curves of the tests on ranks. However, the results of the tests of H: $\rho = 0$ based on the *t*-statistic are almost identical to those of the test of the same hypothesis based on the *r* to *Z* transformation, for both scores and ranks.

Figure 5 shows the probability of Type I errors as a function of ρ for scores and ranks, using the *r* to *Z* transformation with the estimate of the standard deviation of *Z* based on $\sqrt{1/(n-3)}$ and for ranks with the estimate based on $\sqrt{1.060/(n-3)}$.

Estimation and significance testing under violation of bivariate normality.

Table 3 presents means and standard deviations of sample correlation coefficients obtained from distributions that violate bivariate normality. Table 4 presents Type I error probabilities of the significance tests described previously, for significance levels of .01, .05, and .10 and for sample sizes of 10, 20, and 40. Tests were performed on scores and on ranks. Tests of the hypothesis $\rho = 0$ employed the Student *t* test, as well as the test of r_S based on permutations, and tests of the hypothesis $\rho = .50$ employed the *r* to *Z* transformation.

Four distribution shapes were examined. These comprised skewed distributions of the kind often encountered in practice and mixed distributions that are models of outliers in research data. First, a mixture of a normal and exponential distribution, sometimes called an ex-Gaussian distribution, was included. This highly skewed distribution, N(0,1) + E(0,15), is the sum of a normal component with mean 0 and standard deviation 1 and an exponential component with mean 0 and standard deviation 15. Second, a similar mixture of a normal distribution and a lognormal distribution, both with mean 0 and standard deviation 1, was included. Third, a contaminated-normal, or mixed-normal distribution, frequently used as a model of outliers, consisted of samples taken from N(0,1) with probability .98 and from N(0,10) with probability .02. Finally, a

mixture of N(0,1) and a rectangular, or uniform, distribution with mean 0 and standard deviation 1 was examined.

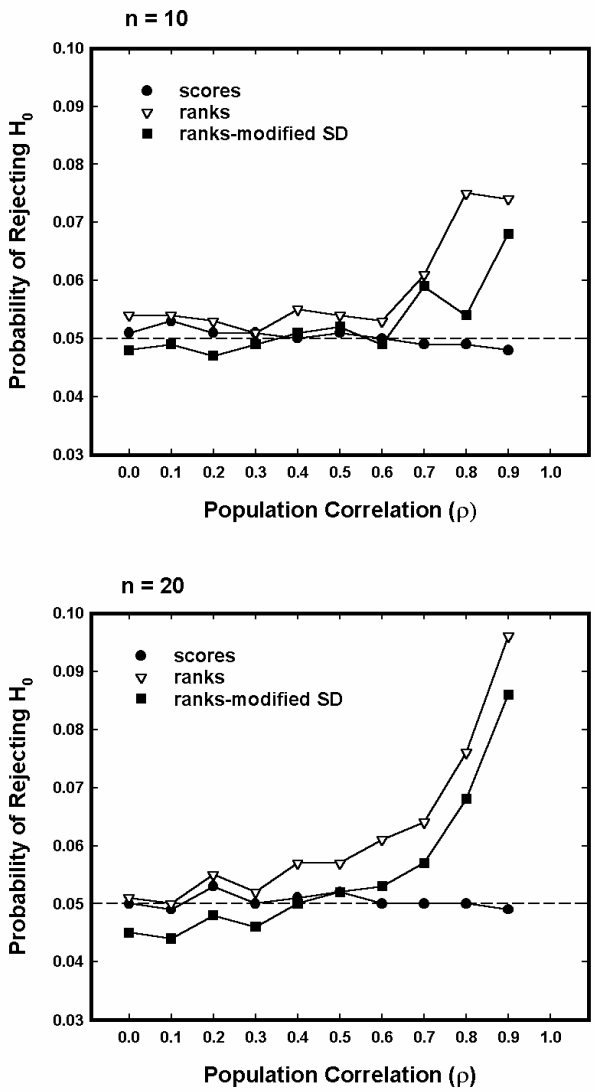

Figure 5. Probability of Type I errors of tests of non-null hypotheses about ρ for scores and ranks using r to Z transformation.

Table 3. Means and standard deviations of sample correlation coefficients based on scores and ranks under violation of the assumption of bivariate normality ($\rho = .50$).

		$n = 10$		$n = 20$		$n = 40$	
Distribution		scores	ranks	scores	ranks	scores	ranks
Mixture of normal and exponential (ex-Gaussian)	M	.497	.476	.500	.496	.501	.507
	SD	.271	.274	.185	.184	.128	.126
Mixture of normal and lognormal	M	.540	.527	.534	.547	.527	.559
	SD	.267	.259	.192	.174	.143	.120
Contaminated normal— mixture of N(0,1) and N(0,10).	M	.554	.532	.549	.554	.537	.565
	SD	.260	.255	.193	.170	.151	.116
Mixture of normal and rectangular	M	.473	.432	.488	.451	.494	.461
	SD	.265	.279	.175	.187	.121	.129

The data in Table 2 indicates that the outcome of significance tests of zero correlation for the various non-normal densities is similar to results found in studies of Type I error probabilities in *t*-tests and *F*-tests of difference in location. The significance levels of the *t* tests on scores are slightly disrupted, except for the case of the mixture of the normal and rectangular distribution, while those of the tests on ranks are not affected to the same extent. Previous studies have shown that the significance levels of tests of location on rectangular distributions, are not distorted.

On the other hand, the significance levels of the tests of the hypothesis $\rho = .50$, based on the *r* to *Z* transformation, are severely distorted for these heavy-tailed distributions. In many cases, the tests on ranks are distorted to a greater extent than the tests on scores. In the case of the lognormal and contaminated normal distributions, inflation of the Type I error probability is extreme, and it becomes more severe as sample size increases. All evidence appears to indicate that the *r* to *Z* transformation is not robust to non-normality.

Further insight into how non-normality influences correlation is obtained from Figures 6 and 7, which give relative frequency distributions of values of r and Z for population correlations of .25 and .85. The samples represented in Figure 6 were obtained from a normal distribution, and those in Figure 7 were obtained from a contaminated normal distribution, as

defined previously. Table 5 gives the means and standard deviations of the sample distributions. Under violation of bivariate normality, it appears that the form of the sample distributions of r and Z are not appreciably altered, but that the means of the distributions are changed enough to substantially modify the probability of Type I and Type II errors.

Table 4. Type I error probabilities of tests of the hypotheses $\rho = 0$ and $\rho = .50$ under violation of bivariate normality.

		$n = 10$			$n = 20$			$n = 40$		
Distribution	α	t-scores	t-ranks	r_S	t-scores	t-ranks	r_S	t-scores	t-ranks	r_S
Mixture of normal and exponential (ex-Gaussian)	.01	.014	.012	.007	.014	.011	.010	.012	.009	.009
	.05	.050	.053	.048	.049	.050	.049	.049	.051	.050
	.10	.095	.105	.088	.094	.101	.099	.096	.099	.099
Mixture of normal and lognormal	.01	.024	.013	.008	.024	.011	.010	.021	.010	.009
	.05	.060	.056	.050	.055	.050	.049	.050	.051	.050
	.10	.098	.108	.091	.088	.101	.099	.082	.101	.101
Contaminated normal—mixture of N(0,1) and N(0,10)	.01	.012	.013	.007	.013	.011	.010	.014	.011	.010
	.05	.052	.054	.048	.051	.051	.049	.049	.051	.050
	.10	.101	.105	.088	.098	.100	.099	.094	.101	.101
Mixture of normal and rectangular	.01	.011	.013	.007	.011	.011	.010	.011	.011	.010
	.05	.052	.054	.049	.050	.051	.049	.050	.050	.049
	.10	.101	.105	.088	.102	.100	.099	.099	.098	.098

		$n = 10$		$n = 20$		$n = 40$	
Distribution	α	Z-scores	Z-ranks	Z-scores	Z-ranks	Z-scores	Z-ranks
Mixture of normal and exponential (ex-Gaussian)	.01	.027	.029	.028	.035	.026	.041
	.05	.085	.081	.089	.098	.089	.114
	.10	.144	.140	.151	.159	.153	.180
Mixture of normal and lognormal	.01	.089	.085	.111	.144	.131	.245
	.05	.192	.180	.224	.268	.251	.409
	.10	.273	.259	.307	.357	.334	.511
Contaminated normal—mixture of N(0,1) and N(0,10)	.01	.070	.059	.141	.124	.218	.263
	.05	.188	.147	.297	.265	.370	.457
	.10	.285	.230	.400	.364	.459	.565
Mixture of normal and rectangular	.01	.011	.015	.009	.012	.009	.014
	.05	.045	.052	.043	.054	.044	.062
	.10	.089	.100	.088	.104	.090	.118

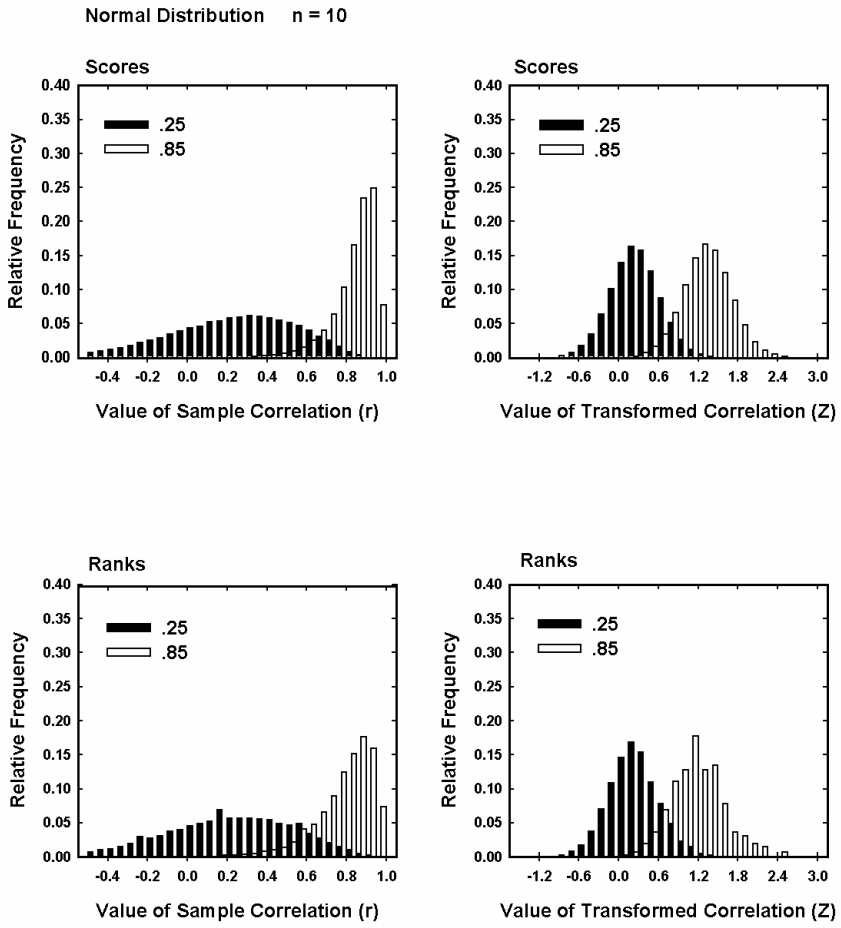

Figure 6. Relative frequency distributions of values of *r* and *Z* for scores and ranks and for population correlations of .25 and .85 (normal distribution).

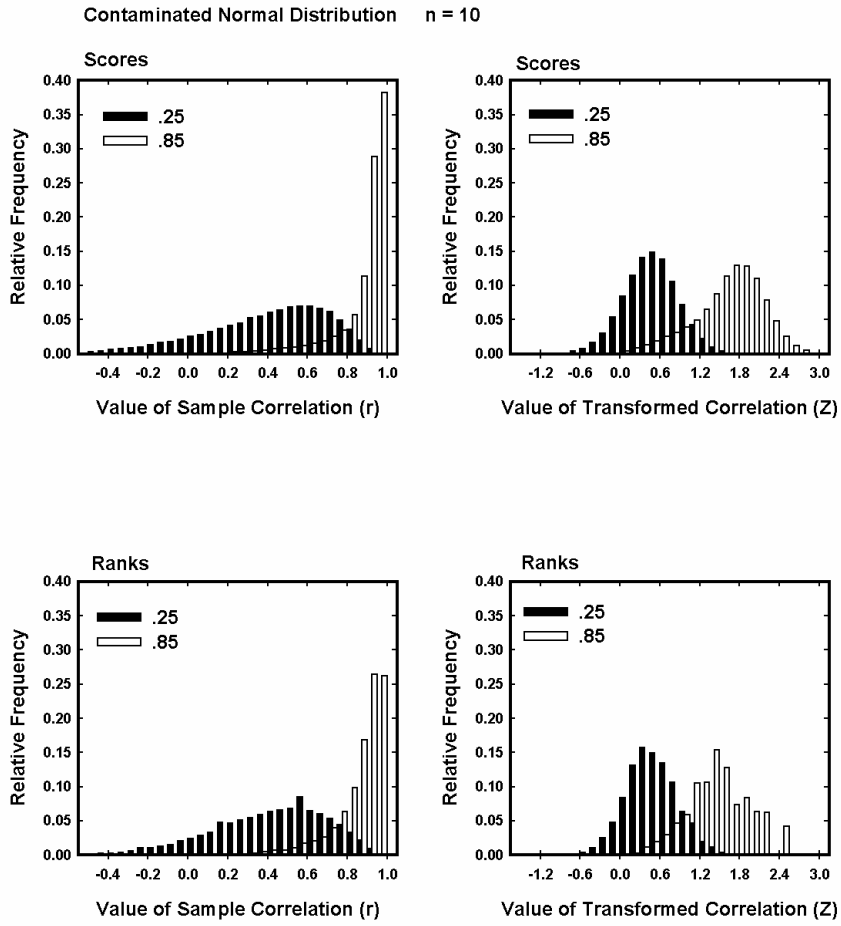

Figure 7. Relative frequency distributions of values of *r* and *Z* for scores and ranks and for population correlations of .25 and .85 (contaminated normal distribution).

Table 5. Means and standard deviations of sample values (*r*) and transformed values (*Z*) of correlation coefficients obtained from normal and contaminated-normal populations ($\rho = .25$ and $\rho = .85$).

Scores

		$\rho = .25$		$\rho = .85$	
Distribution		*r*	*Z*	*r*	*Z*
Normal	Mean	.238	.272	.835	1.311
	SD	.317	.376	.116	.367
Contaminated Normal	Mean	.406	.489	.878	1.628
	SD	.311	.410	.159	.530

Ranks

		$\rho = .25$		$\rho = .85$	
Distribution		*r*	*Z*	*r*	*Z*
Normal	Mean	.219	.252	.786	1.182
	SD	.320	.381	.148	.418
Contaminated Normal	Mean	.416	.500	.862	1.503
	SD	.287	.389	.131	.530

Some Practical Implications for Psychological Measurement

These findings have some practical implications for research in psychology, education, and other fields. First, if one is troubled by the slight bias in the correlation coefficient for normal populations, it is clear that it can be largely eliminated by the Fisher approximate unbiased estimator or by the Olkin and Pratt estimator. It is a simple matter to employ one of these formulas routinely in calculating correlation coefficients.

For many purposes in educational and psychological research, the bias revealed in the present study may not be large enough to cause concern. If one's research involves simply establishing the existence of correlation, the bias probably is not excessive. However, if one is concerned with the accuracy of specific degrees of non-zero correlation, especially higher

degrees of correlation, then correcting for the bias may be desirable. In the past, investigators in some areas have paid close attention to precise numerical values of correlation coefficients, and in these circumstances bias of $-.03$ or $-.04$ could be misleading. This is especially true if sample sizes are small and at the same time the correlation is high.

If sample sizes are large, the bias is relatively small. For example, in studies of test reliability, where reliability coefficients of .80 or higher are common, it is likely that bias is negligible, because the estimates are usually based on thousands of examinees. The same is true for validity coefficients in large samples. On the other hand, if the validity or reliability of a dependent variable in an experimental study with a small number of subjects is found to be .60, it could be .63 or .64 in a larger population. Even in cases where the bias is smaller, however, nothing is lost by using an unbiased estimator. It certainly is possible to incorporate these estimators without undue difficulty into computer programs, and there is no reason why statistical software could not routinely obtain the more accurate estimates.

Another implication of the present findings is that, in practice, the r to Z transformation can be expected to be sensitive to violation of bivariate normality. This fact is relevant to hypotheses testing, finding confidence intervals, and averaging correlation coefficients. In these applications, neither large samples nor conversion to ranks affords protection. For example, significance tests of hypotheses about validity and reliability coefficients or differences between them require an assumption of bivariate normality despite large sample sizes.

Researchers certainly should be aware of this assumption before using the r to Z transformation in data analysis. If it is not tenable, estimates of non-zero values of correlation coefficients can be extremely biased, and significance tests can be invalid. These consequences appear to be more severe than ones typically associated with non-normality in t and F tests of differences in location.

Spearman rank correlation frequently is used when research data initially is in the form of ranks, and numerical measures underlying the ranks are unavailable or meaningless. For such data, it is immaterial whether one uses the Pearson formula to calculate the correlation between the ranks, or the Spearman computational formula instead. On the other hand, if the initial data is numerical, but the assumption of bivariate normality is not satisfied, transformation to ranks is desirable in order to avoid distortion of significance tests because of the distributional properties of the scores. In this case, the correlation between the ranks is not necessarily the same as the correlation between the scores. It is still possible

to test the hypothesis of zero correlation, although the probabilities of Type I and Type II errors of the test on ranks are not necessarily the same as those of the test on scores.

In this situation, unfortunately, there is no effective method for testing hypotheses about non-zero values of correlation, and further research on this problem is needed. Another topic for further research is suggested by Figure 3. Note that the bias of the rank correlation asymptotically approaches a negative value, about −.02 when ρ is .30 and about .03 when is ρ is .60. These values of the bias could be calculated and tabulated more systematically for a range of correlations and used as correction factors in large-sample research, when *n* is large enough for the bias to be close to the asymptotic value.

REFERENCES

Box, G.E.P., & Muller, M. (1958). A note on the generation of normal deviates. *Annals of Mathematical Statistics, 29*, 610-611.

Charter, R.A., & Larsen, B.S. (1983). Fisher's *z* to *r*. *Educational and Psychological Measurement, 43,* 41-42.

Daniels, H.E. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society, B, 12,* 171-181.

Daniels, H.E. (1951). Note on Durbin and Stuart's formula for $E(r_S)$. *Journal of the Royal Statistical Society, B, 13,* 310.

David, F.N., & Mallows, C.L. (1961). The variance of Spearman's rho in normal samples. *Biometrika, 48,* 19-28.

Devroye, L. (1986). *Non-uniform random variate generation.* New York: Springer-Verlag.

Durbin, J., & Stuart, A. (1951). Inversions of rank correlation coefficients. *Journal of the Royal Statistical Society, B, 13,* 303-309.

Fieller, E.C., Hartley, H.O., & Pearson, E.S. (1957). Tests for rank correlation coefficients: I. *Biometrika, 44,* 470-481.

Fieller, E.C., & Pearson, E.S. (1961). Tests for rank correlation coefficients: II. *Biometrika, 48,* 29-40.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10,* 507-521.

Fisher, R.A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3-32.

Fowler, R.L. (1987). Power and robustness in product-moment correlation. *Applied Psychological Measurement, 11,* 419-428.

Glasser, G.J., & Winter, R.F. (1961). Critical values of the coefficient of rank correlation for testing the hypothesis of independence. *Biometrika, 48,* 444-448.

Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics, 19*, 546-557.

Kendall, M.G., Kendall, F.H., & Smith, B.B. (1939). The distribution of Spearman's coefficient of rank correlation in a universe in which all rankings occur an equal number of times. *Biometrika, 30,* 251-273.

Kendall, M.G., & Gibbons, J.D. (1990). *Rank correlation methods* (5th ed.). New York: Oxford University Press.

Kenney, J.F., & Keeping, E.S. (1951). *Mathematics of statistics, part two* (2nd ed.). New York: Van Nostrand.

Marsaglia, G., & Bray, T.A. (1964). A convenient method for generating normal variables. *SIAM Review, 6,* 260-264.

Marsaglia, G., Zaman, A., & Tsang, W.W. (1990). Toward a universal random number generator. *Statistics & Probability Letters, 8,* 35-39.

Morgan, B.J.T. (1984). *Elements of simulation.* London: Chapman & Hall.

Muchinsky, P.M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56,* 1, 63-65.

Olkin, I., & Pratt, J.W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29,* 201-211.

Pashley, P.J. (1993). On generating random sequences. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 395-415). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sawkins, D.T. (1944). Simple regression and correlation. *Journal and Proceedings of the Royal Society of New South Wales, 77,* 85-95.

Siegel, S., & Castellan, N.J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw-Hill.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Stuart, A. (1954). The correlation between variate-values and ranks in samples from a continuous distribution. *British Journal of Psychology, 7,* 37-44.

Zimmerman, D.W., and Williams, R.H. (1997). Properties of the Spearman correction for attenuation for normal and realistic non-normal distributions. *Applied Psychological Measurement, 21*, 3, 253-270.

(Manuscript received: 27/2/02; accepted: 14/5/02)