Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 24

11-1-2003

Bootstrapping Confidence Intervals For Robust Measures Of Association

Jason E. King
Baylor College of Medicine, jasonk@bcm.tmc.edu

DOI: 10.22237/jmasm/1067646240

Bootstrapping Confidence Intervals For Robust Measures Of Association

Cover Page Footnote
The author acknowledges Professor Bruce Thompson and the doctoral committee for their contributions.

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 512-519

1538 – 9472/03/$30.00

Bootstrapping Confidence Intervals For Robust Measures Of Association

Jason E. King
Baylor College of Medicine

A Monte Carlo simulation study compared four bootstrapping procedures in generating confidence intervals for the robust Winsorized and percentage bend correlations. Results revealed the superior resiliency of the robust correlations over *r*, with neither outperforming the other. Unexpectedly, the bootstrapping procedures achieved roughly equivalent outcomes for each correlation.

Key words: Robust methods, bootstrapping, percentage bend correlation, Winsorized correlation

Introduction

A number of "robust" (Box, 1953) analogs to traditional estimators, population parameters, and hypothesis-testing methods have seen development during the past 40 years. Robust procedures typically retain the statistical interpretations associated with classical procedures, but are more resistant to distributional non-normalities and outliers. The Pearson product-moment correlation is without question the most commonly used measure of linear association, yet is not robust to departures from normality, especially when the bivariate surface is non-normal and dependence exists (King, 2003).

Two new robust alternatives to *r* appear promising. The Winsorized correlation (Devlin, Gnanadesikan, & Kettenring, 1975; Gnanadesikan & Kettenring, 1972; Wilcox, 1993) and the percentage bend correlation (Wilcox, 1994, 1997) yield interpretations analogous to *r* and asymptotically equal zero under bivariate independence, yet possess properties that curb the influence of distributional non-normalities.

This article was based on the doctoral dissertation by Jason E. King. The author acknowledges Professor Bruce Thompson and the doctoral committee for their contributions. Email address: jasonk@bcm.tmc.edu.

The Winsorized correlation (r_w) is computed in an identical fashion to *r* except that a specified proportion of extreme scores in each tail are first Winsorized, that is, deleted and set equal to the most extreme score remaining in the tail of the distribution. The percentage bend correlation (r_{pb}) is based on the percentage bend measures of location and midvariance and is less intuitive. See Wilcox (1994, 1997) for the relevant equations.

Yet few researchers have explored these newer correlations, notably with respect to estimating confidence intervals and defining their sampling distributions. For statistics with no known sampling distribution, Efron's (1979, 1982) *bootstrap* has proven to be effective in a variety of contexts. The conjecture is that the sampling distribution of a statistic can be approximated by the distribution of a large number of resampled estimates of the statistic obtained from a single sample of observations.

The distribution of resampled estimates forms an empirically-derived sampling distribution from which confidence intervals or other indices may be estimated, either for inferential or descriptive purposes (Thompson, 1993). The usefulness of bootstrapping is evident because an increasing number of disciplines are now encouraging or requiring the reporting of confidence intervals (Thompson, 2002; Vacha-Haase, Nilsson, Reetz, Lance, & Thompson, 2000; Wilkinson & APA Task Force on Statistical Inference, 1999).

An "almost bewildering array" (Hall, 1988, p. 927) of bootstrapping procedures is now available. These vary in the accuracy with which the bootstrap-generated interval spans the true interval. Accuracy is also contingent on the

type of statistic under examination. At the current level of knowledge, it is unknown which bootstrapping procedure produces the most accurate confidence intervals for r_{pb} and r_w. Although Wilcox (1993, 1994, 1997) compared Type I error rates for these robust correlations, only two studies (Wilcox, 1997; Wilcox & Muska, 2001) have examined the accuracy of bootstrapped confidence intervals for r_{pb}, and none for r_w. Clearly, more research is needed.

The goal of this simulation study was to compare various means of bootstrapping confidence intervals for r_w and r_{pb} across a variety of conditions. The study compared four bootstrapping procedures, each of which has proven useful in some contexts: the ordinary *percentile bootstrap* (Efron, 1979), an *adjusted bootstrap* (Strube, 1988), the *bias-corrected bootstrap* (BC; Efron, 1981, 1982, 1985), and the *bias-corrected and accelerated bootstrap* (BC_a; Efron, 1987). The Pearson *r* and Fisher's inverse hyperbolic tangent transformation of *r*, r_z, were included for comparative purposes, although the latter frequently fails to produce even asymptotically correct results (Duncan & Layard, 1973).

Methodology

The simulation procedure began by randomly generating 1,000,000 observations from a population with known characteristics, serving as a derived population. This step was necessary because the Winsorized and percentage bend correlation parameters (ρ_w and ρ_{pb}) will not necessarily exactly equal ρ under dependence conditions. The second step involved drawing $m = 100$ samples, each of size *n*, from the derived population and calculating sample estimates for each of the four correlational measures. Lastly, $B = 500$ bootstrap samples were drawn by sampling with replacement from each of the *m* samples and 95% confidence intervals calculated via each of the four bootstrapping procedures. Gamma (γ) and beta (β) are two constants that must be fixed in computing the Winsorized and percentage bend correlations, respectively. These were each set to .2 for all simulations.

Real data often demonstrate excessive distributional non-normality (Bradley, 1977, 1978; Micceri, 1989; Rasmussen, 1986; Stigler, 1973; Wilcox, 1990) and such can moderate the accuracy of a bootstrapping procedure for a given statistic (Hall, 1988; Wilcox, 1997). Thus, the present study compared bootstrapped correlations across a wide range of conditions including nine distributional shape variations, one contaminated distribution, six mixed distributions, three independence and dependence conditions (i.e., population correlations of .0, .4, .8), and four sample sizes (i.e., *n*s of 20, 50, 100, 250).

Four indices served as points of comparison for the bootstrapped correlations: Type I error rate, bias, efficiency, and interval width. The latter was constructed by modifying a ratio proposed by Efron (1988) such that the width of each bootstrap-estimated interval was divided by the width of a "true" (i.e., Monte Carlo-estimated) confidence interval. This required drawing an additional 10,000 samples, each of size *n*, from each simulated population to create the "true" sampling distributions.

Simulation studies typically compare Type I error rates and other indices in an informal manner; however, a more formal analysis is useful for processing the large number of indices obtained in the present study. Analysis of Variance (ANOVA) is well suited for quantifying sources of variation. This procedure allowed for partitioning the systematic variance components affecting the indices (viz., correlational measure, bootstrapping procedure, distributional shape and type, sample size, and strength of bivariate relationship).

Results

Tables 1-5 and Figures 1-2 display representative results averaged across distributional shape. Disaggregated data and fuller explanations are available in King (2000). Efficiency varied little across the correlational measures and is not presented.

Comparisons Among Bootstrapping Procedures

As regards Type I error rate (see Tables 1, 2, and Figure 1) and bias (see Tables 3, 4, and Figure 2), no bootstrapping procedure emerged as unmistakably superior across a majority of conditions for either robust correlation (e.g., a Bootstrap by Correlation effect is absent in Tables 2 and 4).

Table 1. Type I Error Rates Averaged Across All Distributional Conditions

	n = 20				n = 50				n = 100				n = 250			
	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}
ρ = 0																
Percentile	.06	.06	.03	.04	.07	.07	.06	.07	.05	.05	.05	.05	.07	.07	.04	.04
Adjusted	.06	.06	.03	.04	.07	.07	.06	.07	.05	.05	.05	.05	.07	.07	.04	.04
BC	.05	.05	.03	.05	.07	.07	.06	.06	.05	.05	.06	.05	.06	.06	.03	.04
BC_a	.05	.04	.03	.05	.07	.07	.05	.06	.06	.06	.06	.05	.07	.07	.04	.04
ρ = .4																
Percentile	*.11*	*.11*	.03	.07	.08	.08	.04	.06	.07	.07	.04	.04	.08	.08	.05	.05
Adjusted	*.11*	*.11*	.04	.08	.08	.08	.04	.06	.08	.08	.04	.04	.08	.08	.05	.05
BC	.08	.08	.03	.06	.08	.08	.03	.06	.08	.08	.04	.04	.09	.09	.05	.05
BC_a	.09	.08	.04	.07	.09	.09	.04	.06	*.10*	*.10*	.04	.03	*.11*	*.11*	.04	.05
ρ = .8																
Percentile	.09	.09	.06	.07	.06	.06	.06	.06	.06	.06	.06	.04	.07	.07	.05	.05
Adjusted	*.15*	*.15*	.09	*.12*	.06	.06	.06	.06	.08	.08	.07	.07	.08	.08	.04	.05
BC	*.10*	*.10*	.05	.05	.06	.06	.06	.07	.07	.07	.06	.04	.08	.08	.06	.05
BC_a	*.12*	*.12*	.06	.07	.07	.07	.05	.06	*.10*	*.10*	.06	.04	.09	.09	.06	.06

Note. Italicized values are greater than two standard errors beyond the nominal .05 level.

Table 2. Analysis of Variance for Type I Error Rate by Correlation and Bootstrapping Procedure

Source	*df*	*F*	*p*	η^2
Model	15	11.028	<.001	.088
CORR	3	50.511	<.001	.081
BOOT	3	2.735	.042	.004
CORR * BOOT	9	.631	.772	.003
Error	1712	(.002)		
Total	1727			

Note. Mean square error enclosed in parentheses.

Figure 1. Mean Type I error rate by correlation and bootstrapping procedure. Reference line indicates the nominal alpha rate of .05.

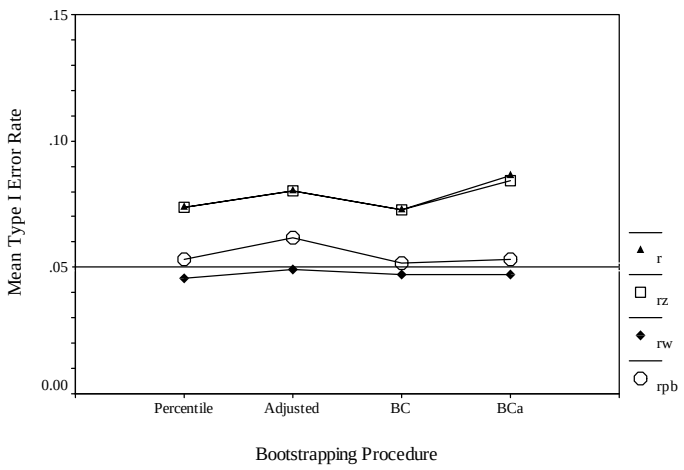

Table 3. Interval Bias Averaged Across All Distributional Conditions

	n = 20				*n* = 50				*n* = 100				*n* = 250			
	r	r_z	r_w	r_{pb}	*r*	r_z	r_w	r_{pb}	*r*	r_z	r_w	r_{pb}	*r*	r_z	r_w	r_{pb}
ρ = 0																
Percentile	.33	.33	.30	.31	.24	.24	.22	.22	.17	.17	.16	.16	.11	.11	.10	.10
Adjusted	.36	.36	.35	.34	.24	.24	.23	.23	.17	.17	.16	.16	.11	.11	.10	.10
BC	.32	.32	.31	.31	.23	.23	.22	.22	.16	.16	.16	.16	.11	.11	.10	.10
BC_a	.34	.33	.31	.31	.24	.24	.22	.22	.17	.17	.16	.16	.11	.11	.10	.10
ρ = .4																
Percentile	.36	.36	.33	.33	.24	.24	.20	.20	.21	.21	.15	.15	.15	.15	.10	.09
Adjusted	.38	.38	.37	.37	.24	.24	.21	.21	.21	.21	.15	.15	.15	.15	.10	.09
BC	.36	.36	.33	.32	.24	.24	.21	.20	.21	.21	.15	.14	.16	.16	.10	.10
BC_a	.38	.37	.33	.33	.26	.25	.21	.20	.23	.23	.15	.15	.17	.17	.10	.10
ρ = .8																
Percentile	.26	.26	.25	.23	.17	.17	.14	.13	.13	.13	.09	.08	.10	.10	.06	.05
Adjusted	.31	.31	.31	.30	.17	.17	.15	.15	.13	.13	.09	.09	.10	.10	.06	.06
BC	.26	.26	.28	.24	.17	.17	.14	.14	.14	.14	.09	.09	.11	.11	.06	.06
BC_a	.28	.28	.28	.25	.18	.18	.15	.15	.15	.15	.09	.09	.12	.12	.06	.06

Table 4. Analysis of Variance for Bias by Correlation and Bootstrapping Procedure

Source	*df*	*F*	*p*	η^2
Model	15	3.497	<.001	.030
CORR	3	15.558	<.001	.026
BOOT	3	1.551	.199	.003
CORR * BOOT	9	.125	.999	.001
Error	1712	(.010)		
Total	1727			

Note. Mean square error enclosed in parentheses.

Figure 2. Mean bias by correlation and bootstrapping procedure.

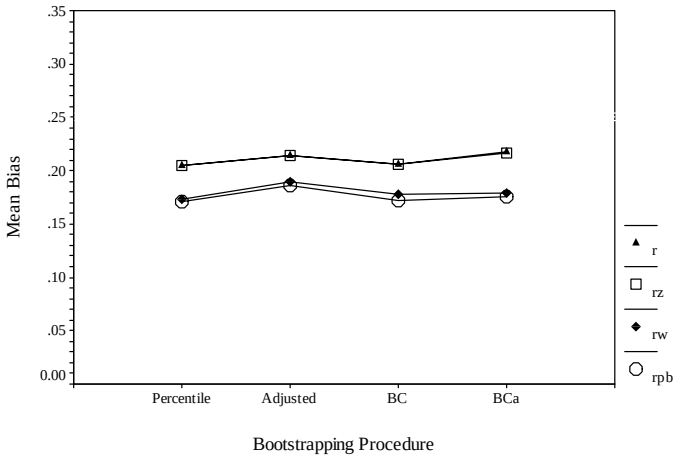

Table 5. Confidence Interval Ratios Averaged Across All Distributional Conditions

	n = 20				*n* = 50				*n* = 100				*n* = 250			
	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}	r	r_z	r_w	r_{pb}
ρ = 0																
Percentile	.91	.91	1.11	1.02	.91	.91	1.04	1.00	.92	.92	1.03	1.00	.93	.93	1.01	.99
Adjusted	.98	.98	1.20	1.10	.94	.94	1.07	1.03	.94	.94	1.04	1.02	.94	.94	1.01	1.00
BC	.92	.92	1.11	1.02	.91	.91	1.04	1.00	.92	.92	1.03	1.00	.93	.93	1.01	.99
BC_a	.92	.92	1.11	1.02	.92	.92	1.04	1.00	.93	.93	1.02	1.00	.94	.94	1.01	.99
ρ = .4																
Percentile	.84	.84	1.09	1.00	.84	.84	1.04	.99	.92	.92	1.03	1.00	.86	.86	1.01	1.00
Adjusted	.91	.91	1.17	1.08	.86	.86	1.07	1.02	.93	.93	1.04	1.02	.87	.87	1.02	1.00
BC	.86	.86	1.11	1.02	.84	.84	1.05	1.00	.91	.91	1.02	1.00	.86	.86	1.01	1.00
BC_a	.85	.86	1.11	1.02	.84	.84	1.05	1.01	.92	.92	1.03	1.00	.86	.86	1.01	1.00
ρ = .8																
Percentile	.81	.81	1.08	.98	.85	.85	1.05	1.02	.81	.81	1.00	.98	.84	.84	1.01	1.00
Adjusted	.87	.87	1.16	1.06	.88	.88	1.08	1.05	.83	.83	1.01	.99	.84	.84	1.01	1.00
BC	.85	.85	1.17	1.04	.87	.87	1.08	1.04	.82	.82	1.01	.99	.84	.84	1.01	1.00
BC_a	.83	.86	1.16	1.05	.85	.87	1.10	1.06	.81	.82	1.02	1.01	.84	.85	1.01	1.01

Note: Ratios greater than 1.0 indicate a bootstrap-estimated interval wider than the "true" interval, and conversely.

Under a few conditions, the BC and ordinary percentile procedures procured slightly more accurate intervals than did the BC_a. In addition, the adjusted bootstrap intervals were, by and large, unacceptable, regardless of the robust measure under examination.

Regarding the width of the estimated intervals (see Table 5), no bootstrapping procedure clearly bettered the others. For small sample size conditions the adjusted bootstrap averaged relatively wider intervals. This widening effect improved accuracy for the narrow *r*- and r_z-generated intervals, but penalized r_w and r_{pb}. The BC_a intervals frequently ran short, the BC intervals shorter still, and the percentile bootstrap the shortest of the four. These trends were slight and not unexpected (e.g., it is widely known that the percentile bootstrap tends to produce narrow intervals).

Comparisons Among Correlations

Confidence intervals formed for r_w and r_{pb} generally outperformed those for *r* and r_z for both Type I error rate and bias. Although the present paper does not depict the data

disaggregated by distributional shape, predictable results surfaced. Under normality all four correlations produced similar Type I error rates, although the Pearson *r* and its transform saw slightly lower levels of bias, at least under small sample conditions. However, as distributional shape diverged from normality or included contaminated or mixed distributions, the robust correlations surpassed *r*. As an aside, the bias index generally produced neater, more theoretically consistent results than did Type I error rate. This is probably due to the dichotomous nature of the latter, that is, a given interval either does or does not enclose the parameter of interest and cause a Type I error, whereas bias is measured on the more sensitive ratio scale of measurement.

Regarding interval *width*, *r* and its transform consistently underestimated the "true" endpoints, more so under non-normal conditions. At times, such intervals were little more than half the "true" width. Bootstrapped confidence intervals for the percentage bend correlation closely mimicked the "true" intervals in almost every instance, and intervals for the Winsorized correlation tended to run slightly wide.

Conclusion

This study confirmed that the Winsorized and percentage bend correlations are useful alternatives to the Pearson correlation and are preferred when resilience to distributional non-normality is needed. Results for three of the four comparative indices (efficiency was virtually a constant) confirmed the robustness of the two robust measures under non-normal, mixed, and contaminated distributional conditions, with neither outperforming the other. The percentage bend and Winsorized correlations reduced bias, more accurately reflected theoretical Type I error probabilities, and more faithfully reproduced the width of true (Monte Carlo simulated) intervals. The robust measures compared favorably to *r* even under the bivariate normal conditions.

Interestingly, across a wide range of simulation conditions the four bootstrapping procedures achieved roughly equivalent outcomes as applied to either robust correlation. The complex BC and BC_a procedures failed to offer sizeable improvements in interval accuracy over the percentile bootstrap, and the "adjusted" bootstrap may have even inflated bias and Type I error rate. While this finding may be interpreted as disappointing because the more elaborate procedures did not offer increased accuracy, researchers can be more confident that the ordinary percentile bootstrap is capable of delivering relatively precise confidence intervals for these robust measures.

It may be that the more complicated procedures did not surpass the percentile bootstrap due to the technical specifications of the simulation. The original study design entailed drawing 1,000 samples for each condition, but this number was reduced to 100 given excessive computational demands. Even though the goal in this component of the simulation procedure is not to fully reproduce a sampling distribution, more samples may be necessary to achieve stable asymptotic dynamics. Similarly, the number of bootstrap samples had to be reduced considerably (e.g., setting *B* to 3,000 produced only 25 samples in eight hours due to the large number of simulated conditions and the involved calculations for r_{pb}). However, for this simulation component, the objective is indeed to model a theoretical sampling distribution, $F(\theta)$, via a bootstrapped sampling distribution, $\hat{F}(\hat{\theta}^*)$. Five hundred bootstrap samples may be sufficient for estimating standard errors (Efron, 1987; Efron & Tibshirani, 1993; but cf. Booth & Sarkar, 1998), but not for forming tight confidence bands (Lunneborg, 2000). Follow-up studies should increase these quantities if possible.

The study also revealed that Fisher's transformation of *r* did not appreciably improve either Type I error rate or bias. When bootstrapping the Pearson correlation, it seems that the *r*-to-*z* transformation merely increases computational time without concomitantly affecting accuracy, as supported by Seivers (1996) in his conclusions about r_z.

In sum, the robust measures may be recommended for general use when it is desired to quantify the linear association underlying the majority of the sample observations, while excluding outliers. Each of the bootstrapping procedures reviewed maintained similar levels of accuracy and may be applied in estimating confidence intervals for the robust correlations, excepting the adjusted bootstrap.

References

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40,* 318-335.

Booth, J. G., & Sarkar, S. (1998). Monte Carlo approximation and bootstrap variances. *American Statistician, 52,* 354-357.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. *American Statistician, 31,* 147-50.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144-152.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika, 62,* 531-545.

Duncan, G. T., & Layard, M. W. (1973). A Monte-Carlo study of asymptotically robust tests for correlation. *Biometrika, 60,* 551-558.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1-26.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics, 9,* 139-172.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika, 72,* 45-58.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82,* 171-185.

Efron, B. (1988). Bootstrap confidence intervals: Good or bad? *Psychological Bulletin, 104,* 293-296.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* New York: Chapman & Hall.

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics, 28,* 81-124.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics, 16,* 927-953.

Huber, P. J. (1981). *Robust Statistics.* New York: Wiley.

King, J. E. (2000). Bootstrapping robust measures of association (Doctoral dissertation, Texas A&M University, 2000). *Dissertation Abstracts International, 61*(11), 5948B.

King, J. E. (2003, February). *What have we learned from 100 years of robustness studies on r*? Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio, TX.

Lunneborg, C. E., (2000). *Data analysis by resampling: Concepts and applications.* Pacific Grove, CA: Brooks/Cole.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

Rasmussen, J. L. (1986). An evaluation of parametric and non-parametric tests on modified and non-modified data. *British Journal of Mathematical and Statistical Psychology, 39,* 213-220.

Sievers, W. (1996). Standard and bootstrap confidence intervals for the correlation coefficient. *British Journal of Mathematical and Statistical Psychology, 49,* 381-396.

Stigler, S. M. (1973). Simon Newcomb, Percy Daniell, and the history of robust estimation 1885-1920. *Journal of the American Statistical Association, 68,* 872-879.

Strube, M. J. (1988). Bootstrap Type I error rates for the correlation coefficient: An examination of alternate procedures. *Psychological Bulletin, 104,* 290-292.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61*(4), 361-377.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24-31.

Vacha-Haase, T., Nilsson, J. E., Reetz, D. R., Lance, T. S., & Thompson, B. (2000). Reporting practices and APA editorial policies regarding statistical significance and effect size. *Theory & Psychology, 10,* 413-425.

Wilcox, R. R. (1990). Comparing variances and means when distributions have non-identical shapes. *Communications in Statistics-Simulation and Computation, 19,* 155-173.

Wilcox, R. R. (1993). Some results on a Winsorized correlation coefficient. *British Journal of Mathematical and Statistical Psychology, 46,* 339-349.

Wilcox, R. R. (1994). The percentage bend correlation coefficient. *Psychometrika, 59,* 601-616.

Wilcox, R. R. (1997). Tests of independence and zero correlations among *p* random variables. *Biometrical Journal, 39,* 183-193.

Wilcox, R. R., & Muska, J. (2001). Inferences about correlations when there is heteroscedasticity. *British Journal of Mathematical & Statistical Psychology, 54,* 39-47.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604.