Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes

A tutorial using simulations and empirical data

de Winter, Joost C F; Gosling, S.D.; Potter, J.P.

DOI
10.1037/met0000079

Publication date
2016

Document Version
Final published version

Published in
Psychological Methods

Citation (APA)
de Winter, J. C. F., Gosling, S. D., & Potter, J. P. (2016). Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, *21*(3), 273-290. https://doi.org/10.1037/met0000079

Psychological Methods
2016, Vol. 21, No. 3, 273–290

1082-989X/16/$12.00 http://dx.doi.org/10.1037/met0000079

Comparing the Pearson and Spearman Correlation Coefficients Across Distributions and Sample Sizes: A Tutorial Using Simulations and Empirical Data

Joost C. F. de Winter
Delft University of Technology

Samuel D. Gosling
University of Texas at Austin and University of Melbourne

Jeff Potter
Atof Inc., Cambridge, Massachusetts

The Pearson product–moment correlation coefficient (r_p) and the Spearman rank correlation coefficient (r_s) are widely used in psychological research. We compare r_p and r_s on 3 criteria: variability, bias with respect to the population value, and robustness to an outlier. Using simulations across low ($N = 5$) to high ($N = 1{,}000$) sample sizes we show that, for normally distributed variables, r_p and r_s have similar expected values but r_s is more variable, especially when the correlation is strong. However, when the variables have high kurtosis, r_p is more variable than r_s. Next, we conducted a sampling study of a psychometric dataset featuring symmetrically distributed data with light tails, and of 2 Likert-type survey datasets, 1 with light-tailed and the other with heavy-tailed distributions. Consistent with the simulations, r_p had lower variability than r_s in the psychometric dataset. In the survey datasets with heavy-tailed variables in particular, r_s had lower variability than r_p, and often corresponded more accurately to the population Pearson correlation coefficient (R_p) than r_p did. The simulations and the sampling studies showed that variability in terms of standard deviations can be reduced by about 20% by choosing r_s instead of r_p. In comparison, increasing the sample size by a factor of 2 results in a 41% reduction of the standard deviations of r_s and r_p. In conclusion, r_p is suitable for light-tailed distributions, whereas r_s is preferable when variables feature heavy-tailed distributions or when outliers are present, as is often the case in psychological research.

Keywords: correlation, outlier, rank transformation, nonparametric versus parametric

Supplemental materials: http://dx.doi.org/10.1037/met0000079.supp

The Pearson product–moment correlation coefficient (r_p; Pearson, 1896) and the Spearman rank correlation coefficient (r_s; Spearman, 1904) were developed over a century ago (for a review see Lovie, 1995). Both coefficients are widely used in psychological research. According to a search of ScienceDirect, of the 18,419 articles published in psychology in 2014, 24.7% reported an effect size measure of some kind. As shown in Table 1, r_p and r_s are particularly popular in sciences involving the analysis of human behavior (social sciences, psychology, neuroscience, medicine). Table 1 further shows that r_p is reported about twice as frequently as r_s. Moreover, Table 1 almost certainly underestimates the prevalence of r_p, because r_p is the default option in many statistical packages; so when the type of correlation coefficient goes unreported, it is likely to be r_p.

Many more researchers use r_p rather than r_s, perhaps because r_p appears to match more closely the linear relationship they aim to estimate. Other reasons why most researchers choose r_p could be because r_p allows for inferences such as calculation of the variance accounted for, or because it is consistent with the methods of available follow-up analyses, such as linear regression (or ANOVA) by least squares or factor analysis by maximum likelihood. Yet another reason for the widespread use of r_p may be that statistical practices are very much determined by what SPSS, R, SAS, MATLAB, and other software manufacturers implement as their default option (Steiger, 2001, 2004). For example, in MATLAB, the command *corr(x,y)* yields the Pearson correlation coefficient between the vectors *x* and *y*. It requires a longer command (*corr(x,y),'type','spearman'*) to calculate the Spearman correlation. Thus, the software may im-

This article was published Online First May 23, 2016.
Joost C. F. de Winter, Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology; Samuel D. Gosling, Department of Psychology, University of Texas at Austin, and School of Psychological Sciences, University of Melbourne; Jeff Potter, Atof Inc., Cambridge, Massachusetts.
The datasets used in this research were obtained from the Transport Research Laboratory (2008), the Bureau of Labor Statistics (2002), and the Gosling-Potter Internet Personality Project. The principal investigator of the Gosling-Potter Internet Personality Project can be contacted to access the data from this project (samg@austin.utexas.edu).
Correspondence concerning this article should be addressed to Joost C. F. de Winter, Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD, Delft, the Netherlands. E-mail: j.c.f.dewinter@tudelft.nl

Table 1
Percentage of the Papers With Abstract Published in 2014 That Contain a Correlation or Effect Size Term, for Eight Selected Subject Areas

Search queries	1. Psychology	2. Neuroscience	3. Medicine and Dentistry	4. Social Sciences	5. Economics, Econometrics, and Finance	6. Computer Sciences	7. Engineering	8. Chemistry	All eight subject areas
Any of the keywords below	24.70%	19.18%	18.62%	12.56%	6.61%	4.15%	1.94%	1.17%	10.42%
ABS({.}) AND ALL("odds ratio" OR "risk ratio" OR "relative risk RR")	6.80%	5.60%	10.37%	4.21%	1.76%	0.46%	0.35%	0.08%	4.88%
ABS({.}) AND ALL("Pearson correlation" OR "Pearson product-moment" OR "Pearson r" OR "Pearson's correlation" OR "Pearson's product-moment" OR "Pearson's r")	9.37%	7.97%	4.21%	4.58%	2.85%	1.98%	0.97%	0.80%	3.01%
ABS({.}) AND ALL("Spearman rank" OR "Spearman correlation" OR "Spearman rho" OR "Spearman's rank" OR "Spearman's correlation" OR "Spearman's rho" OR "rank-order correlation")	3.36%	3.87%	3.11%	1.85%	1.70%	0.79%	0.39%	0.20%	1.81%
ABS({.}) AND ALL("intraclass correlation" OR "intra-class correlation" OR "intraclass r" OR "intra-class r")	3.24%	1.66%	1.63%	1.32%	0.20%	0.19%	0.11%	0.03%	0.85%
ABS({.}) AND ALL("Cohen's d" OR "Cohen d" OR "Cohen's effect size")	4.47%	2.18%	0.73%	1.17%	0.08%	0.22%	0.06%	0.00%	0.52%
ABS({.}) AND ALL("Cohen's kappa" OR "kappa statistic" OR "Cohen's k" OR "k-statistic")	1.12%	0.54%	0.73%	0.81%	0.27%	0.54%	0.11%	0.02%	0.44%
ABS({.}) AND ALL("Kendall tau" OR "Kendall correlation" OR "Kendall's tau" OR "Kendall's correlation")	0.23%	0.20%	0.10%	0.17%	0.45%	0.25%	0.09%	0.01%	0.11%
ABS({.}) AND ALL("Hedges's g" OR "Hedges g" OR "Hedges effect size")	0.58%	0.23%	0.10%	0.06%	0.01%	0.03%	0.01%	0.00%	0.06%
ABS({.}) AND ALL("Cramer's V" OR "Cramer's phi")	0.34%	0.13%	0.07%	0.21%	0.08%	0.03%	0.01%	0.00%	0.06%
ABS({.}) AND ALL("point biserial" OR "point bi-serial")	0.34%	0.14%	0.08%	0.12%	0.02%	0.03%	0.01%	0.00%	0.05%
ABS({.}) AND ALL("concordance correlation")	0.01%	0.02%	0.07%	0.03%	0.01%	0.02%	0.01%	0.01%	0.04%
ABS({.}) AND ALL("polychoric correlation" OR "tetrachoric correlation" OR "tetrachoric coefficient")	0.33%	0.11%	0.05%	0.12%	0.10%	0.01%	0.00%	0.00%	0.04%

Table 1 (*continued*)

Search queries	1. Psychology	2. Neuroscience	3. Medicine and Dentistry	4. Social Sciences	5. Economics, Econometrics, and Finance	6. Computer Sciences	7. Engineering	8. Chemistry	All eight subject areas
ABS({.}) AND ALL("RV coefficient" OR "congruence coefficient" OR "distance correlation" OR "Brownian correlation" OR "Brownian covariance")	0.09%	0.13%	0.02%	0.03%	0.04%	0.06%	0.03%	0.04%	0.04%
ABS({.}) AND ALL("Fleiss kappa")	0.11%	0.04%	0.06%	0.07%	0.01%	0.06%	0.01%	0.00%	0.03%
ABS({.}) AND ALL("correlation phi" OR "phi correlation" OR "mean square contingency coefficient" OR "Matthews correlation")	0.08%	0.04%	0.03%	0.03%	0.02%	0.14%	0.03%	0.04%	0.03%
ABS({.}) AND ALL("correlation ratio" OR "eta correlation")	0.02%	0.03%	0.02%	0.04%	0.02%	0.04%	0.01%	0.00%	0.02%
Total number of publications in 2014	18,419	33,758	131,076	32,137	12,261	26,120	64,616	53,604	297,669*

Note. This table is based on a full-text search of ScienceDirect conducted on October 9, 2015. Searching for "correlation coefficient" while excluding all search terms in Table 1 yielded 9,443 articles; in other words, the type of correlation coefficient often goes unreported.
* "All eight subject areas" is not the sum of the eight columns, but the number of articles retrieved when searching in all eight subject areas simultaneously. This number is smaller than the sum of the publications in the eight individual subject areas because some articles are classified in two or more subject areas.

plicitly give the impression that r_p is the preferred option and it also requires more knowledge of the software commands to calculate r_s.

Some Well-Known and Less Well-Known Properties of r_p and r_s

The sample Pearson correlation coefficient r_p is defined according to Equation 1. Here, we have first performed a mean centering procedure on the x and y vectors.

$$r_p = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i^2}} \tag{1}$$

The sample Spearman correlation coefficient r_s is calculated in the same manner as r_p, except that r_s is calculated after both x and y have been rank transformed to values between 1 and N (Equation 2). When calculating r_s, a so-called fractional ranking is used, which means that the mean rank is assigned in case of ties. For example, suppose that the two smallest numbers of x are equal, then they will be both ranked as 1.5 (i.e., [1 + 2]/2). Again, a mean centering is first performed (by subtracting $N/2 + 1/2$ from each of the two ranked vectors).

$$r_s = \frac{\sum_{i=1}^{N} x_{i,r} y_{i,r}}{\sqrt{\sum_{i=1}^{N} x_{i,r}^2 \sum_{i=1}^{N} y_{i,r}^2}} \tag{2}$$

Assuming there are no ties, Equation 2 can be rewritten in various formats (Equation 3).

$$r_s = \frac{\sum_{i=1}^{N} x_{i,r}^2 - \frac{1}{2}\sum_{i=1}^{N} (x_{i,r} - y_{i,r})^2}{\sum_{i=1}^{N} x_{i,r}^2} = 1 - \frac{\sum_{i=1}^{N} (x_{i,r} - y_{i,r})^2}{2\sum_{i=1}^{N} x_{i,r}^2}$$

$$= 1 - \frac{6\sum_{i=1}^{N} (x_{i,r} - y_{i,r})^2}{N(N^2-1)} = \frac{12}{N(N^2-1)}\sum_{i=1}^{N} x_{i,r} y_{i,r} \tag{3}$$

It can be inferred from Equations 1–3 that r_p will be high when the individual points lie close to a straight line, whereas r_s will be high when both vectors have a similar ordinal relationship. As mathematically shown by Yuan and Bentler (2000), the distribution of r_p depends only on the fourth-order moments (or kurtoses) of the two variables, not on their skewness (see also Yuan, Bentler, & Zhang, 2005). After all, r_p is a function of second-order sample moments, and so the variance of r_p is determined by fourth-order moments. The nonparametric measure r_s, on the other hand, is relatively robust to heavy-tailed distributions and outliers; all data are transformed to values ranging from 1 to N, so the influence function is bounded (Croux & Dehon, 2010). Several of the above characteristics of r_p and r_s are covered in many introductory statistics books and graduate-level psychology programs. Furthermore, a large number of research papers have previously described the differences between r_p and r_s, and have confirmed that r_s has attractive robustness properties (e.g., Bishara & Hittner, 2015; Fowler, 1987; Hotelling & Pabst, 1936).

Nonetheless, several characteristics of r_p and r_s may not be well known to researchers, even for the standard scenario of normally distributed variables. The derivation of the probability density function of r_p for bivariate normal variables can be traced back to contributions by Fisher (1915), Sawkins (1944), Hotelling (1951, 1953), and Kenney and Keeping (1951), and was reported more recently by Shieh (2010):

$$f(r_p) = \frac{(N-2)\left(1-R_p^2\right)^{\frac{(N-1)}{2}}\left(1-r_p^2\right)^{\frac{(N-4)}{2}}}{\sqrt{N}(N-2)\beta\left(\frac{1}{2}, N-\frac{1}{2}\right)\left(1-R_p r_p\right)^{N-\frac{3}{2}}} \times {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; N-\frac{1}{2}; \frac{R_p r_p + 1}{2}\right) \tag{4}$$

Here, R_p is the population Pearson correlation coefficient, β is the beta function, and ${}_2F_1$ is Gauss' hypergeometric function. The hypergeometric function is available in software packages (e.g., *hypergeom ([1/2 1/2], N-1/2, (R_p * r_p+1)/2)* in MATLAB), but can also be readily calculated according to a power series, with Γ being the gamma function:

$${}_2F_1\left(\frac{1}{2}, \frac{1}{2}; N-\frac{1}{2}; \frac{R_p r_p + 1}{2}\right) = \sum_{i=0}^{\infty}\left(\frac{\Gamma\left(\frac{1}{2}+i\right)^2 \Gamma\left(N-\frac{1}{2}\right)}{\pi \cdot \Gamma\left(N-\frac{1}{2}+i\right)} \frac{\left(\frac{R_p r_p + 1}{2}\right)^i}{i!}\right) \tag{5}$$

Shieh (2010) stated "It is not well understood that the underlying probability distribution function of r is complicated in form, under the classical assumption that the two variables follow a bivariate normal distribution. The complexity incurs continuous investigation" (p. 906). Figure 1 illustrates the probability density function of r_p for two sample sizes (N = 5 and 50) and three population correlation coefficients (R_p = .2, .4, and .8). It can be seen that the mode of the distribution is greater than R_p and that the distribution is negatively skewed, with the skew being stronger for higher R_p and for smaller N.

Equation 4 allows one to calculate exact p values and confidence intervals. However, the popular and considerably more straightforward Fisher transformation can also be used in statistical inference (e.g., Fisher, 1921; Fouladi & Steiger, 2008; Hjelm & Norris, 1962; Hotelling, 1953; Winterbottom, 1979). For r_s, exact probability density functions are available for small sample sizes, and over the years various approximations (in terms of bias, mean squared error, and relative asymptotic efficiency) of the distribution and its moments have been published (Best & Roberts, 1975; Bonett & Wright, 2000; Croux & Dehon, 2010; David & Mallows, 1961; David, Kendall, & Stuart, 1951; Fieller, Hartley, & Pearson, 1957; Xu, Hou, Hung, & Zou, 2013). Furthermore, several variance-stabilizing transformations have been developed for r_s. These transformations, which can be applied in analogous fashion to the Fisher z transformation for r_p, may be practical for statistical inference purposes (Bonett & Wright, 2000; Fieller et al., 1957; but see Borkowf, 2002 demonstrating limitations of this concept).

Typically in psychology, investigators undertake research on samples (i.e., a subset of the population) with the aim of estimating the true relationships in the population. It is useful to point out that the expected values of both r_p and r_s are biased estimates of their respective population coefficients R_p and R_s (Ghosh, 1966; Zim-

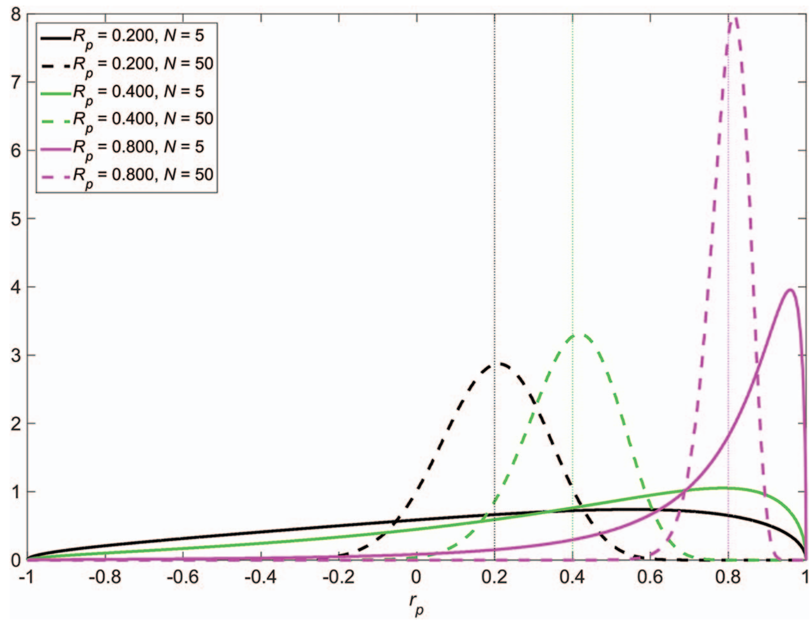

Figure 1. Probability density function of the Pearson correlation coefficient (r_p) for three levels of the population Pearson correlation coefficient ($R_p = .2$, $R_p = .4$, $R_p = .8$) and two levels of sample size ($N = 5$, $N = 50$). The area under each curve equals 1. See the online article for the color version of this figure.

merman, Zumbo, & Williams, 2003). Zimmerman et al. (2003) stated "It is not widely recognized among researchers that this bias can be as much as .03 or .04 under some realistic conditions" (p. 134). Equation 6 provides the expected value of r_p (Ghosh, 1966), while Equation 7 provides the expected value of r_s (Moran, 1948; Xu et al., 2013; Zimmerman et al., 2003). Both these equations indicate that the population value is underestimated, especially for small N. This underestimation is relatively small if R_p is small or moderate. For example, if $R_p = .2$ (corresponding $R_s = .191$, calculated using Equation 9), then $E(r_p)$ and $E(r_s)$ are .177 and .160, respectively at $N = 5$, and .195 and .182 at $N = 20$. The underestimation is more severe for R_p between .3 and .9. If $R_p = .8$ ($R_s = .786$), then $E(r_p)$ and $E(r_s)$ are .754 and .688 at $N = 5$, and .792 and .758 at $N = 20$.

$$E(r_p) = \frac{2\left(\Gamma\left(\frac{N}{2}\right)\right)^2}{(N-1)\left(\Gamma\left(\frac{N-1}{2}\right)\right)^2} R_p \cdot {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; \frac{N+1}{2}; R_p^2\right) \quad (6)$$

$$E(r_s) = \frac{6}{\pi(N+1)}\left(\arcsin(R_p) + (N-2)\arcsin\left(\frac{R_p}{2}\right)\right) \quad (7)$$

Equation 7 can be rewritten into a form that clarifies how the expected value of r_s relates to the population value of the Spearman coefficient and another well-known rank coefficient, Kendall's tau (Durbin & Stuart, 1951; Hoeffding, 1948).

$$E(r_s) = \frac{(N-2)R_s + 3R_t}{N+1} \quad (8)$$

The Pearson, Spearman, and Kendall correlation coefficients at the population level (i.e., R_p, R_s, R_t) for normally distributed variables can be described by a closed-form expression (e.g., Croux & Dehon, 2010; Pearson, 1907). In other words, for an infinite sample size, the Pearson, Spearman, and Kendall correlation coefficients differ when the two variables are normally distributed (Equations 9, 10, and 11).

$$R_s = \frac{6}{\pi}\arcsin\left(\frac{R_p}{2}\right) \quad (9)$$

$$R_t = \frac{2}{\pi}\arcsin(R_p) \quad (10)$$

$$R_s = \frac{6}{\pi}\arcsin\left(\frac{\sin\left(\frac{1}{2}\pi R_t\right)}{2}\right) \quad (11)$$

The maximum difference between R_p and R_s is .0181 and occurs at $R_p = .594$ $\left(\frac{\sqrt{4\pi^2 - 36}}{\pi}\right)$ and $R_s = .576$ $\left(\frac{6}{\pi}\arcsin\left(\frac{\sqrt{\pi^2 - 9}}{\pi}\right)\right)$, see also Guérin, De Oliveira, and Weber (2013). Figure S1 of the supplementary material illustrates the relationships between R_p, R_s, and R_t (see also Kruskal, 1958).

Aim of the Present Study

As shown above, the definitions and essential characteristics of r_p and r_s are probably well known. However, r_p and r_s exhibit a variety of interesting features in the case of bivariate normality. Of course, in real-life scenarios, psychologists are likely to encounter non-normal data as well.

In light of the widespread use of correlations in psychology and the predominance of r_p over r_s, the goal of this contribution is to review the properties of the r_p versus r_s, and to clarify the situations in which r_p or r_s should be preferred. We examine the properties of both coefficients with the aim of providing researchers with empirically derived guidance about which coefficient to use.

We use simulations and analyses of existing datasets to compare r_p with r_s for conditions that are representative of those found in psychological research. We start out by comparing r_p versus r_s for normally distributed variables, which as we indicated above, may have various unfamiliar properties. We aim to depict the characteristics of r_p and r_s in an intuitive, graphical manner. Next, we evaluate r_p versus r_s when the two variables have a non-normal distribution, a situation that is common in psychological research. We also graphically illustrate the strength of r_s when one or more outliers are present. Finally, we provide a demonstration of the differences of r_p versus r_s for typical psychological data. The main contribution of these sampling studies is to explain the relative performance of r_p versus r_s as a function of item/scale characteristics and sample size. In all cases, we compare the two coefficients in terms of variability, bias with respect to the population value, and robustness to an outlier.

r_p Versus r_s With a Normally Distributed Population

Normally Distributed Variables in Psychological Research

The central limit theorem states that the sum of a large number of independent random variables conforms to a normal distribution. Psychologists often aggregate data into constructs, and furthermore, various types of human attributes (such as personality and intelligence) may be seen as the effect of a large number of unobserved random processes. Hence, the central limit theorem can explain why certain psychological variables are *approximately* normally distributed (see Lyon, 2014, for a discussion on the factors that contribute to normality). Intelligence and physical ability are prime examples of human attributes that follow an approximately normal distribution (Burt, 1957; Plomin & Deary, 2015). The normal distribution occurs empirically regardless of whether the attribute is measured on an ordinal scale (e.g., a paper and pencil intelligence test) or on a ratio scale (e.g., intelligence defined chronometrically; Jensen, 2006). Let us therefore first evaluate how r_p and r_s behave when the two variables are normally distributed.

Selected Population Correlation Coefficients

To describe the behavior of r_p and r_s for bivariate normal variables and finite sample sizes, we undertook a simulation study. To ensure that the ranges of coefficient sizes were representative of those potentially encountered in psychological research, we consulted the literature. In published research, correlations among psychometric test scores, and correlations between psychological assessment scores and performance criteria, generally range between 0 and .5 (cf. Jensen, 2006; Meyer et al., 2001; Tett, Jackson, & Rothstein, 1991). One review of 322 meta-analyses showed that the absolute correlation coefficients in social psychology average at .21, with 95% of the coefficients between 0 and .5, and the remaining 5% between .5 and .8 (Richard, Bond, & Stokes-Zoota, 2003). Only variables that are conceptually similar to one another, such as intelligence test scores and scholastic performance, will correlate as highly as .8 (Deary, Strand, Smith, & Fernandes, 2007; Frey & Detterman, 2004). In short, population correlations between 0 and .8 reflect the range found in virtually all psychological/behavioral research. Therefore, simulation studies were performed with population Pearson correlation coefficients that were zero ($R_p = 0$), moderate ($R_p = .2$), strong ($R_p = .4$), and very strong ($R_p = .8$). The corresponding population Spearman correlation coefficients (R_s) were calculated according to Equation 9.

Selected Sample Sizes

Sample sizes used by psychologists are known to vary widely. One analysis of hundreds of articles (Marszalek, Barber, Kohlhart, & Holmes, 2011) showed that in the *Journal of Experimental Psychology* in the year 2006, the median total sample size was 18 ($Q_1 = 10$, $Q_3 = 32$), whereas in the *Journal of Applied Psychology*, the mean sample size was 148 ($Q_1 = 45$, $Q_3 = 269$). Fraley and Vazire (2014) showed that the median sample size in five high-impact psychological journals in the years 2006–2010 ranged between 73 ($Q_1 = 41$, $Q_3 = 143$) for *Psychological Science* and 178 ($Q_1 = 100$, $Q_3 = 344$) for the *Journal of Personality* (we calculated the interquartile ranges from the supplementary material of Fraley & Vazire, 2014). Here we note that personality psychology is more likely than experimental psychology to use correlation coefficients (e.g., Cronbach, 1957; Tracy, Robins, & Sherman, 2009), and so a sample size of about 200 is regarded as typical for correlational analyses. This sample size is in line with a recent simulation study that investigated at which sample size correlations stabilize, and which concluded that "there are few occasions in which it may be justifiable to go below $n = 150$ and for typical research scenarios reasonable trade-offs between accuracy and confidence start to be achieved when n approaches 250" (Schönbrodt & Perugini, 2013, p. 611).

To cover the range of sample sizes found in psychological research, we used 25 sample sizes (Ns) logarithmically spaced between 5 and 1,000. To generate stable estimates of r_p and r_s, for each sample size, 100,000 samples of variable 1 (hereafter called x) and variable 2 (hereafter called y) were drawn, and r_p and r_s were calculated for each of the 100,000 samples.

Results of the Simulations

The simulation results for $R_p = .2$ are shown in Figure 2. The mean r_s is slightly lower than the mean r_p, for all sample sizes. For small sample sizes, the mean r_p and mean r_s are both slight underestimates of their respective population values R_p and R_s (see also Equations 6 and 7). Figure 2 also shows how the absolute variability decreases with sample size for both r_p and r_s. However, r_s has a slightly higher variability, with the standard deviation of r_s being about 0.7% greater than the standard deviation of r_p, for each tested sample size. Similarly, the root mean squared error (RMSE) of r_s with respect to R_s is 0.7% greater than the RMSE of r_p with respect to R_p.

Note that r_s can take on only a distinct number of values, rapidly increasing with increasing N (Sloane, 2003; sequence A126972). For example, for $N = 5$, r_s can be only 1 of 21 different values (−1, −.9, −.8, . . . , .8, .9, 1; see Figure S2 for an illustration of the distribution of r_p and r_s at $N = 5$). The supplementary material (Figures S3, S4, and S5) includes the distributions of r_p and r_s for $R_p = 0$, $R_p = .4$, and $R_p = .8$. For $R_p = 0$, r_p and r_s behave almost identically. For $R_p = .4$, the standard deviation of r_s is 3% to 4% higher than the standard deviation of r_p, and for $R_p = .8$, the standard deviation of r_s is as much as 18% higher than the standard

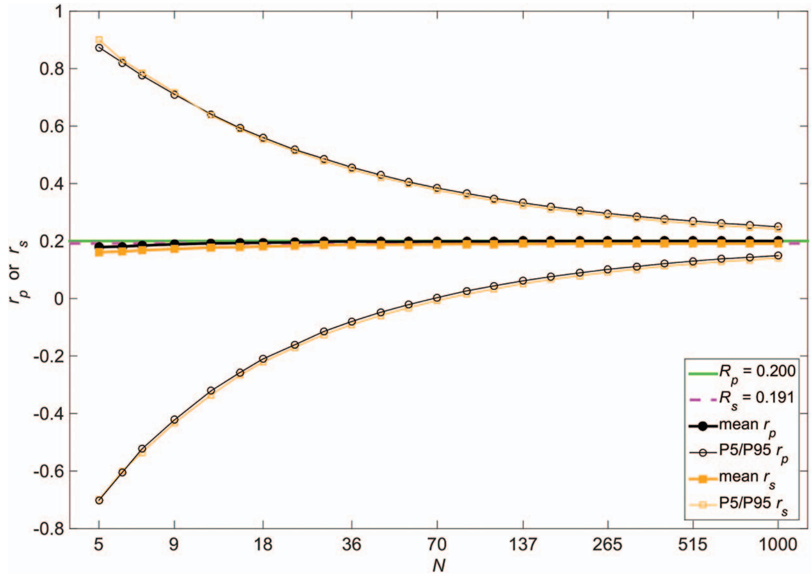

Figure 2. Simulation results for normally distributed variables having a population Pearson correlation coefficient of .2 (R_p = .2). The population Spearman correlation coefficient (R_s) was calculated according to Equation 9. The figure shows the mean, 5th percentile (P5), and 95th percentile (P95) of the Pearson correlation coefficient (r_p) and the Spearman correlation coefficient (r_s) as a function of sample size (N). See the online article for the color version of this figure.

deviation of r_p. The smaller variability of r_p compared with r_s is consistent with previous research (Bonett & Wright, 2000; Croux & Dehon, 2010; Fieller et al., 1957) and suggests that when both population variables are known to have approximately normal distributions, r_p should be used instead of r_s, especially when the correlation is thought to be strong.

r_p Versus r_s With a Non-Normally Distributed Population

Non-Normally Distributed Variables in Psychological Research

It frequently happens that psychological measurements feature a non-normal distribution. For example, it is known that psychiatric and other types of disorders follow a skewed distribution among individuals (Delucchi & Bostrom, 2004; Keats & Lord, 1962; McGrath et al., 2004). Yet in other cases, measurement scales may be limited by artifacts such as ceiling and floor effects (Van den Oord, Pickles, & Waldman, 2003). One analysis of 693 distributions of cognitive measures and other psychological variables with sample sizes ranging from 10 to 30 showed that 39.9% of the distributions were considered as slightly non-normal, 34.5% as moderately non-normal, 10.4% as highly non-normal, and a further 9.6% as extremely non-normal (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013). Another analysis of 440 large-sample distributions of achievement and psychometric data classified 31% of the distributions as extremely asymmetric, and 49% as having at least one extremely heavy tail (Micceri, 1989).

Selected Kurtosis of the Marginal Distributions

In light of these kinds of observations, we explored the behavior of r_p and r_s for two correlated variables having leptokurtic distributions, meaning that kurtosis was greater than would be expected from a normal distribution (see Figure 3 for illustration, and DeCarlo, 1997, for an explanation of kurtosis). The variables x and y were approximately exponentially distributed (hence, skewness = 2 and kurtosis = 9) and strongly correlated (R_p = .4). We

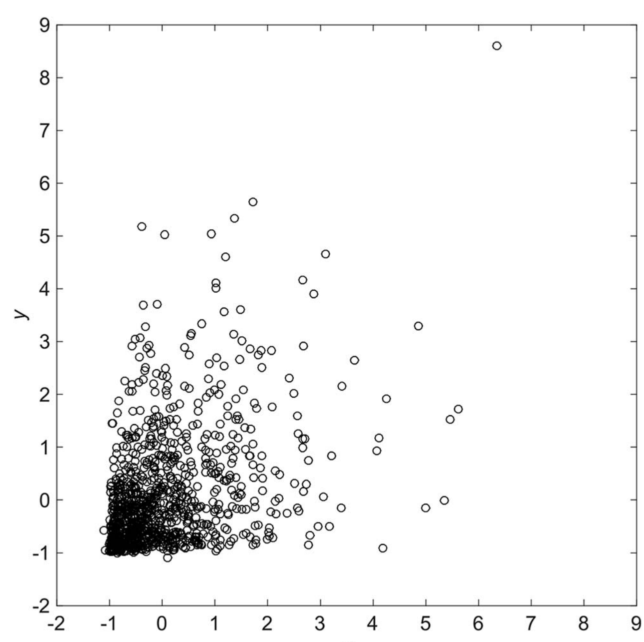

Figure 3. Depiction (using N = 1,000) of two correlated variables having an exponential distribution with population Pearson correlation coefficient (R_p) of .4. R_p was obtained by calculating r_p for a sample of $N = 10^7$.

used a fifth-order polynomial transformation method for generating the correlated non-normally distributed variables (Headrick, 2002). Because R_p and R_s could not be determined exactly, we defined these parameters by calculating the correlation coefficients for a very large sample size ($N = 10^7$).

Results of the Simulations

Figure 4 shows the distributions of r_p and r_s for the same range of sample sizes as those used to create Figure 2. It can be seen that the expected values of r_p and r_s are about the same and unbiased with respect to their respective population values, but r_p is more variable than r_s. Specifically, the standard deviation of r_p is 13.5%, 26.0%, and 27.3% greater than the standard deviation of r_s, for $N = 18$, $N = 213$, and $N = 1,000$, respectively. Similarly, the RMSE of r_p with respect to R_p is 13.0%, 25.9%, and 27.3% greater than the RMSE of r_s with respect to R_s, for $N = 18$, $N = 213$, and $N = 1,000$, respectively.

Additional Simulation Results With Other Kurtosis and r_p

If the two variables have greater kurtosis than exponentially distributed variables, then r_p is likely to be even more variable (see Figure S6 of the supplementary material). Also note that the size of the correlation coefficient is an important determinant of the behavior of r_p and r_s. For example, when choosing $R_p = .2$ instead of $R_p = .4$, the standard deviation of r_p is only 8.0%, 14.5%, and 15.5% greater than the standard deviation of r_s, for $N = 18$, $N = 213$, and $N = 1,000$, respectively. However, for $R_p = .8$, the standard deviation of r_p is 13.5%, 36.0%, and 38.9% greater than the standard deviation of r_s, for $N = 18$, $N = 213$, and $N = 1,000$, respectively (see Figures S10–S13).

In summary, our simulations showed that when the two variables have leptokurtic distributions, r_p is likely to be more variable than r_s. These observations are consistent with theory showing that the standard deviation of r_p is proportional to the kurtosis of the variables (Yuan & Bentler, 2000). Moreover, our results are in line with several simulation studies which demonstrated lower variability of r_s compared with r_p for (severely) non-normal distributions (Bishara & Hittner, 2015; Chok, 2010; Kowalski, 1972). Obviously, our set of simulations provide only a snapshot of the constellation of the bivariate relationships that may occur in psychological research. Furthermore note that when the two variables are mesokurtic or platykurtic (i.e., kurtosis ≤ 3), r_p will tend to be more stable than r_s.

r_p Versus r_s When There are Outliers

It has been well documented that the Pearson correlation coefficient is sensitive to outliers (e.g., Chok, 2010; Croux & Dehon, 2010). Formal treatments of so-called "influence functions" or "expected resistance" of r_p and r_s can be found in Blair and Lawson (1982), Zayed and Quade (1997), and Croux and Dehon (2010). Herein, we graphically and numerically illustrate how r_p and r_s respond to adding a spurious data point in conditions that are likely to occur in psychological research.

Although sample sizes in psychological research vary widely, we used $N = 200$ because this is in line with typical sample sizes used in applied and personality psychology (Fraley & Vazire, 2014; Marszalek et al., 2011). A sample ($N = 200$) was drawn from two standard normal distributions having a moderate inter-

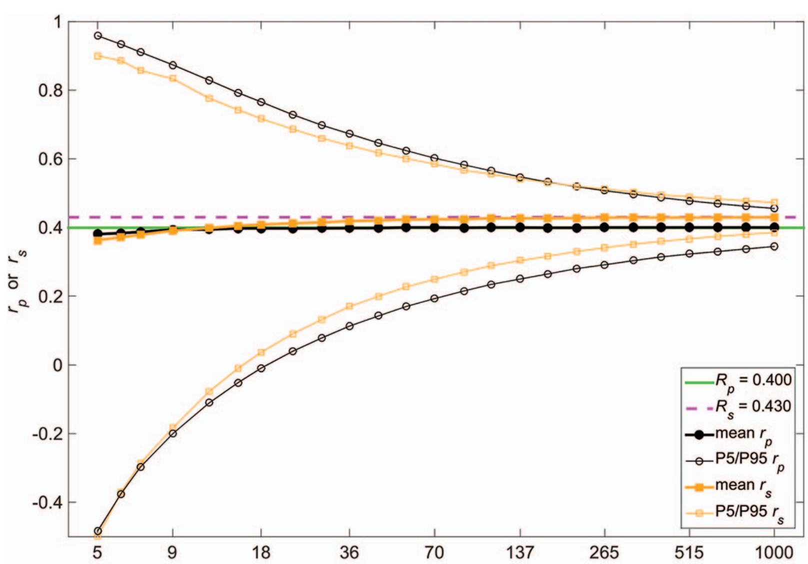

Figure 4. Simulation results for two correlated variables having an exponential distribution (see Figure 3 for a large-sample illustration of the distribution). The figure shows the mean, 5th percentile (P5), and 95th percentile (P95) of the Pearson correlation coefficient (r_p) and the Spearman correlation coefficient (r_s) as a function of sample size (N). The population coefficients R_p and R_s were obtained by calculating r_p and r_s, respectively, for a sample of $N = 10^7$. See the online article for the color version of this figure.

relationship in the population (R_p = .2). Next, one data point was added so that N = 201. The value of the spurious data point was systematically varied from −5 to 5 with a resolution of 0.05 for the two variables, x and y. Accordingly, 40,401 (i.e., 201 × 201) r_ps and 40,401 r_ss were determined.

Figure 5 illustrates the influence of the added (201st) data point on the obtained r_p and r_s, respectively. It can be seen that r_p is sensitive to this data point. Specifically, r_p equaled .231 without the data point, and has values between .100 (at $x = -5$, $y = 5$) and .312 (at $x = 5$, $y = 5$) by including it, with 19% of the r_ps differing by more than .05 from the original r_p of .231. In contrast, r_s is robust: r_s equals .222 without the extra data point, and adding it results in r_s values between .204 and .233. r_s is robust to outliers because the data in x and y are transformed to integers between 1 and N. This means it is impossible for very low or very high values in x or y to have a large effect on r_s.

Of course, in most real data there may be more than one outlier. Suppose, for example, that one outlier is located at $x = 5$ and $y = 5$, then adding a second outlier at all possible positions between −5 and 5 results in an r_p ranging between .186 and .377 (N = 202), with 77% of the r_ps differing by more than .05 from the original r_p of .231. Now suppose that the first outlier is at $x = 5$ and $y = -5$, then adding the second outlier results in an r_p between −.003 and .191. Again, r_s is robust, and always between .186 and .245 when two outliers were present. So, having more than one outlier can create even more problems for r_p, as the second outlier does not alleviate the distortive effect of the first outlier.

Five Demonstrations Using Empirical Data

The simulations above are indicative of the differences between r_p and r_s for normally and non-normally distributed variables. However, the simulations do not necessarily reflect situations encountered by empiricists. To test r_p versus r_s on data likely to be found in psychological studies, we undertook a sampling study using empirical data.

Selected Datasets

Three large datasets were used: a psychometric test battery (Armed Services Vocational Aptitude Battery; ASVAB), and two survey-based datasets: 5-point Likert scale data from the Big Five Inventory (BFI) and 6-point scale data from the Driver Behaviour Questionnaire (DBQ). The ASVAB, BFI, and DBQ datasets were all large (N = 11,878, N = 1,895,753, and N = 9,077, respectively), and were therefore used as populations from which we could draw samples to calculate sample correlation coefficients. The ASVAB consists of 10 very strongly intercorrelated test results, each symmetrically distributed with light tails (see Table 2). Recall that the simulation results above showed that r_p is less variable than r_s for normally distributed variables that are strongly correlated, so we expected the ASVAB sampling results to reflect these findings. The primary difference between the BFI and DBQ is that the BFI items have low kurtosis because the means of all 44 items are close to the middle option on the 5-point scale (see Table 2). In contrast, the DBQ items are leptokurtic, with the majority of participants reporting that they "never" make a certain error or violation in traffic (see also Mattsson, 2012). In light of the above simulation results, we expected r_s to outperform r_p for the DBQ dataset, and to a lesser extent for the BFI dataset.

Sampling Study 1: ASVAB. The ASVAB dataset is a psychometric dataset consisting of 11,878 subjects who, in the framework of the National Longitudinal Survey of Youth 1979, had taken a test battery (Bureau of Labor Statistics, 2002; Frey & Detterman, 2004; Maier & Sims, 1986; Morgan, 1983). The population included 5,951 men and 5,927 women. The mean age of the subjects was 18.8 years (SD = 2.3). The ASVAB consists of 10 tests (general science [25 items], arithmetic reasoning [30 items], word knowledge [35 items], paragraph comprehension [15 items],

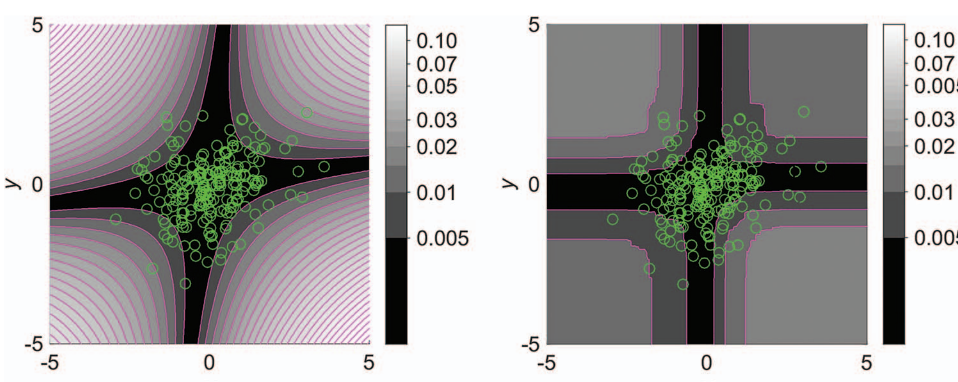

Figure 5. Simulation results demonstrating the influence of a spurious data point at location (x, y) on the Pearson correlation coefficient (left figure) and on the Spearman correlation coefficient (right figure). The circles represent a sample (N = 200) drawn from two standard normal distributions with population Pearson correlation coefficient (R_p) = .2. The sample Pearson correlation coefficient (r_p) = .231. The sample Spearman correlation coefficient (r_s) = .222. The grayscale background represents the absolute deviation from r_p (left figure) and the absolute deviation from r_s (right figure), after adding one data point so that N = 201. Isolines are drawn at every 0.005 increment. The vertical bars next to each figure signify the numeric values corresponding to a particular level of grayness. The value of the data point was systematically varied from −5 and +5 with a resolution of 0.05 for the two variables, x and y. See the online article for the color version of this figure.

Table 2
Means, Standard Deviations, Minima, and Maxima of Absolute Population Correlation Coefficients, and of Population Skewness and Population Kurtosis of the Items/Scales

Measure	ASVAB, 45 correlations	BFI items, 946 correlations	BFI scales, 10 correlations	DBQ items, 561 correlations	DBQ scales, 10 correlations
$\|R_p\|$					
Mean	.6273	.1206	.1778	.1713	.4197
SD	.1205	.1146	.0985	.0790	.1363
Min	.3317	.0002	.0283	.0003	.1511
Max	.8247	.5973	.3158	.5106	.5805
$\|R_s\|$					
Mean	.6281	.1223	.1690	.1622	.4157
SD	.1207	.1148	.0935	.0748	.1155
Min	.3362	.0001	.0309	.0024	.1742
Max	.8336	.6029	.3035	.4747	.5362
	ASVAB, 10 tests	BFI, 44 items	BFI, 5 scales	DBQ, 34 items	DBQ, 5 scales
Skewness					
Mean	−0.02	−0.37	−0.21	2.19	1.65
SD	0.40	0.42	0.10	1.44	0.72
Min	−0.59	−1.33	−0.30	0.50	0.83
Max	0.50	0.42	−0.06	6.42	2.46
Kurtosis					
Mean	2.32	2.54	2.89	11.96	8.89
SD	0.18	0.67	0.15	13.90	5.17
Min	2.03	1.80	2.74	3.16	4.03
Max	2.73	4.73	3.08	60.89	16.61

Note. R_p = population Pearson correlation coefficient; R_s = population Spearman correlation coefficient; ASVAB = Armed Services Vocational Aptitude Battery; BFI = Big Five Inventory; DBQ = Driver Behaviour Questionnaire. Skewness was defined as the third central moment divided by the cube of the standard deviation. Kurtosis was defined as the fourth central moment divided by the fourth power of the standard deviation. Kurtosis of a normal distribution = 3. R_p and R_s were defined as the correlation coefficients for the total sample (i.e., N = 11,878 for the ASVAB, N = 1,895,753 for the BFI, and N = 9,077 for the DBQ). The population skewness and population kurtosis have a strong correlation (ASVAB: r_s between skewness and kurtosis = −.50 [N = 10 items]), BFI items: r_s = −.83 [N = 44 items], BFI scales: r_s = −.70 [N = 5 scales], DBQ items: r_s = .99 [N = 34 items], DBQ scales: r_s = 1.00 [N = 5 scales]).

numerical operations [50 items], coding speed [84 items], auto and shop information [25 items], mathematics knowledge [25 items], mechanical comprehension [25 items], and electronics information [10 items]). The Pearson correlation matrix among the 10 variables contained 45 (= 10 ∗ (10 − 1)/2) unique elements. The maximum R_p was .825, occurring between "general science" and "word knowledge" (corresponding R_s = .834). The distribution of the variables was symmetric and platykurtic, that is, having somewhat lighter tails than would be expected from a normal distribution (see Table 2).

Sampling Study 2: BFI items. The BFI is a 44-item personality questionnaire answered on a Likert scale from 1 = *disagree strongly* to 5 = *agree strongly*. The BFI data (N = 3,093,144) were obtained via noncommercial, advertisement-free Internet web sites between 1999 and 2013 as part of the Gosling-Potter Internet Personality Project (e.g., Bleidorn et al., 2013; Gosling, Vazire, Srivastava, & John, 2004; Obschonka, Schmitt-Rodermund, Silbereisen, Gosling, & Potter, 2013; Rentfrow et al., 2013; Srivastava, John, Gosling, & Potter, 2003). Only participants who filled in the English version of the inventory, who answered all items without giving identical answers to all 44 items, and who were between 18 to 98 years were included, leaving a dataset of 1,895,753 respondents. The mean age of the respondents was 28.2 (median = 25.0, *SD* = 10.4). The population included 921,670 women and 651,914 men, and the sex was unknown for a further 322,169 respondents. The average mean response across the 44 items was 3.45 (*SD* = 0.47), with a minimum mean of 2.48 for the item "is depressed blue" and a maximum mean of 4.33 for the item "is a reliable worker." The BFI correlation matrix contained 946 (= 44 ∗ (44 − 1)/2) unique off-diagonal elements. The maximum R_p was .597, occurring between "is talkative" and "is outgoing, sociable" (corresponding R_s = .595). The variables were symmetric with low kurtosis (see Table 2).

Sampling Study 3: BFI scales. Psychological researchers often conduct their analysis at the scale level instead of the item level, so we also carried out the sampling study based on the five BFI scales. The following five sum scores were calculated: agreeableness (9 items), conscientiousness (9 items), extraversion (8 items), openness (10 items), and neuroticism (8 items). The 10 off-diagonal R_ps ranged between −.32 (for agreeableness vs. neuroticism; corresponding R_s = −.30) and .28 (for agreeableness vs. conscientiousness; corresponding R_s = .28). Table 2 shows that the five scales were fairly symmetric with low kurtosis.

Sampling Study 4: DBQ items. The DBQ dataset consisted of 9,077 respondents who, as part of a cohort study of learner and new drivers, had responded to the query "when driving, how often do you do each of the following?" with respect to 34 items

(Transport Research Laboratory, 2008; Wells, Tong, Sexton, Grayson, & Jones, 2008). The responses ranged from 1 = *never* to 6 = *nearly all the time.* The mean age of the respondents was 22.6 years (median = 18.7; *SD* = 8.1). The population consisted of 5,754 women and 3,323 men. The average mean response across the 34 items was 1.46 (*SD* = 0.26), with a minimum mean of 1.05 and a maximum mean of 2.06. The correlation matrix contained 561 (= 34 ∗ (34 − 1)/2) unique off-diagonal elements. The maximum R_p was .511 (between "Disregard the speed limit on a motorway" and "Disregard the speed limit on a residential road") with a corresponding R_s of .475. Items were highly skewed and leptokurtic (see Table 2).

Sampling Study 5: DBQ scales. The DBQ analysis was repeated at the scale level. The following five sum scales were calculated (as in Wells et al., 2008): violations (6 items), errors (8 items), aggressive violations (6 items), inexperience errors (7 items), and slips (7 items). The 10 off-diagonal R_ps ranged between .151 (between aggressive violations and inexperience errors; corresponding R_s = .174) and .581 (between violations and aggressive violations; corresponding R_s = .536). As with the DBQ items, the DBQ scales had high kurtosis, but the scale data were more strongly intercorrelated than the item data (see Table 2).

Sampling Methods

For each of the five datasets (i.e., ASVAB, BFI items, BFI scales, DBQ items, and DBQ scales), 50,000 random sample of N = 200 were drawn with replacement. For each drawn sample, the Pearson and Spearman correlation matrices were calculated. Next, for each element of the correlation matrices, we calculated the absolute of the mean and the standard deviation across the 50,000 samples. To assess how accurately the sample correlation coefficients corresponded to the population values, we calculated the mean absolute difference of each r_p and r_s with respect to the population values (R_p and R_s). R_p and R_s were defined as the correlation coefficients for the full population (N = 11,878 for the ASVAB, N = 1,895,753 for the BFI, and N = 9,077 for the DBQ).

Results of the Five Sampling Studies

A numerical comparison between the performance of r_p and r_s is provided in Table 3. It can be seen that for the ASVAB data, r_p gives the same average values as r_s, with about 6% lower variability (i.e., lower *SD*). For the BFI and DBQ data, the opposite results were found: the mean absolute difference between r_s and R_s is smaller than the mean absolute difference between r_p and R_p. In other words, Spearman correlation coefficients are closer to their population value than are Pearson correlation coefficients. Furthermore, for the DBQ data in particular, the mean absolute difference between r_s and R_p is smaller than the mean absolute difference between r_p and R_p. That is, r_s even outperformed r_p in recovering r_p's own population value.

Table 3 further shows that the superior performance of r_s is evident for the DBQ dataset (featuring kurtosis > 3 for all items) and is less evident for the BFI dataset (featuring average kurtosis < 3). r_p on average has 2% higher variability (i.e., higher *SD*) than r_s for the BFI items, 4% higher variability for the BFI scales, 18% higher variability for the DBQ items, and 24% higher variability for the DBQ scales.

The mean absolute difference of r_p (and to a lesser extent of r_s) with respect to the population value is particularly large for pairs of DBQ items that have distributions with high kurtosis (see Figure S7 of the supplementary material). The distributions of r_p and r_s for the two DBQ items having the highest kurtosis (60.9 and 57.2, respectively) are illustrated in Figure 6. It can be seen that for this selected pair of variables, r_p was considerably more variable than r_s, with the standard deviation at N = 1,000 being .071 for r_p and .049 for r_s. Figure 7 illustrates the variability of r_p and r_s as a function of R_p for each of the five sampling studies. It can be seen

Table 3
Means and Standard Deviations of Sample Correlation Coefficients, and Mean Absolute Difference Between Sample Correlation Coefficients and Population Correlation Coefficients (N = 200)

Measure	ASVAB Mean across 45 correlations	BFI items Mean across 946 correlations	BFI scales Mean across 10 correlations	DBQ items Mean across 561 correlations	DBQ scales Mean across 10 correlations
$\lvert\text{Mean } r_p\rvert$	.6269	.1205	.1772	.1694	.4178
$\lvert\text{Mean } r_s\rvert$	.6258	.1221	.1683	.1616	.4144
$\lvert\text{Mean } r_p\rvert - \lvert R_p\rvert$	−.0005	−.0001	−.0005	−.0019	−.0019
$\lvert\text{Mean } r_s\rvert - \lvert R_s\rvert$	−.0022	−.0002	−.0008	−.0006	−.0013
$SD\ r_p$	.0411	.0732	.0741	.0872	.0750
$SD\ r_s$	.0436	.0715	.0714	.0742	.0605
Mean $\lvert r_p - R_p\rvert$	.0327	.0585	.0592	.0697	.0596
Mean $\lvert r_p - R_s\rvert$	.0352	.0587	.0602	.0701	.0628
Mean $\lvert r_s - R_p\rvert$	.0371	.0574	.0579	.0606	.0522
Mean $\lvert r_s - R_s\rvert$	.0347	.0571	.0570	.0593	.0483

Note. R_p = population Pearson correlation coefficient; R_s = population Spearman correlation coefficient; ASVAB = Armed Services Vocational Aptitude Battery; BFI = Big Five Inventory; DBQ = Driver Behaviour Questionnaire. Skewness was defined as the third central moment divided by the cube of the standard deviation. Kurtosis was defined as the fourth central moment divided by the fourth power of the standard deviation. Kurtosis of a normal distribution = 3. R_p and R_s were defined as the correlation coefficients for the total sample (i.e., N = 11,878 for the ASVAB, N = 1,895,753 for the BFI, and N = 9,077 for the DBQ).

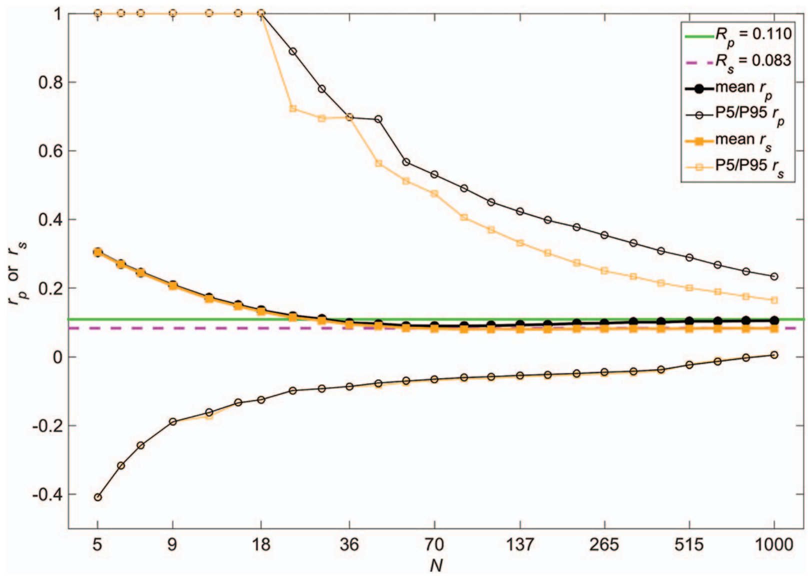

Figure 6. Sampling results for the two variables of the Driver Behaviour Questionnaire (DBQ) having the highest kurtosis of the 34 items (population kurtosis = 60.9 and 57.2, respectively; population skewness = 6.42 and 6.05, respectively). The figure shows the mean, 5th percentile (P5), and 95th percentile (P95) of the Pearson correlation coefficient (r_p) and the Spearman correlation coefficient (r_s) as a function of sample size (*N*). The population coefficients R_p and R_s were defined as the correlation coefficients for the total sample (N = 9,077). The results were based on 50,000 samples. Note that 8,272 of 9,077 respondents answered "never" to both items, and hence the correlation coefficient could often not be calculated when the sample size was small. The sampling was repeated when the correlation coefficient could not be calculated. See the online article for the color version of this figure.

that r_s is considerably less variable than r_p, especially for the BFI scales, DBQ items, and DBQ scales.

Additional Simulations With $N = 25$ and $N = 1{,}000$

The results in Table 3 and Figure 7 were based on a sample size of 200. To test whether the results depend on sample size, the simulations were repeated for $N = 25$ and $N = 1{,}000$ (see online supplemental materials Tables S1 and S2). For $N = 25$, the variabilities of r_p and r_s are obviously higher than for $N = 200$, but the pattern of differences between r_p and r_s is the same. For $N = 1{,}000$, the variabilities of r_p and r_s are considerably lower than for $N = 200$, but again the pattern of differences is the same, with r_s having a lower standard deviation than r_p for the BFI and DBQ datasets. For $N = 1{,}000$ it is less likely that the mean absolute difference between r_s and R_p is smaller than the mean absolute difference between r_p and R_p, because at such high sample size, the correlation coefficients r_p and r_s are close to their own respective population values.

Discussion

The Pearson product–moment correlation coefficient (r_p) and the Spearman rank correlation coefficient (r_s) are widely used in psychology, with r_p being the most popular. The two coefficients have different goals: r_p is a measure of the degree of linearity between two vectors of data, whereas r_s measures their degree of monotonicity.

The characteristics of r_p and r_s have been widely studied for over 100 years, and in the case of bivariate normality, the distribution of r_p is known exactly (Equation 4). The influence functions of the Pearson and Spearman correlations have been described exactly as well (e.g., Croux & Dehon, 2010). However, several of these features of r_p and r_s may not be known among substantive researchers, and hence our simulations of normally distributed variables are presented as a helpful tutorial. In other words, we illustrated in an intuitive graphical manner the variability, bias, and robustness properties of both coefficients, with a focus on the effect sizes and sample sizes that are likely to occur in psychological research. The relative performance of r_p and r_s in real psychological datasets for different item characteristics, sample sizes, and aggregation methods (i.e., item and scale levels) is intended to facilitate informed decision making regarding when to select r_p and when to select r_s.

Our computer simulations showed that for normally distributed variables r_s behaves approximately the same as r_p, with r_s being slightly lower and more variable than r_p. The difference between the standard deviation of r_p and r_s was minor (< 1%) when the association was weak or moderate in the population ($R_p = 0$ and $R_p = .2$). However, r_s had a substantially higher standard deviation than r_p when the correlation was strong (i.e., a 3% to 4% higher standard deviation when $R_p = .4$) or very strong (i.e., 18% higher standard deviation when $R_p = .8$).

In psychological research, near-normally distributed data, such as the ASVAB test scores, do occur. We showed that for the

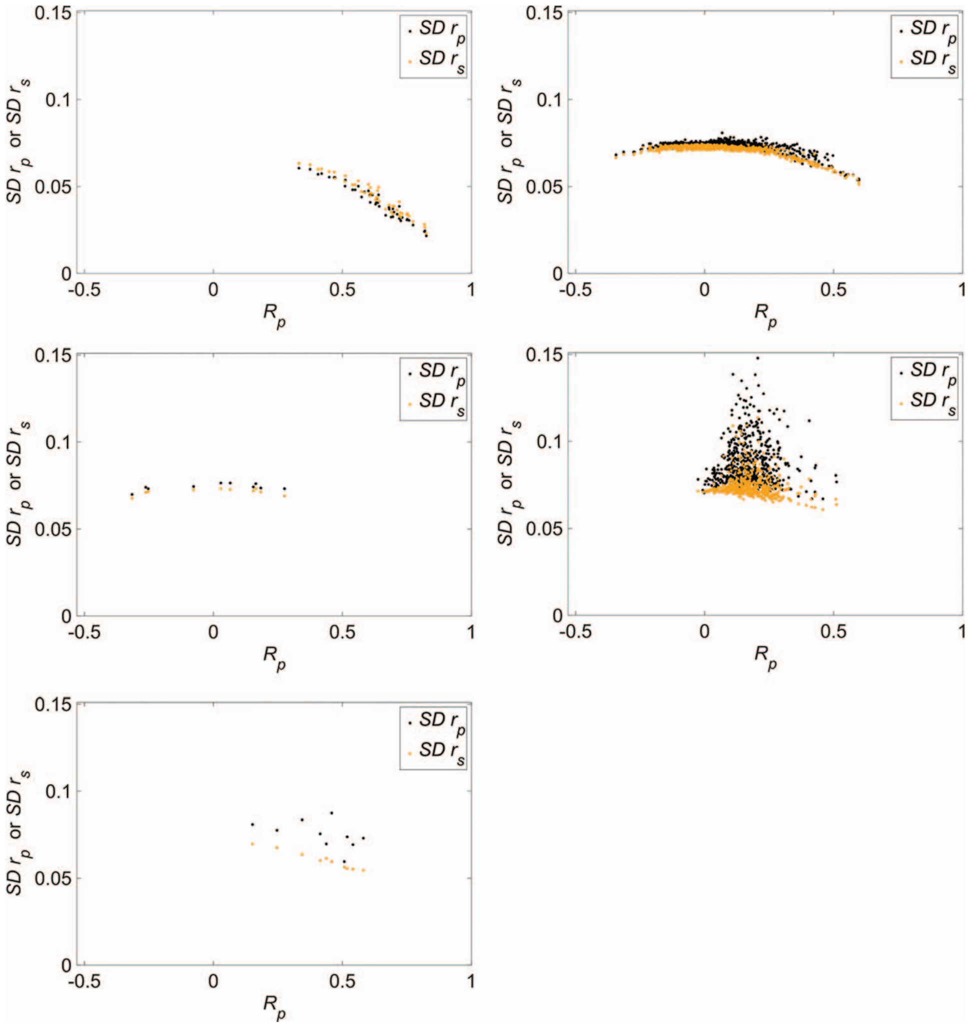

Figure 7. Standard deviation (*SD*) of the Pearson correlation coefficient (r_p) and the standard deviation of the Spearman correlation coefficient (r_s; N = 200) as a function of the population Pearson correlation coefficient (R_p). The population coefficient R_p was defined as the correlation coefficients for the total sample (N = 11,878 for the ASVAB, N = 1,895,753 for the BFI, and N = 9,077 for the DBQ). Top left: Armed Services Vocational Aptitude Battery (ASVAB; 45 correlation coefficients). Top right: Big Five Inventory (BFI) items (946 correlation coefficients). Middle left: BFI scales (10 correlation coefficients). Middle right: Driver Behaviour Questionnaire (DBQ) items (561 correlation coefficients). Bottom left: DBQ scales (10 correlation coefficients). See the online article for the color version of this figure.

strongly intercorrelated and approximately normally distributed variables of the ASVAB, r_p slightly outperformed r_s in terms of variability. The expected values of r_p and r_s were almost the same, but the standard deviation of r_p was about 6% lower than the standard deviation of r_s. The similarity of r_s and r_p for normally distributed psychometric variables is consistent with empirical sampling research in the physical sciences, where normally distributed variables tend to be common (McDonald & Green, 1960). However, in psychological research, heavy-tailed distributions are common (Blanca et al., 2013; Micceri, 1989). Using a simulation of two correlated variables with heavy-tailed distributions, we showed that r_s was between 13 and 27% less variable than was r_p.

The comparative efficacy of r_p versus r_s was further explored in a sampling study of BFI and DBQ survey data at both the item and scale levels. For these survey datasets, r_s turned out to be between 2% and 24% less variable than r_p. In fact, for the DBQ dataset, we found that the sample Spearman correlation coefficient (r_s) was a more accurate approximation of the population Pearson correlation coefficient (R_p) than was the sample Pearson correlation coefficient (r_p). This inaccuracy of r_p with respect to R_p was particularly large when the two variables had heavy-tailed distributions (see Figure S7 of supplementary material).

Our simulations further made clear that r_s is robust, while r_p is sensitive to an outlier, even for a sample size as high as 200. Outliers may be caused by a recording error, an error in the experimental procedure, or an accurate representation of a rare case (Cohen et al., 2013). It is likely that real-life data are contaminated with "faulty data" (Spearman, 1910) or an "accidental error" (Spearman, 1904, p. 81), and therefore the robustness of the Spearman estimator (r_s) is a virtue for empirical researchers. Using Anscombe's (1960) insurance policy analogy, r_s yields a slight loss of efficiency when bivariate normality assumptions are met, but this seems a small premium given the impressive protection it provides against outliers (see Figure 5).

Our study also illustrated the dramatic effect of sample size on the variability of the correlation coefficients. A sample size of 25 yields average errors that are often even larger than the absolute magnitude of the correlation coefficient (e.g., Figure 2; Table S1), which essentially means that the observed correlations are almost meaningless. The standard deviations of r_s and r_p decrease approximately according to the square root of sample size, which means that the standard deviations reduce by approximately 41% when sample size is doubled (cf. Figure S6). In other words, although substantial efficiency gains can be achieved by choosing r_s instead of r_p, the effect of sample size is much more dramatic, and therefore we urge researchers to always monitor the confidence interval of their obtained effects.

So, should a practitioner use r_p or r_s? Of course, the two correlation coefficients have different goals: r_p represents the strength of the linear relationship between two vectors of data, whereas r_s describes their degree of monotonicity. Because r_p and r_s have different goals, they strictly ought not to be seen as competing approaches. That is, if one's aim is solely to assess whether the individual sample data points are linearly related (regardless of any nonlinearity that exists), and one's sample size is very large, then r_p should be used. However, it is likely that practitioners are interested in obtaining a high quality correlational measure in terms of low variability, low bias, and high robustness. In such case, r_s clearly has attractive properties compared to r_p. If one expects that the two variables have low kurtosis (i.e., normal or platykurtic distributions) and outliers are unlikely to be present, r_p is to be recommended. In other circumstances, r_s seems to be the preferred method because of its superior performance in terms of variability and robustness. The 'embarrassing' failure of r_p to accurately estimate its own population value (R_p) in the DBQ dataset, both at the item and at the scale levels, strongly argues in favor of using r_s for heavy-tailed survey data. Note that the behavior of r_p and r_s depends not just on kurtosis, but also on sample size, the population correlation coefficient, and the type of nonlinear relationship between the two variables (see supplementary material). These factors may explain some of the idiosyncratic behaviors of the datasets (see Table 3). Ambiguity arises when having to analyze a large set of variables, whereby half of the data are platykurtic and the other half leptokurtic. In this case, again using Anscombe's (1960) insurance parallel, we recommend using r_s instead of r_p, because the premium-protection trade-off is not symmetric. After all, there is a relatively small increase of variability for the variables that are indeed platykurtic, while r_s offers marked robustness to heavy tails and outliers.

There are, of course, a large number of other types of data transformations, such as a logarithmic, multiplicative inverse, or power transformation, that can be successfully applied prior to calculating the Pearson correlation coefficient (Bishara & Hittner, 2012). However, whereas the rank transformation as used in the Spearman correlation coefficient is broadly applicable, other types of data transformation are not. For example, a logarithmic or square root transformation is impossible on negative numbers (unless applying an arbitrary offset), and the multiplicative inverse transformation dilutes any meaningful association when some of the numbers are close to zero. In other words, it is quite possible to mess up one's data by choosing the 'wrong' type of transformation, so that, for example, a normal distribution becomes highly non-normal. As a result, selecting an appropriate nonlinear data transformation requires either prior knowledge of the population distribution or the ethically dubious practice of 'peeking' at the data (Sagarin, Ambler, & Lee, 2014), and it is therefore difficult to come up with systematic meaningful guidelines. In contrast, the Spearman correlation appears to be applicable across a broad array of normal and non-normal distributions.

Alternative measures of association, such as the percentage bend correlation (Wilcox, 1994), the Winsorized correlation (Wilcox, 1993), and the Kendall's tau rank correlation coefficient (r_t), may be even more robust and efficient than r_s (see Croux & Dehon, 2010). r_t is attractive because it can be interpreted intuitively as the proportion of pairs of observations that are in the same order on both variables minus the proportion that are opposite (Cliff, 1996; Noether, 1981). Other attractive properties of r_t are that it is an unbiased estimator of its population value and that the variance is given in closed form (Esscher, 1924; Fligner & Rust, 1983; Hollander, Wolfe, & Chicken, 2013; Kendall, 1948; Kendall, Kendall, & Babington Smith, 1939; Xu, Hou, Hung, & Zou, 2013). However, Xu et al. (2013) argued that r_s has a lower computational load than r_t, and that the variance of r_s can be approximated with high numerical accuracy, leading the authors to conclude that the mathematical advantage of r_t over r_s is not of great importance. Another issue is that r_t converges to markedly different population values than r_p and r_s. For typical bivariate normal distributions, r_p and r_s are about 50% greater than r_t (Equations 9 and 10, Fredricks & Nelsen, 2007, see also Figure S16). Because present-day researchers are familiar with interpreting r_p (see Table 1), it seems unlikely that r_t could replace r_p. r_s on the other hand has the potential to be used in place of r_p, because, as we showed, r_s can surpass r_p in estimating R_p. Corrected correlations, such as polychoric correlations, may also be useful alternatives to the Spearman correlation, especially for multivariate applications. Although multivariate methods using the polychoric correlation matrix have been implemented in almost all SEM packages, and are still under scrutiny (e.g., Rhemtulla, Brosseau-Liard, & Savalei, 2012; Yuan, Wu, & Bentler, 2011), the polychoric correlation has not yet caught on among substantive researchers (see Table 1).

There are established ways of dealing with outliers, including outlier removal and robust approaches such as least absolute deviation, least trimmed squares, M-estimates, and bounded inference estimators (Cohen et al., 2013; Rousseeuw & Leroy, 2005), or procedures that take into account the structure of the data (Wilcox & Keselman, 2012, see Pernet, Wilcox, & Rousselet, 2012 for an open source MATLAB toolbox). However, removing outliers is an inherently subjective procedure, and retaining too much flexibility could easily lead to inflated effect sizes and false positive inferences (Bakker & Wicherts, 2014; Cohen et al., 2013). It is noted that high kurtosis and outliers can be indicative of problems in the measurement procedure. Subtle changes in questionnaire wording or anchoring can have large effects on the obtained results (Schwarz, 1999). We recommend that researchers remedy the root causes of outliers and high kurtosis before they continue their study.

The choice of correlation coefficient is important not only for establishing bivariate relationships. Psychologists often intend to do follow-up analyses, such as to calculate a percentage of variance explained, to perform an ANOVA or MANOVA, to carry out a meta-analysis of correlation coefficients, or to establish a matrix of correlation coefficients to be submitted to a multivariate statistical method such as principal component analysis, factor analysis, or structural equation modeling. Cliff (1996) argued that perhaps most of the answers that psychologists want to get from their data are ordinal ones, and the data they work with have, at best, ordinal justification. He concluded that ordinal questions should be answered ordinally, instead of trying to answer them with Pearson correlations, mean differences, and parametric techniques. Using ordinal statistics has the added benefit that the inferences remain unchanged if the variables are monotonically transformed (Cliff, 1996). Unfortunately, purely ordinal multivariate statistical methods are rare and generally less developed than traditional parametric methods (for a possible exception using Kendall's tau, see Cliff, 1996).

Indeed, there has been considerable controversy about the use of a rank transformation, because corresponding statistical procedures in complex research designs are sometimes unavailable, inexact, and difficult to interpret (e.g., Fligner, 1981; Sawilowsky, 1990; Zimmerman, 2012). In some cases, the rank transformation may be even entirely inappropriate. For example, when testing the null hypothesis of no interactions in a multifactorial layout, the rank transformation can yield a test statistic that goes to infinity as the sample size increases (Thompson, 1991; see also Akritas, 1993; Sawilowsky, Blair, & Higgins, 1989). Hence, our present results, which favor r_s over r_p, seem to lead to a "cul de sac" for researchers in psychology.

However, one could set aside such theoretical constraints, and adopt "a pragmatic sanction" (Stevens, 1951, p. 26). We argue that there is no good reason to stick to r_p for the mere reason that it is consistent with follow-up analyses such as ANOVA and principal component analysis. It is easily forgotten that the assumption of normality is almost always violated in the population, and that calculating r_p on ordinal data, such as those obtained from Likert items, is not strictly permissible anyway (Stevens, 1946). The debate of representational versus pragmatic measurement is a long and bitter one with deep philosophical roots (e.g., Hand, 2004; Michell, 2008; Velleman & Wilkinson, 1993). We support Lord's (1953) pragmatic view that "the numbers don't remember where they came from" (p. 21), and we argue that if r_s outperforms r_p in terms of bias, variability, and robustness, then there is no justifiable reason for not using r_s. We illustrate this point by submitting an r_s correlation matrix and an r_p correlation matrix of the DBQ data to a principal component analysis (and see Babakus, Ferguson, & Jöreskog, 1987 and Mittag, 1993, for a similar approach). Results showed that the first six eigenvalues of the r_p correlation matrix were between 26% and 68% more variable than the eigenvalues of the r_s correlation matrix (see online supplemental materials Table S3), which means that the factor structure is more stable if researchers simply base their multivariate analyses on the r_s matrix. In some software packages, it is relatively easy to submit the r_s matrix to a multivariate analysis (e.g., in MATLAB *factoran(corr(X,'type','spearman'),2,'xtype','covariance')* performs a maximum-likelihood factor analysis on the X matrix, extracting two factors). However, in SPSS, for example, this analysis requires extensive scripting (Garcia-Granero, 2002). Therefore, we recommend the simpler approach of transforming all variables to ranks prior to running the multivariate analysis (e.g., *factoran(tiedrank(X),2)* in the MATLAB command window or *Transform* > *Rank Cases* from SPSS's pull-down menu). Summarizing, a rank-transformation is an appropriate bridge between nonparametric and parametric statistics (Conover & Iman, 1981).

References

Akritas, M. G. (1993). Limitations of the rank transform procedure: A study of repeated measures design, Part II. *Statistics & Probability Letters, 17,* 149–156. http://dx.doi.org/10.1016/0167-7152(93)90009-8

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics, 2,* 123–146. http://dx.doi.org/10.1080/00401706.1960.10489888

Babakus, E., Ferguson, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research, 24,* 222–228. http://dx.doi.org/10.2307/3151512

Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods, 19,* 409–427. http://dx.doi.org/10.1037/met0000014

Best, D. J., & Roberts, D. E. (1975). Algorithm AS 89: The upper tail probabilities of Spearman's rho. *Applied Statistics, 24,* 377–379. http://dx.doi.org/10.2307/2347111

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods, 17,* 399–417. http://dx.doi.org/10.1037/a0028087

Bishara, A. J., & Hittner, J. B. (2015). Reducing bias and error in the correlation coefficient due to nonnormality. *Educational and Psychological Measurement, 75,* 785–804. http://dx.doi.org/10.1177/0013164414557639

Blair, R. C., & Lawson, S. B. (1982). Another look at the robustness of the product-moment correlation coefficient to population non-normality. *Florida Journal of Educational Research, 24,* 11–15.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9,* 78–84. http://dx.doi.org/10.1027/1614-2241/a000057

Bleidorn, W., Klimstra, T. A., Denissen, J. J., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2013). Personality maturation around the world: A cross-cultural examination of social-investment theory. *Psychological Science, 24,* 2530–2540. http://dx.doi.org/10.1177/0956797613498396

Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika, 65,* 23–28. http://dx.doi.org/10.1007/BF02294183

Borkowf, C. B. (2002). Computing the nonnull asymptotic variance and the asymptotic relative efficiency of Spearman's rank correlation. *Computational Statistics & Data Analysis, 39,* 271–286. http://dx.doi.org/10.1016/S0167-9473(01)00081-0

Bureau of Labor Statistics. (2002). *U.S. Department of Labor. National Longitudinal Survey of Youth 1979 cohort, 1979–2002* [Computer file]. Produced and distributed by the Center for Human Resource Research, The Ohio State University. Columbus, OH. Retrieved from https://www.nlsinfo.org/investigator/pages/search.jsp?s=NLS79

Burt, C. (1957). The distribution of intelligence. *British Journal of Psychology, 48,* 161–175. http://dx.doi.org/10.1111/j.2044-8295.1957.tb00614.x

Chok, N. S. (2010). *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data* (Doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics. *Multivariate Behavioral Research, 31,* 331–350. http://dx.doi.org/10.1207/s15327906mbr3103_4

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124–129.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684. http://dx.doi.org/10.1037/h0043943

Croux, C., & Dehon, C. (2010). Influence functions of the Spearman and Kendall correlation measures. *Statistical Methods and Applications, 19,* 497–515. http://dx.doi.org/10.1007/s10260-010-0142-z

David, F. N., & Mallows, C. L. (1961). The variance of Spearman's rho in normal samples. *Biometrika, 48,* 19–28. http://dx.doi.org/10.1093/biomet/48.1-2.19

David, S. T., Kendall, M. G., & Stuart, A. (1951). Some questions of distribution in the theory of rank correlation. *Biometrika, 38,* 131–140. http://dx.doi.org/10.1093/biomet/38.1-2.131

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35,* 13–21. http://dx.doi.org/10.1016/j.intell.2006.02.001

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods, 2,* 292–307. http://dx.doi.org/10.1037/1082-989X.2.3.292

Delucchi, K. L., & Bostrom, A. (2004). Methods for analysis of skewed data distributions in psychiatric clinical studies: Working with many zero values. *The American Journal of Psychiatry, 161,* 1159–1168. http://dx.doi.org/10.1176/appi.ajp.161.7.1159

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

Esscher, F. (1924). On a method of determining correlation from the ranks of the variates. *Skandinavisk Aktuarietidskrift, 7,* 201–219.

Fieller, E. C., Hartley, H. O., & Pearson, E. S. (1957). Tests for rank correlation coefficients. I. *Biometrika, 44,* 470–481. http://dx.doi.org/10.1093/biomet/44.3-4.470

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.

Fligner, M. A. (1981). Comment. *The American Statistician, 35,* 131–132.

Fligner, M. A., & Rust, S. V. (1983). On the independence problem and Kendall's tau. *Communications in Statistics Theory and Methods, 12,* 1597–1607. http://dx.doi.org/10.1080/03610928308828554

Fouladi, R. T., & Steiger, J. H. (2008). The Fisher transform of the Pearson product moment correlation coefficient and its square: Cumulants, moments, and applications. *Communications in Statistics Simulation and Computation, 37,* 928–944. http://dx.doi.org/10.1080/03610910801943735

Fowler, R. L. (1987). Power and robustness in product-moment correlation. *Applied Psychological Measurement, 11,* 419–428. http://dx.doi.org/10.1177/014662168701100407

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS ONE, 9,* e109019. http://dx.doi.org/10.1371/journal.pone.0109019

Fredricks, G. A., & Nelsen, R. B. (2007). On the relationship between Spearman's rho and Kendall's tau for pairs of continuous random variables. *Journal of Statistical Planning and Inference, 137,* 2143–2150. http://dx.doi.org/10.1016/j.jspi.2006.06.045

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science, 15,* 373–378. http://dx.doi.org/10.1111/j.0956-7976.2004.00687.x

Garcia-Granero, M. (2002). *How to perform factor analysis with Spearman correlation thru a matrix*. Retrieved from http://spsstools.net/en/plaintext/FAwithSpearmanCorrelation.txt/

Ghosh, B. K. (1966). Asymptotic expansions for the moments of the distribution of correlation coefficient. *Biometrika, 53,* 258–262. http://dx.doi.org/10.2307/2334076

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59,* 93–104. http://dx.doi.org/10.1037/0003-066X.59.2.93

Guérin, R., De Oliveira, J. C., & Weber, S. (2013). Adoption of bundled services with network externalities and correlated affinities. *ACM Transactions on Internet Technology, 14,* 13.

Hand, D. J. (2004). *Measurement: Theory and practice*. London, UK: Arnold.

Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics & Data Analysis, 40,* 687–711. http://dx.doi.org/10.1016/S0167-9473(02)00072-5

Hjelm, H. F., & Norris, R. C. (1962). Empirical study of the efficacy of Fisher's z-transformation. *Journal of Experimental Education, 30,* 269–277. http://dx.doi.org/10.1080/00220973.1962.11010715

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics, 19,* 293–325. http://dx.doi.org/10.1214/aoms/1177730196

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods* (3rd ed.). Hoboken, NJ: Wiley.

Hotelling, H. (1951). The impact of R. A. Fisher on statistics. *Journal of the American Statistical Association, 46,* 35–46. http://dx.doi.org/10.1080/01621459.1951.10500765

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society Series B. Methodological, 15,* 193–232.

Hotelling, H., & Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *Annals of Mathematical Statistics, 7,* 29–43. http://dx.doi.org/10.1214/aoms/1177732543

Jensen, A. R. (2006). *Clocking the mind: Mental chronometry and individual differences*. Amsterdam, the Netherlands: Elsevier.

Keats, J. A., & Lord, F. M. (1962). A theoretical distribution for mental test scores. *Psychometrika, 27,* 59–72. http://dx.doi.org/10.1007/BF02289665

Kendall, M. G. (1948). *Rank correlation methods*. Oxford, UK: Griffin.

Kendall, M. G., Kendall, S. F. H., & Babington Smith, 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.

Kenney, J. F., & Keeping, E. S. (1951). *Mathematics of statistics: Part two*. Toronto, Canada: Van Nostrand.

Kowalski, C. J. (1972). On the effects of non-normality on the distribution of the sample product-moment correlation coefficient. *Journal of the Royal Statistical Society Series C, Applied Statistics, 21,* 1–12.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association, 53,* 814–861. http://dx.doi.org/10.1080/01621459.1958.10501481

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751. http://dx.doi.org/10.1037/h0063675

Lovie, A. D. (1995). Who discovered Spearman's rank correlation? *British Journal of Mathematical and Statistical Psychology, 48,* 255–269. http://dx.doi.org/10.1111/j.2044-8317.1995.tb01063.x

Lyon, A. (2014). Why are normal distributions normal? *The British Journal for the Philosophy of Science, 65,* 621–649. http://dx.doi.org/10.1093/bjps/axs046

Maier, M. H., & Sims, W. H. (1986). *The ASVAB score scales: 1980 and World War II*. Alexandria, VA: Center for Naval Analyses.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112,* 331–348. http://dx.doi.org/10.2466/03.11.PMS.112.2.331-348

Mattsson, M. (2012). Investigating the factorial invariance of the 28-item DBQ across genders and age groups: An exploratory structural equation modeling study. *Accident Analysis and Prevention, 48,* 379–396. http://dx.doi.org/10.1016/j.aap.2012.02.009

McDonald, J. E., & Green, C. R. (1960). A comparison of rank-difference and product-moment correlation of precipitation data. *Journal of Geophysical Research, 65,* 333–336. http://dx.doi.org/10.1029/JZ065i001p00333

McGrath, J., Saha, S., Welham, J., El Saadi, O., MacCauley, C., & Chant, D. (2004). A systematic review of the incidence of schizophrenia: The distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC Medicine, 2,* 13. http://dx.doi.org/10.1186/1741-7015-2-13

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., . . . Reed, G. M. (2001). Psychological testing and psychological assessment. A review of evidence and issues. *American Psychologist, 56,* 128–165.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166. http://dx.doi.org/10.1037/0033-2909.105.1.156

Michell, J. (2008). Is psychometrics pathological science? *Measurement, 6,* 7–24.

Mittag, K. C. (1993, January). *Scale-free nonparametric factor analysis: A user-friendly introduction with concrete heuristic examples*. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Moran, P. A. P. (1948). Rank correlation and product-moment correlation. *Biometrika, 35,* 203–206. http://dx.doi.org/10.1093/biomet/35.1-2.203

Morgan, W. R. (1983). Learning and student life quality of public and private school youth. *Sociology of Education, 56,* 187–202. http://dx.doi.org/10.2307/2112548

Noether, G. E. (1981). Why Kendall Tau? *Teaching Statistics, 3,* 41–43. http://dx.doi.org/10.1111/j.1467-9639.1981.tb00422.x

Obschonka, M., Schmitt-Rodermund, E., Silbereisen, R. K., Gosling, S. D., & Potter, J. (2013). The regional distribution and correlates of an entrepreneurship-prone personality profile in the United States, Germany, and the United Kingdom: A socioecological perspective. *Journal of Personality and Social Psychology, 105,* 104–122. http://dx.doi.org/10.1037/a0032275

Pearson, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions A, 373,* 253–318.

Pearson, K. (1907). *On further methods of determining correlation. Drapers' Company Research Memoirs Biometric Series IV*. Retrieved from https://archive.org/details/onfurthermethod00peargoog

Pernet, C. R., Wilcox, R., & Rousselet, G. A. (2012). Robust correlation analyses: False positive and power validation using a new open source Matlab toolbox. *Frontiers in Psychology, 3,* 606.

Plomin, R., & Deary, I. J. (2015). Genetics and intelligence differences: Five special findings. *Molecular Psychiatry, 20,* 98–108. http://dx.doi.org/10.1038/mp.2014.105

Rentfrow, P. J., Gosling, S. D., Jokela, M., Stillwell, D. J., Kosinski, M., & Potter, J. (2013). Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology, 105,* 996–1012.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods, 17,* 354–373. http://dx.doi.org/10.1037/a0029315

Richard, F. D., Bond, C. F., Jr., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331

Rousseeuw, P. J., & Leroy, A. M. (2005). *Robust regression and outlier detection*. New York, NY: Wiley.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science, 9,* 293–304. http://dx.doi.org/10.1177/1745691614528214

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research, 60,* 91–126. http://dx.doi.org/10.3102/00346543060001091

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the Type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational and Behavioral Statistics, 14,* 255–267. http://dx.doi.org/10.3102/10769986014003255

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

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47,* 609–612. http://dx.doi.org/10.1016/j.jrp.2013.05.009

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105. http://dx.doi.org/10.1037/0003-066X.54.2.93

Shieh, G. (2010). Estimation of the simple correlation coefficient. *Behavior Research Methods, 42,* 906–917. http://dx.doi.org/10.3758/BRM.42.4.906

Sloane, N. J. (2003). *The on-line encyclopedia of integer sequences*. Retrieved from https://oeis.org/A126972

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15,* 72–101. http://dx.doi.org/10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3,* 271–295.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84,* 1041–1053.

Steiger, J. H. (2001). Driving fast in reverse: The relationship between software development, theory, and education in structural equation modeling. *Journal of the American Statistical Association, 96,* 331–338. http://dx.doi.org/10.1198/016214501750332893

Steiger, J. H. (2004). Paul Meehl and the evolution of statistical methods in psychology. *Applied & Preventive Psychology, 11,* 69–72. http://dx.doi.org/10.1016/j.appsy.2004.02.012

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, 103,* 677–680. http://dx.doi.org/10.1126/science.103.2684.677

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York, NY: Wiley.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44,* 703–742. http://dx.doi.org/10.1111/j.1744-6570.1991.tb00696.x

Thompson, G. L. (1991). A note on the rank transform for interactions. *Biometrika, 78,* 697–701. http://dx.doi.org/10.1093/biomet/78.3.697

Tracy, J. L., Robins, R. W., & Sherman, J. W. (2009). The practice of psychological science: Searching for Cronbach's two streams in social-personality psychology. *Journal of Personality and Social Psychology, 96,* 1206–1225. http://dx.doi.org/10.1037/a0015173

Transport Research Laboratory. (2008). *Transport Research Laboratory, Safety, Security and Investigations Division [TRL], 2008. Cohort II: A study of learner and novice drivers, 2001–2005* [Computer file].

Retrieved from http://discover.ukdataservice.ac.uk/catalogue/?sn=5985&type=Data%20catalogue

van den Oord, E. J. C. G., Pickles, A., & Waldman, I. D. (2003). Normal variation and abnormality: An empirical study of the liability distributions underlying depression and delinquency. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 44,* 180–192. http://dx.doi.org/10.1111/1469-7610.00112

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47,* 65–72.

Wells, P., Tong, S., Sexton, B., Grayson, G., & Jones, E. (2008). *Cohort II: A study of learner and new drivers. Vol. 1. Main report (Report no. 81)*. London, UK: Department for Transport.

Wilcox, R. R. (1993). Some results on a Winsorized correlation coefficient. *The British Journal of Mathematical and Statistical Psychology, 46,* 339–349. http://dx.doi.org/10.1111/j.2044-8317.1993.tb01020.x

Wilcox, R. R. (1994). The percentage bend correlation coefficient. *Psychometrika, 59,* 601–616. http://dx.doi.org/10.1007/BF02294395

Wilcox, R. R., & Keselman, H. J. (2012). Modern regression methods that can substantially increase power and provide a more accurate understanding of associations. *European Journal of Personality, 26,* 165–174. http://dx.doi.org/10.1002/per.860

Winterbottom, A. (1979). A note on the derivation of Fisher's transformation of the correlation coefficient. *The American Statistician, 33,* 142–143.

Xu, W., Hou, Y., Hung, Y. S., & Zou, Y. (2013). A comparative analysis of Spearman's rho and Kendall's tau in normal and contaminated normal models. *Signal Processing, 93,* 261–276. http://dx.doi.org/10.1016/j.sigpro.2012.08.005

Yuan, K.-H., & Bentler, P. M. (2000). Inferences on correlation coefficients in some classes of nonnormal distributions. *Journal of Multivariate Analysis, 72,* 230–248. http://dx.doi.org/10.1006/jmva.1999.1858

Yuan, K.-H., Bentler, P. M., & Zhang, W. (2005). The effect of skewness and kurtosis on mean and covariance structure analysis: The univariate case and its multivariate implication. *Sociological Methods & Research, 34,* 240–258. http://dx.doi.org/10.1177/0049124105280200

Yuan, K.-H., Wu, R., & Bentler, P. M. (2011). Ridge structural equation modelling with correlation matrices for ordinal and continuous data. *British Journal of Mathematical and Statistical Psychology, 64,* 107–133. http://dx.doi.org/10.1348/000711010X497442

Zayed, H., & Quade, D. (1997). On the resistance of rank correlation. *Journal of Statistical Computation and Simulation, 58,* 59–81. http://dx.doi.org/10.1080/00949659708811822

Zimmerman, D. W. (2012). A note on consistency of non-parametric rank tests and related rank transformations. *British Journal of Mathematical and Statistical Psychology, 65,* 122–144. http://dx.doi.org/10.1111/j.2044-8317.2011.02017.x

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicológica: Revista de Metodología y Psicología Experimental, 24,* 133–158.

Received May 29, 2014
Revision received December 30, 2015
Accepted January 26, 2016 ■