Received: 11 January 2024 | Accepted: 14 May 2024

DOI: 10.1111/bmsp.12354

the british psychological society
promoting excellence in psychology

ARTICLE

Are alternative variables in a set differently associated with a target variable? Statistical tests and practical advice for dealing with dependent correlations

Miguel A. García-Pérez

Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Madrid, Spain

Correspondence
Miguel A. García-Pérez, Departamento de Metodología, Facultad de Psicología, Universidad Complutense, Campus de Somosaguas, 28223 Madrid, Spain.
Email: miguel@psi.ucm.es

Funding information
Ministerio de Ciencia e Innovación, Grant/Award Number: PID2019-11083GB-I00

Abstract

The analysis of multiple bivariate correlations is often carried out by conducting simple tests to check whether each of them is significantly different from zero. In addition, pairwise differences are often judged by eye or by comparing the *p*-values of the individual tests of significance despite the existence of statistical tests for differences between correlations. This paper uses simulation methods to assess the accuracy (empirical Type I error rate), power, and robustness of 10 tests designed to check the significance of the difference between two dependent correlations with overlapping variables (i.e., the correlation between X_1 and Y and the correlation between X_2 and Y). Five of the tests turned out to be inadvisable because their empirical Type I error rates under normality differ greatly from the nominal alpha level of .05 either across the board or within certain sub-ranges of the parameter space. The remaining five tests were acceptable and their merits were similar in terms of all comparison criteria, although none of them was robust across all forms of non-normality explored in the study. Practical recommendations are given for the choice of a statistical test to compare dependent correlations with overlapping variables.

KEYWORDS

dependent correlations, Monte Carlo simulation, power, robustness, statistical tests, Type I error

1 | INTRODUCTION

The statistical analysis of correlations among variables has not changed much over decades of psychological research, nor has it been essentially altered by an increasing concern about matching statistical methods to the intended inferences. The most common approach to dealing with large numbers of correlations (e.g., intercorrelations among all pairs of variables in a set or correlations between each variable in a set and each variable in another set) consists of reporting all of them accompanied by a description of which of them, considered in isolation, are significantly different from zero. In addition, inferences about differences in correlation are usually made by comparing the *p*-values of the respective isolated tests of significance (for a relatively old but representative example, see Thorson & Powell, 1993). The inadequacy of this strategy for comparing correlations is easy to understand on considering that the boundary for a test that a correlation is significantly different from zero is sharp. For instance, with sample sizes of 120 observations and a two-sided test with $\alpha = .05$, a correlation of .180 in one group will be significantly different from zero whereas a correlation of .179 in another group will not. While these separate claims about each correlation are unquestionable, they cannot sustain a claim that the strength of correlation between the two variables of concern is different in each group, a claim that seemingly calls for a direct test that the two population correlations differ from one another. Similar considerations hold when separate tests reject the null that the correlation is zero or when they do not reject it, whether with similar or with different accompanying *p*-values.

Empirical scenarios requiring tests of equality of two correlations may have a number of distinct characteristics. The case mentioned in the previous paragraph involves equality of correlation between the same two variables in different populations, a case usually termed the statistical comparison of independent correlations because each correlation is computed using data from a different sample of bivariate observations. A variant thereof involves equality of the correlation between two variables and the correlation between two other variables, all four of them measured in a sample from the same population. This scenario pertains to what is generally termed the statistical comparison of two dependent correlations with non-overlapping variables, where the term captures the fact that all observations come from the same multivariate sample but none of the observations is used twice (e.g., for the computation of more than one of the correlations of concern). A somewhat different scenario also involving a single multivariate sample is referred to as the comparison of dependent correlations with overlapping variables, because now some observations are used more than once for the computation of all relevant correlations. A prototypical example involves testing whether two variables, X_1 and X_2, are equally correlated with a third variable Y (see, for example, Blake & Palmisano, 2021; Guo et al., 2022). In such cases, observations in Y are used for computation of its correlation with X_1 and, once again, for computation of its correlation with X_2. Scenarios involving a comparison of independent or dependent correlations can both involve omnibus tests of equality of more than two correlations.

The study reported in this paper focuses on statistical tests for the comparison of two dependent correlations with overlapping variables. Cases in which the comparisons of concern involve more than two correlations will be discussed in Section 6. A large number of statistical tests have been proposed for this purpose, 10 of which are presented in Section 2. In principle, all of them are interchangeable although some technicalities of their computation may affect their performance in terms of accuracy (i.e., Type I error rates), power, robustness to violations of distributional assumptions, or even the feasibility of their computation given the values of the sample statistics on hand. Earlier research on the comparative performance of these tests has been fragmentary and somewhat limited in scope. Some studies have compared a subset of these tests according to their accuracy and power with data drawn from trivariate normal distributions, although none of the studies has considered a broad range of values for the population correlations and a broad range of sample sizes (see, for example, Boyer et al., 1983; Cohen, 1989; Dunn & Clark, 1971; Hittner et al., 2003; May & Hittner, 1997a; Neill & Dunn, 1975; Wilcox & Tian, 2008). In addition, some studies have reported fractional results on robustness under violation of trivariate normality, always using identical non-normal (univariate) distributions for X_1, X_2, and Y. For instance, Boyer et al. (1983) used lognormal distributions, Cohen (1989) used chi-square distributions with two and four

degrees of freedom, and May and Hittner (1997a); see also (Hittner et al., 2003; Wilcox & Tian, 2008) used uniform and exponential distributions. Results reported in the studies just mentioned have shown a relatively small deterioration in the accuracy and power of the tests under violation of normality, but the generalizability of this conclusion to the empirically plausible case of arbitrarily different non-normal univariate distributions for X_1, X_2, and Y is uncertain. A further aspect that has not previously been studied is the comparative applicability of the tests, all of which involve computations that may render the test statistic undefined for certain combinations of the values of sample statistics.

The motivation for conducting a comprehensive comparative study of accuracy, power, robustness, and applicability of alternative tests for equality of two dependent correlations is twofold. A first component comes from the foreseeable replacement of current malpractice (i.e., conducting isolated tests of significance of each of the implied correlations and comparing their outcomes) with the use of proper tests of significance of the difference between dependent correlations. If this is about to come, the choice of statistical test for this purpose must be based on evidence of their comparative performance. The second component comes from the availability of computer programs and web applications that carry out and report the outcomes of a number of alternative tests with no accompanying information regarding which test is more dependable than which other, thus leaving users uncertain about which result should be pulled out from the list (see, for example, Diedenhofen & Musch, 2015; Silver et al., 2006; Silver & Merino-Soto, 2016). To investigate the usage of the various tests considered in this paper, on 22 December 2023, we used Google Scholar to find papers published in 2023 in journals containing 'psychology' or 'psychological' in their title and that cited *cocor* (Diedenhofen & Musch, 2015), an R package (https://cran.r-project.org/web/packages/cocor) and web application (http://comparingcorrelations.org) to conduct tests of equality of dependent correlations with overlapping variables among other comparisons of correlations. The search retrieved 52 papers, but nearly half of them reported using *cocor* to conduct tests of equality of independent correlations or equality of dependent correlations with non-overlapping variables, whereas others cited the target paper for reasons not related to actual usage of the tests. Twenty-seven papers remained that reported using *cocor* for comparing dependent correlations with overlapping variables. Table A1 in Appendix A identifies each of these papers and indicates which test was reportedly used. Ten of the papers did not even report which test had been used, one other reported that all tests gave the same results, and the remaining 16 papers varied in their choice of test.

The plan of this paper is as follows. Section 2 describes the 10 tests whose performance will be compared. Section 3 describes the criteria for comparison, which cover the four aspects mentioned above, namely, Type I error rates, power, robustness, and empirical applicability. Section 4 describes the simulation method used to achieve these goals, including a description of the extensive set of scenarios (population correlations, sample sizes, and univariate distributions) under which the performance of the tests is compared. Section 5 reports the results in terms of each of the comparison criteria. Section 6 presents some empirical examples with data from recent papers in which dependent correlations were conventionally compared (i.e., via the p-value of their respective tests of significance against zero). Finally, Section 7 summarizes and discusses the results, also offering practical recommendations regarding the optimal choice of statistical test.

2 | GENERAL NOTATION AND TEST STATISTICS

Let X_1 and X_2 be the two variables whose individual correlations with variable Y are to be compared. We initially assume that all variables have standard normal distributions and a trivariate normal distribution with covariance and correlation matrix

$$P = \begin{pmatrix} 1 & \rho_{X_1X_2} & \rho_{X_1Y} \\ \rho_{X_1X_2} & 1 & \rho_{X_2Y} \\ \rho_{X_1Y} & \rho_{X_2Y} & 1 \end{pmatrix}.$$

We will simplify notation via $\rho_{1Y} := \rho_{X_1Y}$, $\rho_{2Y} := \rho_{X_2Y}$, and $\rho_{12} := \rho_{X_1X_2}$. This section describes briefly a number of methods that have been proposed to test the null hypothesis H_0: $\varrho_{1Y} = \varrho_{2Y}$ using n triplets of observations from which sample correlations r_{1Y}, r_{2Y}, and r_{12} are computed. We name the tests as in Diedenhofen and Musch (2015), where a full presentation is given. Computation of all test statistics involves some square root whose radicand is determined by sample correlations and, thus, it is not guaranteed to be positive-valued. Whenever the radicand turns up negative for the sample of concern, the test statistic is undefined.

2.1 | Pearson–Filon

This statistic (Pearson & Filon, 1898) has a standard normal distribution and is computed as

$$z_{PF} = \frac{(r_{1Y} - r_{2Y})\sqrt{n}}{\sqrt{(1-r_{1Y}^2)^2 + (1-r_{2Y}^2)^2 - 2r_{12}(1 - r_{1Y}^2 - r_{2Y}^2) + (r_{1Y}r_{2Y})(1 - r_{1Y}^2 - r_{2Y}^2 - r_{12}^2)}}. \quad (1)$$

2.2 | Olkin

This statistic (Olkin, 1967; with the correct formula in Hendrickson & Collins, 1970) has a standard normal distribution and is computed as

$$z_{O} = \frac{(r_{1Y} - r_{2Y})\sqrt{n}}{\sqrt{(1-r_{1Y}^2)^2 + (1-r_{2Y}^2)^2 - 2r_{12}^3 - (2r_{12} - r_{1Y}r_{2Y})(1 - r_{1Y}^2 - r_{2Y}^2 - r_{12}^2)}}. \quad (2)$$

2.3 | Hotelling

This statistic (Hotelling, 1940) has a Student's t distribution with $n-3$ degrees of freedom and is computed as

$$t_{H} = \frac{(r_{1Y} - r_{2Y})\sqrt{(n-3)(1+r_{12})}}{\sqrt{2|R|}} \quad (3)$$

with

$$|R| = 1 - r_{1Y}^2 - r_{2Y}^2 - r_{12}^2 + 2r_{1Y}r_{2Y}r_{12}. \quad (4)$$

2.4 | Standard Williams

This statistic (Williams, 1959) has a Student's t distribution with $n-3$ degrees of freedom and is computed as

$$t_{W} = \frac{(r_{1Y} - r_{2Y})\sqrt{(n-3)(1+r_{12})}}{\sqrt{2|R| + \bar{r}^2\frac{n-3}{n-1}(1-r_{12})^3}} \quad (5)$$

with $|R|$ from Equation (4) and

$$\bar{r} = \frac{r_{1Y} + r_{2Y}}{2}. \quad (6)$$

2.5 | Hendrickson–Stanley–Hills

This statistic (Hendrickson et al., 1970) has a Student's t distribution with $n-3$ degrees of freedom and is computed as

$$t_{\mathrm{HSH}} = \frac{(r_{1Y} - r_{2Y})\sqrt{(n-3)(1+r_{12})}}{\sqrt{2|R| + \frac{(r_{1Y}-r_{2Y})^2(1-r_{12})^3}{4(n-1)}}} \tag{7}$$

with $|R|$ from Equation (4).

2.6 | Dunn–Clark

This statistic (Dunn & Clark, 1969) has a standard normal distribution and is computed as

$$z_{\mathrm{DC}} = \frac{(Z_{1Y} - Z_{2Y})\sqrt{n-3}}{\sqrt{2-2c_{\mathrm{DC}}}} \tag{8}$$

with Z_{1Y} and Z_{2Y} given by Fisher's transformation by which

$$Z_{xy} = \frac{1}{2}\ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$$

and

$$c_{\mathrm{DC}} = \frac{r_{12}\left(1-r_{1Y}^2-r_{2Y}^2\right) - \frac{1}{2}r_{1Y}r_{2Y}\left(1-r_{1Y}^2-r_{2Y}^2-r_{12}^2\right)}{\left(1-r_{1Y}^2\right)\left(1-r_{2Y}^2\right)}. \tag{9}$$

2.7 | Steiger

This statistic (Steiger, 1980) has a standard normal distribution and is computed as

$$z_{\mathrm{S}} = \frac{(Z_{1Y} - Z_{2Y})\sqrt{n-3}}{\sqrt{2-2c_{\mathrm{S}}}} \tag{10}$$

with Z_{1Y} and Z_{2Y} given by Fisher's transformation and

$$c_{\mathrm{S}} = \frac{r_{12}\left(1-2\bar{r}^2\right) - \frac{1}{2}\bar{r}^2\left(1-2\bar{r}^2-r_{12}^2\right)}{\left(1-\bar{r}^2\right)^2}, \tag{11}$$

with $\bar{r}$ from Equation (6).

2.8 | Hittner–May–Silver

This statistic (Hittner et al., 2003) has a standard normal distribution and is computed as

$$z_{\mathrm{HMS}} = \frac{(Z_{1Y} - Z_{2Y})\sqrt{n-3}}{\sqrt{2-2c_{\mathrm{HMS}}}} \tag{12}$$

with Z_{1Y} and Z_{2Y} given by Fisher's transformation and

$$c_{\text{HMS}} = \frac{r_{12}\left(1-2\overline{r}_z^2\right) - \frac{1}{2}\overline{r}_z^2\left(1-2\overline{r}_z^2 - r_{12}^2\right)}{\left(1-\overline{r}_z^2\right)^2}, \tag{13}$$

in which

$$\overline{r}_z = \frac{\exp\left(2\overline{Z}-1\right)}{\exp\left(2\overline{Z}+1\right)} \text{ and } \overline{Z} = \frac{Z_{1Y}+Z_{2Y}}{2}.$$

2.9 | Meng–Rosenthal–Rubin

This statistic (Meng et al., 1992) has a standard normal distribution and is computed as

$$z_{\text{MRR}} = \frac{\left(Z_{1Y}-Z_{2Y}\right)\sqrt{n-3}}{\sqrt{2\left(1-r_{12}\right)h}} \tag{14}$$

with Z_{1Y} and Z_{2Y} given by Fisher's transformation and

$$h = \frac{1-f\overline{r^2}}{1-\overline{r^2}} \tag{15}$$

with

$$\overline{r^2} = \frac{r_{1Y}^2 + r_{2Y}^2}{2} \quad \text{and} \quad f = \frac{1-r_{12}}{2\left(1-\overline{r^2}\right)}$$

Meng et al. (1992) stated that $f=1$ should be used whenever the computation of f using sample statistics results in $f>1$, but the justification for this replacement is unclear. At first glance, the only apparent requirement is that the radicand in Equation (14) is positive-valued. This implies $h>0$ or, by Equation (15), $f<1/\overline{r^2}$, which can sometimes result in $f>1$. This study will use this test without replacing any $f>1$ with $f=1$, with the consequence that the test statistic is undefined when sample statistics result in $f \geq 1/\overline{r^2}$.

2.10 | Zou

This test (Zou, 2007) is based on a confidence interval. The lower (L) and upper (U) limits of a 100 $(1-\alpha)$% confidence interval for $\rho_{1Y}-\rho_{2Y}$ are

$$L = r_{1Y} - r_{2Y} - \sqrt{\left(r_{1Y}-l_1\right)^2 + \left(u_2-r_{2Y}\right)^2 - 2c\left(r_{1Y}-l_1\right)\left(u_2-r_{2Y}\right)}, \tag{16}$$

$$U = r_{1Y} - r_{2Y} + \sqrt{\left(u_1-r_{1Y}\right)^2 + \left(r_{2Y}-l_2\right)^2 - 2c_Z\left(u_1-r_{1Y}\right)\left(r_{2Y}-l_2\right)}, \tag{17}$$

with

$$l_i = \frac{\exp\left(2l_i^*\right)-1}{\exp\left(2l_i^*\right)+1}, \quad i \in \{1,2\}, \tag{18}$$

$$u_i = \frac{\exp\left(2u_i^*\right) - 1}{\exp\left(2u_i^*\right) + 1}, \quad i \in \{1,2\}, \tag{19}$$

$$l_i^* = Z_{iY} - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}, \quad i \in \{1,2\}, \tag{20}$$

$$u_i^* = Z_{iY} + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}, \quad i \in \{1,2\}, \tag{21}$$

where Z_{iY} is given by Fisher's transformation of r_{iY}, $z_{1-\alpha/2}$ is the $100(1-\alpha/2)\%$ standard normal quantile, and

$$c_Z = \frac{\left(r_{12} - \frac{1}{2}r_{1Y}r_{2Y}\right)\left(1 - r_{1Y}^2 - r_{2Y}^2 - r_{12}^2\right) + r_{12}^3}{\left(1 - r_{1Y}^2\right)\left(1 - r_{2Y}^2\right)}. \tag{22}$$

Given that $L < U$ by construction, application of this method rejects the null hypothesis of equality of correlations whenever $L > 0$ or $U < 0$.

3 | CRITERIA FOR COMPARISON

Before presenting the criteria for comparison of alternative tests for H_0: $\rho_{1Y} = \rho_{2Y}$, it should be noted that the identity stated in the null can only hold for a restricted range of values determined by the population correlation ρ_{12}. The reason is that a non-degenerate trivariate normal distribution can only exist if $|P| = 1 - \rho_{1Y}^2 - \rho_{2Y}^2 - \rho_{12}^2 + 2\rho_{1Y}\rho_{2Y}\rho_{12} > 0$. Algebraic manipulation of this inequality indicates that the possible range for ρ_{2Y} given arbitrary values for ρ_{12} and ρ_{1Y} is

$$\rho_{12}\rho_{1Y} - \sqrt{\left(1 - \rho_{12}^2\right)\left(1 - \rho_{1Y}^2\right)} < \rho_{2Y} < \rho_{12}\rho_{1Y} + \sqrt{\left(1 - \rho_{12}^2\right)\left(1 - \rho_{1Y}^2\right)}; \tag{23}$$

note that the range is not empty for any pair of values for ρ_{12} and ρ_{1Y}. The shaded areas in Figure 1 show cross-sectional slices of the region of feasibility of ρ_{1Y} and ρ_{2Y} at selected values for ρ_{12}. The intersection of these regions with the diagonal identity line describes the range of values for which a trivariate normal distribution may exist where H_0 is true. A comparison of the accuracy of tests is thus restricted to feasible true nulls (i.e., $\rho_{1Y} = \rho_{2Y}$) for each value of ρ_{12}, whereas a comparison of their power is restricted to feasible false nulls (i.e., $\rho_{1Y} \neq \rho_{2Y}$) for each value of ρ_{12}.

The first criterion for comparison is the region in the three-dimensional space of sample correlations r_{1Y}, r_{2Y}, and r_{12} for which a test statistic is real-valued. Although the region of existence of a trivariate normal distribution in the analogous three-dimensional space of population correlations is well defined and illustrated in Figure 1, it is not clear that all test statistics can accommodate any sample correlations in this region. Consider the Hotelling test statistic in Equation (3), where the radicand (Equation (4)) simply sets on sample correlations the same condition set on population correlations for the existence of a trivariate normal distribution, thus ensuring coverage of the sample space of trivariate correlations. In contrast, the standard Williams statistic in Equation (5) and the Hendrickson–Stanley–Hills statistic in Equation (7) both incorporate into the radicand additive and positive-valued terms that broaden the applicability of these statistics. It is not immediately obvious how the relatively more complex radicands in the remaining test statistics affect their corresponding regions of applicability.

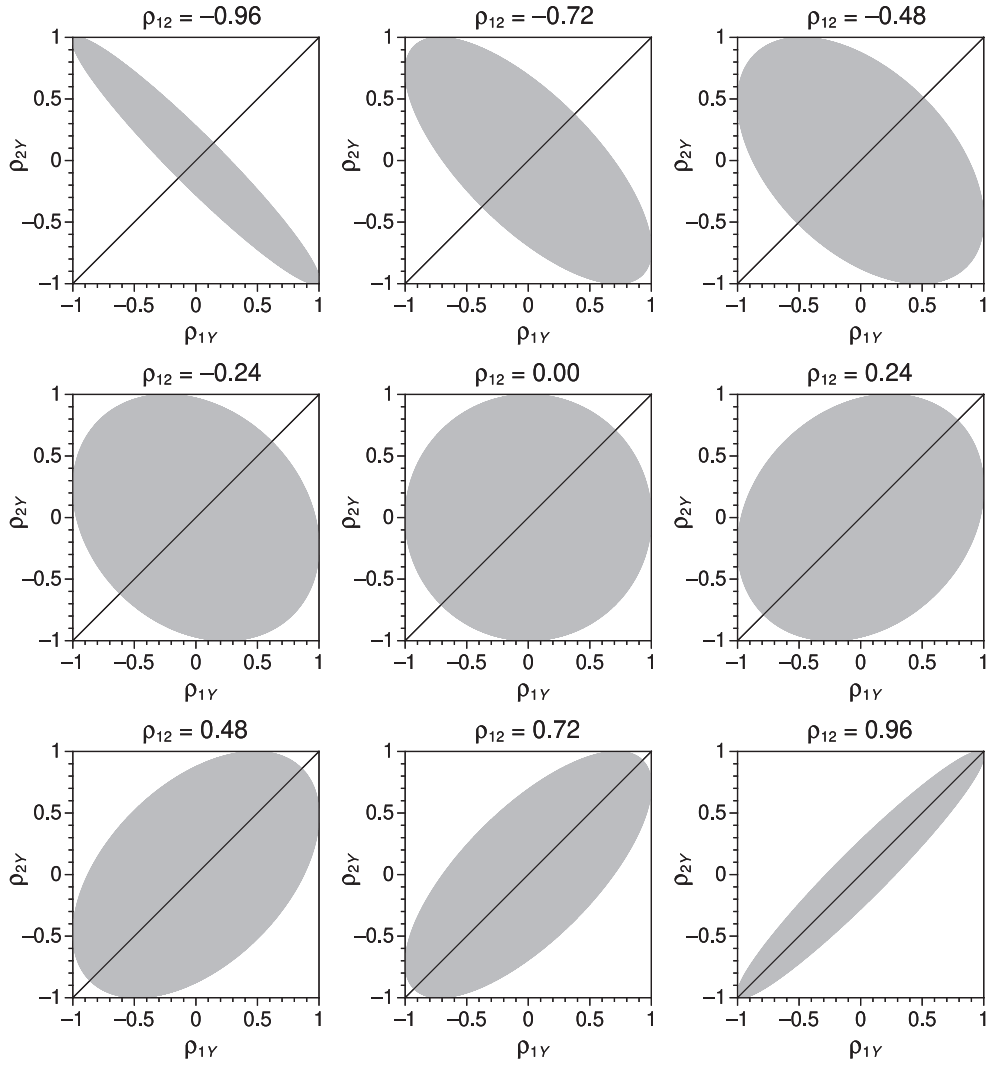

FIGURE 1 Region (shaded area) where pairs of values for ρ_{1Y} and ρ_{2Y} may exist in a trivariate normal distribution, as a function of the value taken on by ρ_{12} (panels). The diagonal in each panel is the identity line $\rho_{1Y}=\rho_{2Y}$.

The second and third criteria for comparison involve the accuracy of the tests (i.e., the match between nominal and actual Type I error rates) and their power, both assessed when observations are drawn from trivariate normal distributions and from non-normal distributions. Comparisons involving these criteria were carried out via simulations described in the next section.

4 | SIMULATION METHOD

Empirical Type I error rates were determined for each test for values of ρ_{12} between −.9 and .9 in steps of .15 and, in each case, within the admissible range of values for $\rho_{1Y}=\rho_{2Y}$ in steps of .02 (i.e., along the diagonal identity line in the panels of Figure 1 within the region indicated by the shaded area). For each combination of ρ_{12} and $\rho_{1Y}=\rho_{2Y}$, we conducted $N=500{,}000$ replications each consisting of n triplets of standard normal observations X_1, X_2, and Y drawn from the corresponding trivariate normal

distribution. In separate simulation runs, sample sizes n were 20, 50, 100, and 200. The 10 test statistics were computed for each sample in each replication, and the empirical Type I error rate for each test was computed as the proportion of replications (out of N) in which the null was rejected in a two-sided test at $\alpha = .05$, keeping separate counts of rejection in each tail.

Power was determined for each test in a similar manner with 200,000 replications per condition and for the same set of values of ϱ_{12}, drawing triplets from trivariate normal distributions with $\varrho_{1Y} \neq \varrho_{2Y}$ in pairings described by the lattice points in Figure 2. Lattice points are located along lines perpendicular to the identity line and with respect to reference points on the diagonal ranging from $\varrho_{1Y} = \varrho_{2Y} = -.96$ to $\varrho_{1Y} = \varrho_{2Y} = .96$ in steps of .04. Successive lattice points along off-diagonal directions are created by progressively subtracting .04 from the reference value of ϱ_{1Y} and adding .04 to the reference value of ϱ_{2Y}. Power at a given lattice point was determined only if a trivariate normal distribution exists under the applicable value of ϱ_{12}, for example, for the lattice points within the region of existence illustrated in Figure 2 for $\varrho_{12} = -.60$ (red ellipse) or for $\varrho_{12} = .90$ (blue ellipse). No lattice points were defined below the diagonal in Figure 2 because of the symmetry of the tests with respect to the sign of the difference between ϱ_{1Y} and ϱ_{2Y}. The off-diagonal lines on which lattice points are placed allow assessment of whether power varies as a simple function of the absolute value of the difference $\varrho_{1Y} - \varrho_{2Y}$ (which is constant for all lattice points falling on the same imaginary line parallel to the identity line) or differs also with the point on the diagonal relative to which $|\varrho_{1Y} - \varrho_{2Y}|$ is placed, that is, with the location of the intersection of the off-diagonal line and the identity line. The location of this intersection is the point at which both ϱ_{1Y} and ϱ_{2Y} equal their average $(\varrho_{1Y} + \varrho_{2Y})/2$. In other words, different off-diagonal lines define lattice points of increasing value of $|\varrho_{1Y} - \varrho_{2Y}|$ at fixed values of $(\varrho_{1Y} + \varrho_{2Y})/2$.

Triplets of observations drawn from non-normal distributions were used to assess accuracy and power under violation of the assumption of trivariate normality. Many forms of non-normality have been described in psychological data at both the univariate and multivariate levels (Cain et al., 2017). Hence, a thorough assessment of robustness to arbitrary violations of trivariate normality is impractical. Our goal here is thus limited to describing results for a number of plausible cases that nevertheless shed light on the robustness (or lack thereof) of the tests under analysis. This was done in three different ways. For the first, data from beta distributions with $a = 1$ and $b = 1$ (i.e., uniform) or $a = 2$ and $b = 5$ (positively skewed) were generated by first drawing triplets of observations in Z_1, Z_2, and Z_y from a trivariate standard normal distribution with correlation matrix **P**. Sample values were

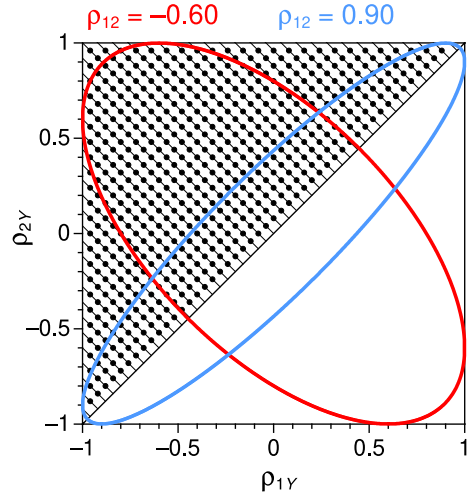

FIGURE 2 Lattice points (small circles) for the analysis of power. For each value of ρ_{12}, only lattice points are used that fall within the region of existence, as illustrated by the red and blue ellipses (Colour is available online).

then transformed via $X_1 = F^{-1}\left(\Phi\left(Z_1\right); a_1, b_1\right)$, $X_2 = F^{-1}\left(\Phi\left(Z_2\right); a_2, b_2\right)$, and $Y = F^{-1}\left(\Phi\left(Z_y\right); a_y, b_y\right)$, where Φ is the standard normal distribution function and F^{-1} is the beta quantile function with parameters *a* and *b*. Note that beta parameters are defined separately for each variable, thus allowing any combination of forms for the univariate distributions of X_1, X_2, and Y. For the second way, triplets Z_1, Z_2, and Z_y generated as described above were transformed via $X_1 = \exp(Z_1)$, $X_2 = \exp(Z_2)$ and $Y = \exp(Z_y)$, which produces data with univariate Lognormal(0, 1) distributions. Finally, normal mixtures were generated whose first component (with probability .9) was the trivariate standard normal distribution with correlation matrix **P** and whose second component (with probability .1) was a trivariate normal distribution with the same correlation matrix but larger variances (2, 4, and 10) for all variables.

Generation of non-normal variables via transformation (or mixture) of normal variables ensured that the former matched the target correlations ϱ_{12}, ϱ_{1Y}, and ϱ_{2Y} included in analogous simulations with normally distributed data. Specifically, for our mixtures of normal variables with the same correlation matrix and differing only in a common variance, the correlation matrix of the mixture is that of the original variables and, then, the components were generated with the applicable target correlations. When standard normal variables X and Y with correlation ϱ_{xy} are exponentially transformed into lognormal variables $\widetilde{X}$ and $\widetilde{Y}$, the correlation between the latter is given by

$$\rho_{\widetilde{x}\widetilde{y}} = \frac{\exp\left(\rho_{xy}\right) - 1}{\exp(1) - 1} \tag{24}$$

(De Veaux, 1976). Then the correlation matrix for the originating normal variables was set by inverting this relation so that the resulting lognormal variables had the required target correlations. This, in turn, reduces the breadth of target correlations that can be explored because the lowest possible correlation between lognormal variables is $-\exp(-1) \approx -.3679$ (De Veaux, 1976). Finally, beta variables generated via the inverse transform method do not generally keep the correlation of the originating normal variables and there is no closed-form expression to compensate for this fact in the correlation matrix of the generating normal variables. Yet, Appendix B shows that the original correlations are almost exactly preserved with the beta parameters of our choice. The skewness and kurtosis of each of the resulting non-normal univariate distributions are listed in Table 1; note that heteroscedasticity generally holds for bivariate distributions with non-normal marginals.

The study includes cases in which all variables had the same non-normal univariate distribution and a number of other cases in which form of distribution varied across variables while keeping the same distribution for X_1 and X_2 to ensure that the null hypothesis remains true or false as needed. In all cases, data were generated under all of the applicable conditions described earlier for similar analyses under trivariate normality, except that we used $N = 250{,}000$ replications in studies of accuracy and $N = 100{,}000$ replications in studies of power.

Because of the extensive and time-consuming computations required for completion of the study, Fortran code was written to carry out all the simulations.

TABLE 1 Skewness and kurtosis of the non-normal univariate distributions used in the simulations.

Distribution	Skewness	Kurtosis
Uniform	0.000	1.800
Beta (2,5)	0.596	2.880
Lognormal (0,1)	6.149	113.936
Mixture 0.9 N(0,1) + 0.1 N(0,2)	0.000	3.223
Mixture 0.9 N(0,1) + 0.1 N(0,4)	0.000	4.438
Mixture 0.9 N(0,1) + 0.1 N(0,10)	0.000	9.059

5 | RESULTS

5.1 | Region of empirical applicability

Obtaining the region of empirical applicability amounts to finding out which triplets (r_{1Y}, r_{2Y}, r_{12}) on the cube $[-1, 1] \times [-1, 1] \times [-1, 1]$ result in a positive-valued radicand on computation of each test statistic. A minimum requirement is that the statistic is real-valued when sample correlations take any values compatible with a trivariate normal distribution, and we have already discussed that the Hotelling statistic in Equation (3) has this exact characteristic. The standard Williams and Hendrickson–Stanley–Hills statistics (both with $n = 100$; see the denominators in Equations (5) and (7)) are similar in this respect and their empirical region of applicability does not differ much from the theoretical region of existence of trivariate normal distributions (see the second and third columns from the left in Figure 3).

At the other end, the Hittner–May–Silver test is applicable for all triplets on the cube and it is thus not displayed in Figure 3. The remaining statistics vary in their region of applicability, as seen in Figure 3. All tests are applicable with any triplet (r_{1Y}, r_{2Y}, r_{12}) compatible with a trivariate normal distribution, except the Meng–Rosenthal–Rubin test without the imposition $f \leq 1$ discussed in Section 2.9 above. As seen in the rightmost column of Figure 3, not replacing an $f \geq 1/\overline{r^2}$ with $f = 1$ prevents computation of the statistic in some areas of the region of existence of trivariate normal distributions, although only when the sample value of r_{12} is negative. The consequences of setting versus not setting a bound on f will be described later.

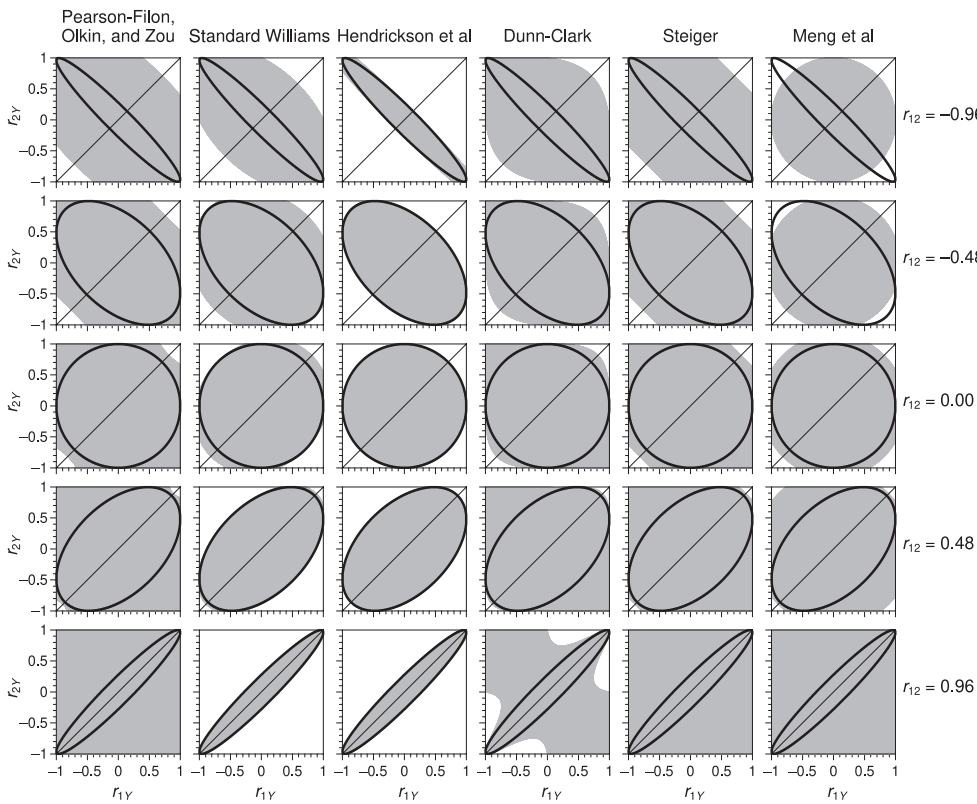

FIGURE 3 Region of applicability (shaded area) of tests (columns) at several sample values for r_{12} (rows). The elliptical region of existence of a trivariate normal distribution is also shown in each panel for reference.

5.2 | Accuracy with trivariate normal observations

For a preliminary illustration, Figure 4 shows the empirical sampling distribution of each of the nine test statistics in the simulation condition with $(\rho_{1Y}, \rho_{2Y}, \rho_{12}) = (.80, .80, .75)$ and $n = 50$ (recall that the Zou test does not come in the form of a test statistic whose sampling distribution can be inspected). The continuous curve in each panel is the applicable asymptotic distribution (i.e., standard normal or Student's t with $n - 3$ degrees of freedom), to which empirical distributions generally match except for the Hittner–May–Silver test. Empirical Type I error rates at the nominal size−.05 test are also shown for each tail in the panels of Figure 4.

The similarity between empirical and nominal rejection rates at each tail indicates an acceptable accuracy for most tests, although some of them seem a little conservative (e.g., Pearson–Filon and Olkin) while others seem a little liberal (e.g., Hotelling and Hendrickson–Stanley–Hills), leaving aside the unacceptably liberal Hittner–May–Silver test. Empirical sampling distributions are also approximately symmetric in Figure 4, as was the case for the all other simulation conditions. For this reason, subsequent graphical results display Type I error rates aggregated over tails.

Figure 5 displays summary results (at $n = 200$) for the three tests whose accuracy proved unacceptable but whose Type I error rates turned out to be essentially invariant with sample size n. The Hittner–May–Silver test (Figure 5a) is accurate only when $\varrho_{1Y} = \varrho_{2Y}$ is very near zero, something that is generally the case when ϱ_{12} is negative (see the top row in Figure 1). Yet, as the positive value of ϱ_{12} increases, the test becomes increasingly liberal as the common value of ϱ_{1Y} and ϱ_{2Y} (under the null) moves away from zero. On the other hand, the Hotelling and Hendrickson–Stanley–Hills tests (Figure 5b) are very (and identically) inaccurate when ϱ_{12} is negative unless $\varrho_{1Y} = \varrho_{2Y}$ is virtually zero, whereas their accuracy increases across the range of values that $\varrho_{1Y} = \varrho_{2Y}$ can actually take as the positive value of ϱ_{12} increases.

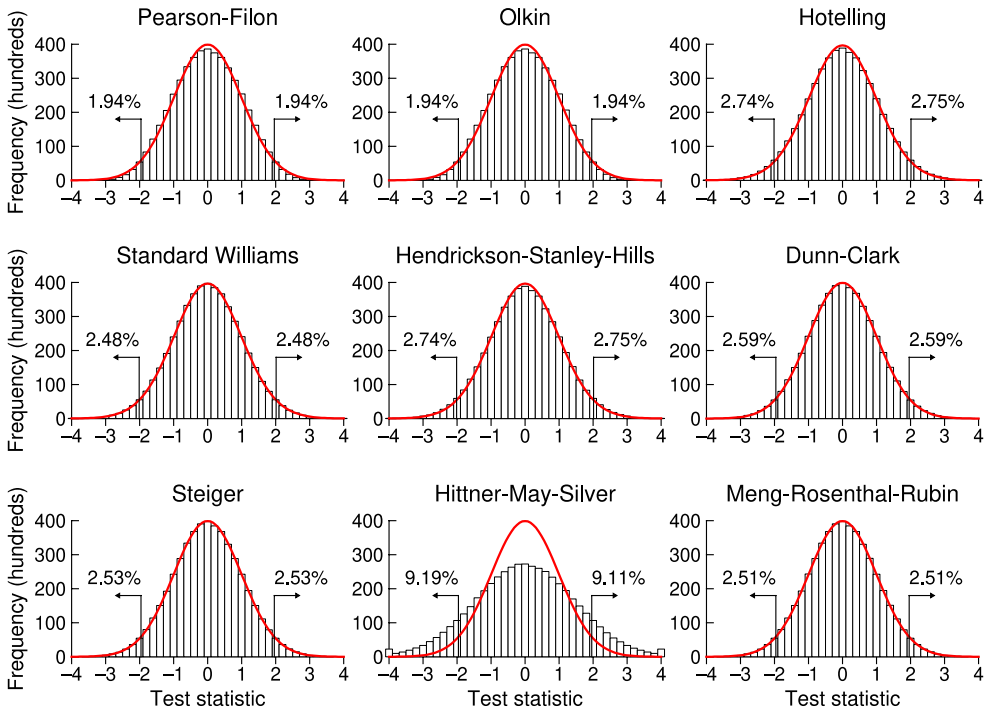

FIGURE 4 Sampling distribution (histogram) of each test statistic when $\rho_{1Y} = \rho_{2Y} = .80$, $\rho_{12} = .75$, and $n = 50$. The asymptotic distribution (red curve) is drawn to scale. Percentage of cases beyond the critical points for a two-sided size-.05 test are indicated at the tails (Colour is available online).

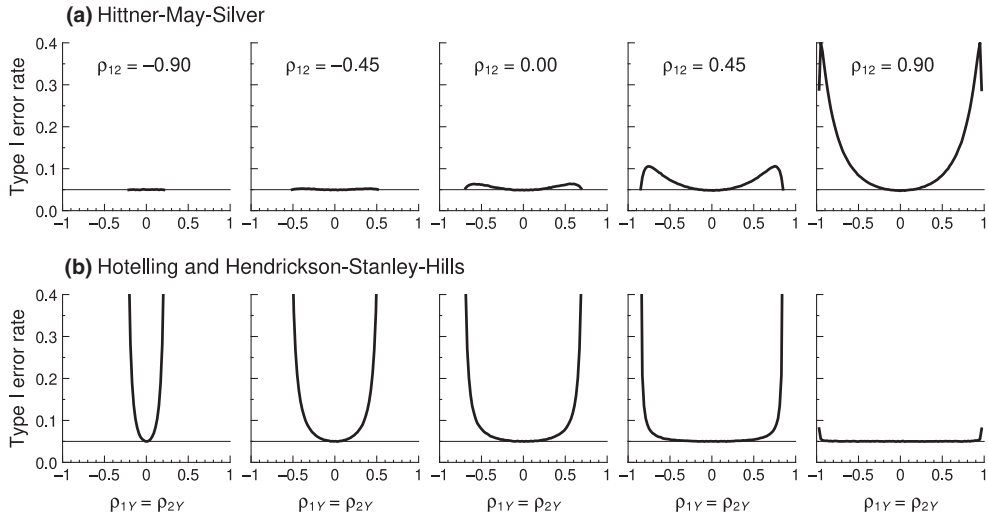

FIGURE 5 Type I error rates of inaccurate tests as a function the actual value of $\rho_{1Y}=\rho_{2Y}$ (horizontal axis) for selected values of ρ_{12} (columns) when sample size is $n=200$. The nominal $\alpha=.05$ level is indicated by a thin horizontal line.

Results for the seven other tests are displayed in Figure 6 in another format, where the vertical axis in each panel also spans a narrower range (from .03 to .07) than that in Figure 5. The Pearson–Filon and Olkin tests (black curves in Figure 6) have identical (in)accuracy that varies with sample size n also identically. Although these two tests are much less inaccurate than those for which results

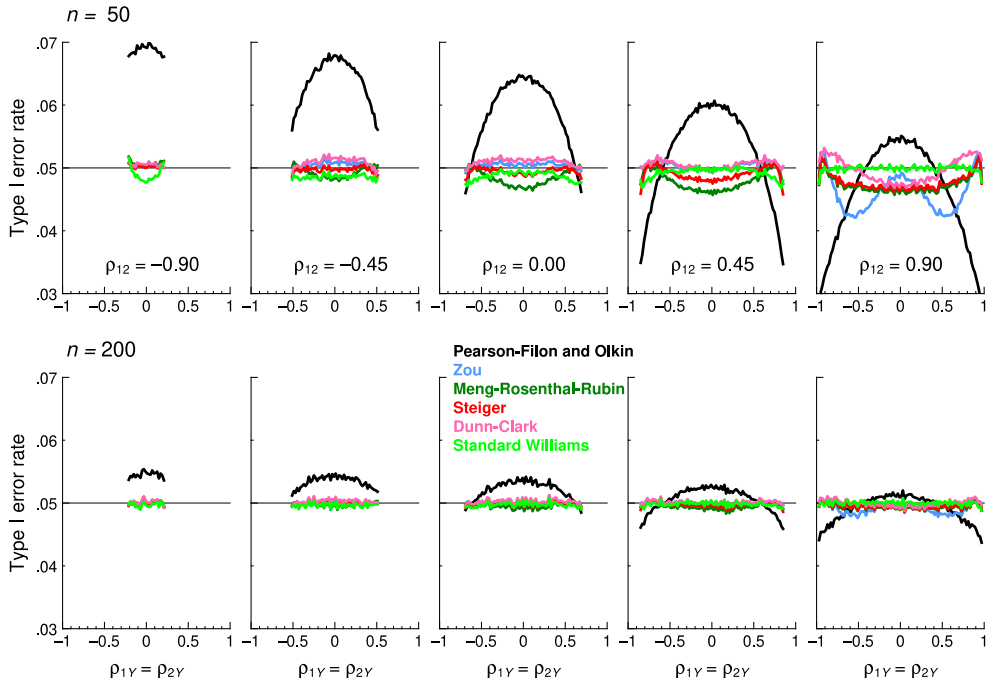

FIGURE 6 Type I error rates of seven tests (see legend) as a function the actual value of $\rho_{1Y}=\rho_{2Y}$ (horizontal axis) for selected values of ρ_{12} (columns). Sample sizes are $n=50$ (top row) and $n=200$ (bottom row). The nominal $\alpha=.05$ level is indicated by a thin horizontal line (Colour is available online).

were displayed in Figure 5, they are both outperformed by the five other tests displayed in Figure 6 (coloured curves), whose accuracies are similar at large sample sizes (bottom row in Figure 6). Yet, with small sample sizes (top row in Figure 6) and positive values of ϱ_{12}, differences can be observed that allow the tests to be ranked in increasing order of accuracy down the list in the legend to Figure 6. At one extreme, the Zou test (blue curves) displays a wavy pattern that departs the most from the .05 line when $\varrho_{12} = .9$; at the other extreme, the standard Williams test (light green curves) stays closest overall to the .05 line. Full graphical results for all tests at all sample sizes are presented in Section A of the Appendix S1.

Based on these results on accuracy, only five tests are advisable (standard Williams, Dunn–Clark, Steiger, Meng–Rosenthal–Rubin, and Zou) although the latter is inferior with small samples and large ϱ_{12}. Yet, it is worth looking at the region within which sample values of r_{1Y} and r_{2Y} result in non-rejection of the null by each of these five tests, a region that is in turn confined within the region where each test can actually be used by the requirement of a positive radicand. Figure 7 shows these regions (green areas) for the tests of concern at selected sample values of r_{12} (rows) when $n = 50$.

All tests appear virtually identical at any given r_{12} within the region of feasible trivariate distributions (ellipse in each panel). It is also worth pointing out that the Dunn–Clark test (second column from the left in Figure 7) displays an anomaly at the top-left and bottom-right corners of the panel (i.e., when r_{1Y} and r_{2Y} are both large in absolute value but have opposite signs), an anomaly that spreads out to a larger region as r_{12} increases and that progressively disappears as the sample size n increases. Admittedly, the applicable combinations of r_{12}, r_{1Y}, and r_{2Y} in this area are empirically impossible, but structural non-rejection of the null in such cases defies logic.

This (inconsequential) anomaly of the Dunn–Clark test raises the question of whether some anomaly may also obtain when the Meng–Rosenthal–Rubin test is applied under the admonition to use $f = 1$ in Equation (15) whenever computation of f using sample statistics renders a value greater than unity. This replacement has two consequences. One is that it turns an otherwise undefined statistic into a real-valued one whose magnitude might result in inadequate rejections or non-rejections; the other is that it alters the value of a well-defined statistic when $1 < f < 1/r^2$.

Results for this (original) version of Meng–Rosenthal–Rubin test are displayed in the fifth column from the right of Figure 7, where it is immediately obvious that the test becomes computable at all points, thus (unnecessarily) extending its applicability. A second feature not apparent in Figure 7 is that the region of non-rejection is minimally larger for the original form within the region of existence

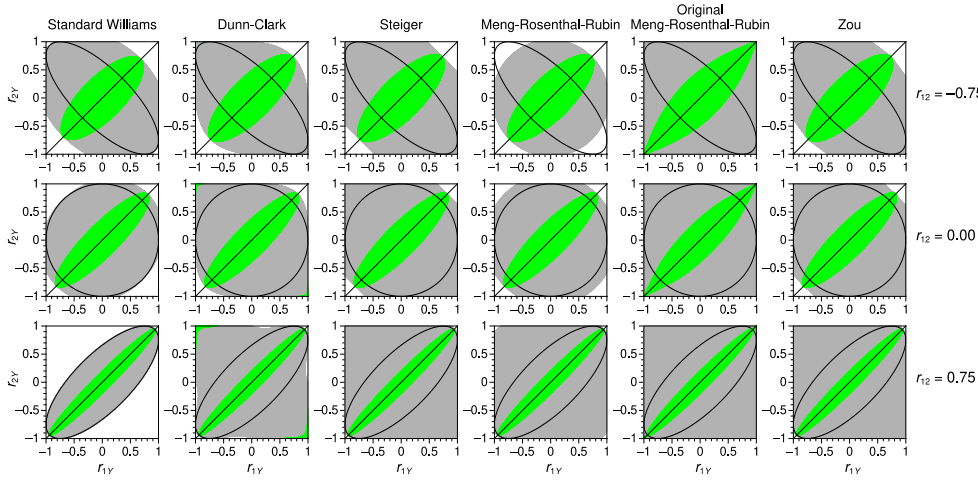

FIGURE 7 Region of non-rejection (green area) in the two-dimensional space with coordinates r_{1Y} and r_{2Y} for accurate tests (columns) at selected values for r_{12} (rows) and with $n = 50$. Each panel also shows the region of applicability of the test (grey area) and the region of existence of trivariate normal distributions (ellipse) (Colour is available online).

of trivariate normal distributions. This is more apparent with small samples and negative ϱ_{12} and it is caused by making $f=1$ when $1<f<1/r^2$ and, thus, even when the test statistic is well defined. The desirability of an artificially larger non-rejection area may be contentious, but we see little reason for the admonition to define f as the smaller of unity and its computed value via Equation (15).

In sum, as far as accuracy and applicability are concerned, the list of recommendable tests in order of preference are standard Williams, Dunn–Clark, Steiger, Meng–Rosenthal–Rubin (without setting a bound for f at unity), and Zou (see accuracies for these tests in Figure 6).

5.3 | Power with trivariate normal observations

Power results are reported here for the five tests just mentioned, but full results for all tests are presented in Section B of the Appendix S1. For a quick glimpse, Figure 8 shows the power of each test (rows) as a function of the absolute difference $|\rho_{1Y}-\rho_{2Y}|$ at selected values of ρ_{12} (columns) along each of the applicable off-diagonal lines in Figure 2 for $n=50$ (red curves) and $n=200$ (blue curves). The most apparent feature is that all tests appear identically powered, as there are no discernible differences across rows. Two other unsurprising characteristics are that power increases with sample size (blue versus red curves in the panels of Figure 8) and with the magnitude of the absolute difference between ρ_{1Y} and ρ_{2Y} (horizontal axis in the panels of Figure 8). Two further characteristics are also apparent. One is that the power to detect an absolute difference of some magnitude between ρ_{1Y} and ρ_{2Y} (i.e., at any arbitrary point along the horizontal axis) increases with increasing ρ_{12} (i.e., rightward each row in Figure 8). The other is that the power curves for different values of $(\rho_{1Y}+\rho_{2Y})/2$ (set of identically coloured curves in each panel) do not superimpose, the less so as ρ_{12} approaches zero from either side. Altogether, this means that power is not a simple function of $|\rho_{1Y}-\rho_{2Y}|$ but that it also depends heavily on $(\rho_{1Y}+\rho_{2Y})/2$ and ρ_{12}.

This dependence is best appreciated in perspective plots of the power surface defined on the two-dimensional space of Figure 2 at different values of ϱ_{12}. Figure 9 displays these plots for the standard Williams test; plots for other tests were virtually identical. Note that the identity line in Figure 2 runs slightly tilted from left to right in each panel of Figure 9 and that the off-diagonal lines in Figure 2, which represent locations with increasing value of $|\varrho_{1Y}-\varrho_{2Y}|$ at a fixed value for $(\varrho_{1Y}+\varrho_{2Y})/2$, run perpendicular to the identity line in the panels of Figure 9, whereas lines parallel to the identity line describe power at locations of constant $|\varrho_{1Y}-\varrho_{2Y}|$ for different values for $(\varrho_{1Y}+\varrho_{2Y})/2$. The U-shaped form of the latter lines within the region where power increases from 0 to 1 indicates that, for any fixed value of $|\varrho_{1Y}-\varrho_{2Y}|$, power increases as $(\varrho_{1Y}+\varrho_{2Y})/2$ moves away from 0.

5.4 | Robustness

Non-normality deteriorated accuracy similarly for all tests, on top of the fact that tests already differed in accuracy under trivariate normality (see Figures 5 and 6). The magnitude and form of deterioration differed greatly across the cases under study. To illustrate the mildest deterioration, Figure 10 shows aggregated Type I error rates for the standard Williams test when all variables had uniform distributions (Figure 10a), Beta(2, 5) distributions (Figure 10b), or mixed uniform and Beta(2, 5) distributions (Figure 10c,d). In these cases, the deterioration was largely invariant with sample size (compare the red and blue curves in each panel of Figure 10) and was milder when the univariate distribution of Y differed from the common univariate distribution of X_1 and X_2 (Figure 10c,d) than it was when the three variables had the same univariate distribution (Figure 10a,b). In comparison to the corresponding plots in Figure 6 for accuracy of the same test (light green curves) under trivariate normality, uniform distributions for all variables (Figure 10a) do not affect accuracy when $\rho_{1Y}=\rho_{2Y}$ are very close to zero but they make the test increasingly liberal as $\rho_{1Y}=\rho_{2Y}$ move away from zero within their possible range of values (which broadens as ρ_{12} increases). Beta(2, 5) distributions for all variables (Figure 10b) also do

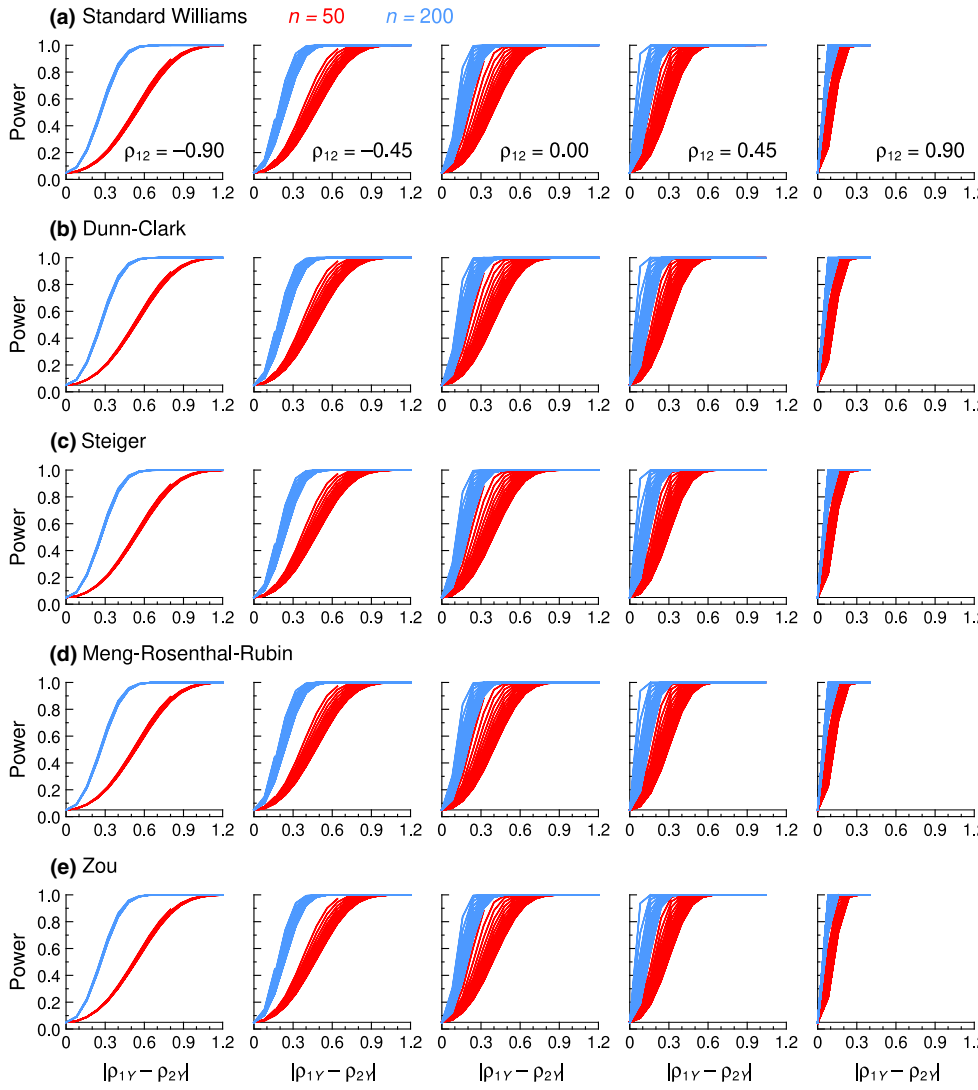

FIGURE 8 Power of five tests (rows) at selected values of ρ_{12} (columns) as a function of the absolute difference between ρ_{1Y} and ρ_{2Y} (horizontal axis) when $n = 50$ (red curves) and $n = 200$ (blue curves) (Colour is available online).

not affect the accuracy of the test when $\rho_{1Y} = \rho_{2Y}$ are virtually zero but, in contrast, they make the test increasingly more conservative (liberal) as $\rho_{1Y} = \rho_{2Y}$ become increasingly negative (positive) within their possible range of values. Mixed forms of distribution across variables (Figure 10c,d) clearly reduced the deterioration, a result that is likely limited to the particular distributions considered here. Non-normality also resulted in symmetric sampling distributions of test statistics that were also centred on zero (i.e., similar to what was shown in Figure 4 under normality) and, thus, Type I error rates in each of the tails were virtually identical. The remaining tests with acceptable accuracy under normality (see Figure 6) displayed a deterioration under non-normality that was similar to that shown in Figure 10 for the standard Williams test.

The lack of generality of the robustness results just discussed is corroborated by the results obtained for the remaining forms of distribution listed in Table 1, which are also illustrated in Figure 11 in a different form and with a broader range of Type I error rates (between 0 and .4, compared to between 0

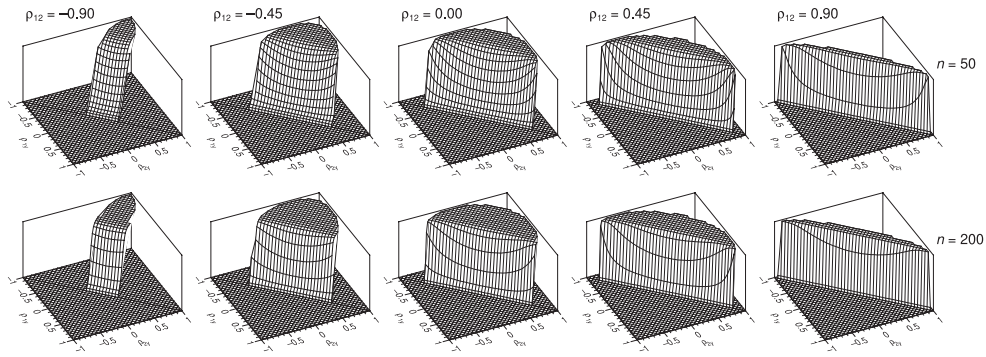

FIGURE 9 Perspective plots of the power curves plotted in Figure 8 for the standard Williams test when $n = 50$ (top row) and $n = 200$ (bottom row).

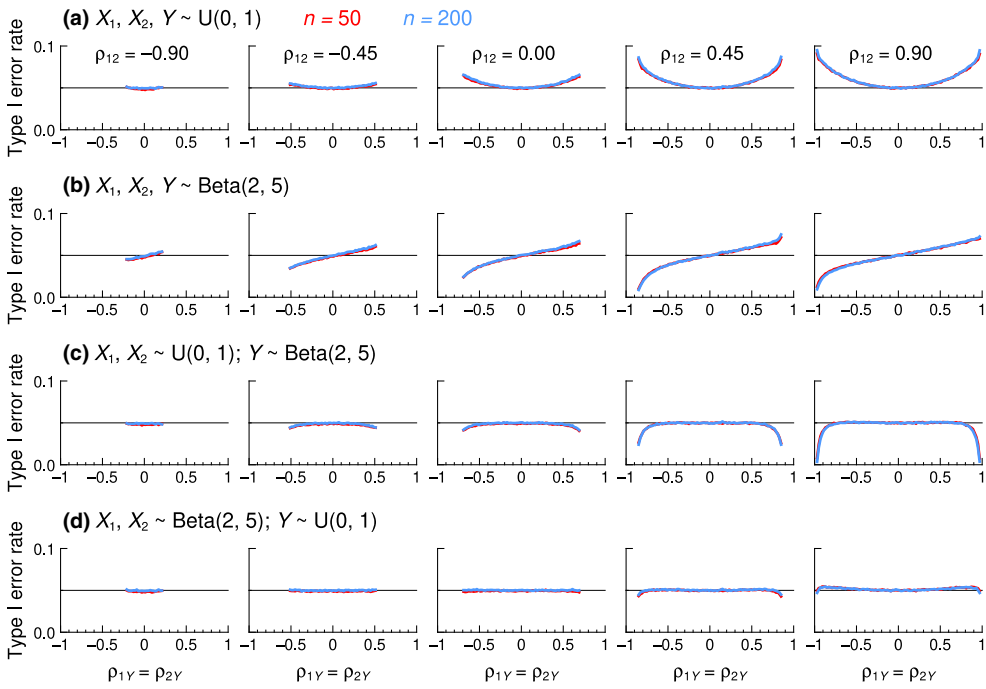

FIGURE 10 Type I error rates of the standard Williams test as a function of the actual value of $\rho_{1Y} = \rho_{2Y}$ (horizontal axis) for selected values of ρ_{12} (columns) when (a) all variables have uniform distributions, (b) all variables have Beta(2, 5) distributions, (c) X_1 and X_2 have uniform distributions while Y has a Beta(2, 5) distribution, and (d) X_1 and X_2 have Beta(2, 5) distributions while Y has a uniform distribution. Sample sizes are $n = 50$ (red curves) and $n = 200$ (blue curves). The nominal $\alpha = .05$ level is indicated by a thin horizontal line (Colour is available online).

and .1 in Figure 10) for the standard Williams test. The lognormal distribution (blue curves in Figure 11) resulted in the most dramatic and nonlinear deterioration. On the other hand, normal mixtures (green, red, and pink curves in Figure 11) resulted in a deterioration whose magnitude was constant across the board but increased with the variance difference between the two components of the mixture. In addition, the magnitude of deterioration increased with sample size (top versus bottom rows in Figure 11). Full results for all tests are presented in Sections C, E, and G–J of the Appendix S1.

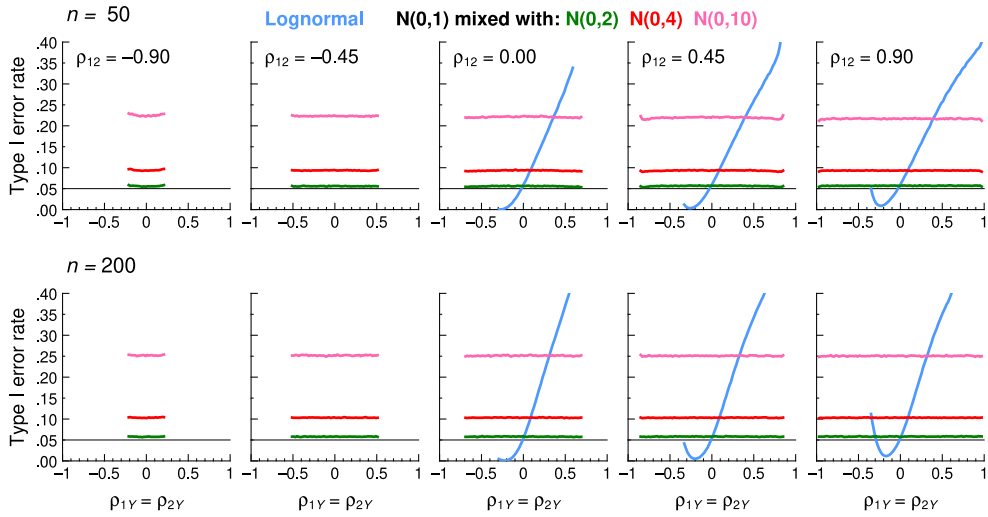

FIGURE 11 Type I error rates of the standard Williams test as a function of the actual value of $\rho_{1Y}=\rho_{2Y}$ (horizontal axis) for selected values of ρ_{12} (columns) when all variables have lognormal distributions (blue curves), or follow normal mixtures as indicated in the legend at the top. The feasible range of values for ρ_{12}, ρ_{1Y}, and ρ_{2Y} with normal mixtures is the same as that for normal variables, whereas the feasible range for lognormal distributions is narrower and renders no possible triplets for the leftmost two panels in each row. Sample sizes are $n=50$ (top row) and $n=200$ (bottom row). The nominal $\alpha=.05$ level is indicated by a thin horizontal line (Colour is available online).

Given the preceding results, it comes as no surprise that power curves for the cases considered in Figure 10 had the same features displayed in Figures 8 and 9 for trivariate normal data. Power curves for the cases considered in Figure 11 were similarly sigmoidal, although they increased from a lower asymptote at the inappropriate Type I error rates depicted in Figure 11. Graphical presentation of these results is deferred to Sections D and F of the Appendix S1.

6 | EMPIRICAL EXAMPLES

Correlational studies are the natural context of application of the tests considered in this paper, but tests are also needed in other types of study in which dependent correlations are compared for whatever reason (e.g., Blake & Palmisano, 2021; Ćoso et al., 2023). A research area in which these tests are frequently needed is the search for cognitive or behavioural correlates of the 2D–4D digit ratio (i.e., the ratio of the lengths of the index and ring fingers), and this is the area from which the coming examples were drawn. Another area of application is research on correlates of state versus trait anxiety (e.g., Guo et al., 2022; Menculini et al., 2023), but no examples will be given from this context. It should be noted that the following examples do not come from papers in which the core of the analysis was a comparison of correlations, and the only criterion guiding the choice of papers was availability of the value of r_{12} in the paper itself, a correlation that is rarely reported when the nominal correlations of concern are only r_{1Y} and r_{2Y}.

Since the 2D–4D digit ratio can be measured on each of the hands, a common question of interest is whether the target variable Y correlates equally with the right-hand ratio (X_1) and the left-hand ratio (X_2). For instance, Brosnan et al. (2011) reported that the correlation between right-hand digit ratio and Java grade in a sample of 73 first-year students taking a course in Java programming reached statistical significance ($r_{1Y}=-.22$; $p=.03$) but not so for the left-hand ratio ($r_{2Y}=-.13$; n.s.), whereas the left- and right-hand ratios were significantly correlated ($r_{12}=.46$; $p<.01$). Use of the standard Williams test yields

$t_W = -0.7420$ with $p = .461$, so that there is no basis for supporting the notion that digit ratios in the right versus the left hand are differently correlated to Java grades, despite the anecdotal result that one of the correlations is significantly different from zero and the other is not.

In another paper from the same research field, Uzun and Tok (2023) reported correlational analyses of right-hand and left-hand digit ratios (variables X_1 and X_2) with attention gathering skill (variable Y) in 112 children aged 60–72 months. The correlation was significant for the right hand ($r_{1Y} = .209$; $p < .05$) but not for the left hand ($r_{2Y} = .086$; n.s.), while the left- and right-hand ratios correlated $r_{12} = .114$ (which is not significantly different from zero). Use of the standard Williams test yields $t_W = 0.9846$ with $p = .327$, so that there is no basis for supporting the notion that digit ratios in the right versus the left hand are differently correlated to attention-gathering skill, despite the anecdotal result that only one of the correlations is significantly different from zero (Incidentally, table 7 in Uzun and Tok (2023) gives $r_{2Y} = -.086$; if this were the actual value, the standard Williams test would yield $t_W = 2.3795$ with $p = .019$, rejecting the null that attention-gathering skill holds the same relation to the right- and left-hand digit ratio).

Also of interest in this area is the search for associations between digit ratio in one of the hands and personality factors, cognitive performance in various tasks, or behavioural or physiological measures. In such cases, digit ratio plays the role of variable Y whereas the other measures define a set of J variables X_j (with $1 \leq j \leq J$ and $J > 2$). At first sight, this situation may seem to imply the null H_0: $\varrho_{1Y} = \varrho_{2Y} = \ldots = \varrho_{JY}$, for which statistical tests have been developed (see, for example, Choi, 1977; Cohen, 1989; Olkin & Finn, 1990). Yet, the context for such a null typically implies that the X_j are measures of the same variable at different time points in a longitudinal study and the goal is to check out whether the correlation between X and Y across the J occasions remains stable. In the context of our examples here, there is no theoretical expectation that digit ratio should hold the same relation with, say, all dimensions of personality. The actual research question is instead whether digit ratio is more strongly associated with some personality dimension than it is with others (see, for example, Fink et al., 2004) and, thus, it requires pairwise analyses with multiple tests of the simple null H_0: $\varrho_{jY} = \varrho_{kY}$ for all $j \neq k$. Although each of these multiple tests leads to an individual claim (and, hence, there is no risk of inflation of Type I error rates that might demand alpha adjustments; see García-Pérez, 2023; Rubin, 2021, 2024), a researcher may still want to set the alpha level different from the usual .05 for other reasons (Maier & Lakens, 2022).

The handling of comparisons of correlations in a case like this one is illustrated here with data from Del Giudice and Angeleri (2016), a study in which 135 boys and 150 girls contributed measures of right-hand digit ratio (variable Y) and three measures of attachment styles: avoidance (variable X_1), preoccupation (variable X_2), and felt security (variable X_3). Intercorrelations among these variables were reported in their Table 1 separately for the samples of boys and girls. Results of our pairwise analyses of equality of correlations with the standard Williams test in the sample of girls are presented in Table 2, with the outcome that equality of correlation with digit ratio is not rejected for any pair of attachment styles. It should be stressed that an analogous and independent set of tests for the sample of boys could be conducted but, depending on the goal of the analysis, it might be more fitting to use the ANOVA-like strategy developed by Bilker et al. (2004) for that purpose.

TABLE 2 Pairwise standard Williams tests of equality of dependent correlations for data from the sample of girls in the study of Del Giudice and Angeleri (2016).

Pairing	r_{jY}	r_{kY}	r_{jk}	t_W	p
$j = 1, k = 2$	−.11	.10	−.45	−1.5067	.134
$j = 1, k = 3$	−.11	−.03	−.33	−0.5979	.551
$j = 2, k = 3$	.10	−.03	−.30	0.9817	.328

7 | DISCUSSION

7.1 | Summary of results

Our comparative analysis of the performance of the 10 tests can be summarized as follows. When variables have a trivariate normal distribution:

1. all tests are applicable under any possible combination of the values that sample correlations may take. This holds for the Meng–Rosenthal–Rubin test only when administered with the original admonition to replace any occurrence of $f > 1$ in Equation (15) with $f = 1$.
2. Type I error rates stay at their nominal level across all possible combinations of population correlations for only five of the tests (in order of preference, standard Williams, Dunn–Clark, Steiger, Meng–Rosenthal–Rubin, and Zou), whose Type I error rates are also barely affected by sample size.
3. power curves are similar for the five tests just mentioned and, across the board, power increases with sample size and with effect size defined as the absolute value of the difference between ρ_{1Y} and ρ_{2Y}. Two additional factors that modulate power are the average of the two population correlations ρ_{1Y} and ρ_{2Y} under comparison (with power increasing faster with effect size as this average correlation moves away from zero) and the value of the population correlation ρ_{12} (with power also increasing faster with effect size as ρ_{12} increases). The former feature was anecdotally noted by May and Hittner (1997b). The presence of these modulating factors raises a warning flag on the results of power analyses that only consider effect size defined as a difference in correlation (see Faul et al., 2009).

When variables are instead non-normally distributed in the limited set of forms explored here, accuracy of the five tests deteriorates to varying extents from negligibly in some cases (see Figure 10) to unacceptably in others (see Figure 11). Further comments on non-normal distributions are deferred to the next subsection.

The results just summarized modulate those reported in previous papers that addressed the analysis for a smaller number of test statistics, under a narrower set of population correlations, with smaller samples, and with fewer replications to estimate Type I error rates and power. By exploring combinations of population correlations more thoroughly than was done in previous studies and by using substantially more replications to estimate Type I error rates and power, this study has identified inadequate performance of some tests whose deficiencies only show within certain ranges of the parameter space (see, for example, accuracy results in Figure 5). Analogously, a presumed inaccuracy of the standard Williams test with trivariate normal observations when $\varrho_{1Y} = \varrho_{2Y} = .7$ and $\varrho_{12} = .1$ for $n \in \{20, 50, 100, 300\}$ and Type I error rates estimated with 2000 replications (see, for example, May & Hittner, 1997a; see also Hittner et al., 2003) was not replicated in our results, which do not show any sign of a singularity at or near this region (see the light green curves in Figure 6).

7.2 | Non-normal distributions

The non-normal distributions included in our study offer a suitable account of the diverse performance of the tests under plausible scenarios, but no study can include all of the non-normal distributions that can be found in empirical data (see Cain et al., 2017). For instance, some variables are subject to order restrictions such that $X_i < Y_i$ for every paired observation i in the sample (e.g., years of marriage and age), a feature that incorporates irrelevant structural components into the magnitude of sample correlations (see García-Pérez & Núñez-Antón, 2013). Our results show that non-normality per se is not necessarily a threat to the accuracy of tests for equality of dependent correlations with overlapping variables (compare Figures 10 and 11), although it seems impossible to determine in advance whether the tests will be robust to the particular (and unknown) form of non-normality of the distributions of any data on hand.

One might think of checking for form of distribution to decide whether any of these tests is safe to use under the circumstances, perhaps leading to preliminary simulations of accuracy under the particular forms of distribution that have been identified. It should nevertheless be noted that such an identification is impossible in strict sense, as one can only test (and then simply reject or fail to reject) a null hypothesis about some particular form of distribution. For instance, with typical small samples, normality may often be incorrectly not rejected when the data actually come from contaminated normal distributions such as those considered here (see Table 1). And this preliminary testing for normality also brings its own inflation of Type I error rates due to the conditional approach (see García-Pérez, 2012).

One could also think of using alternative tests that are robust to violations of normality, none of which have been explored in this paper for reasons discussed next. Wilcox and Tian (2008) considered six such alternative tests in a comparison of accuracy and power with those of the standard Williams test under normality and non-normality (uniform or exponential distributions for all variables). They discarded four of the tests (designated B1, B2, B3, and B4) for various reasons, and their simulations (which were somewhat limited in scope) showed that the two other tests (designated D1 and D2) slightly outperformed the standard Williams test when data were drawn from non-normal distributions, but only by the criterion that Type I error rates do not exceed the nominal *a* level. Yet, neither of these two tests fared well in comparison to the standard Williams test when data were instead drawn from normal distributions. Given the difficulty of ascertaining whether the data on hand come or do not come from normal distributions, using these alternative tests cannot be recommended in general. Alternative tests do not seem to be available that are robust to violations of normality and remain accurate under normality.

Another option is to address the hypothesis of equality of dependent correlations with overlapping variables via replacement of Pearson's correlation with an alternative correlation coefficient for which tests have been developed that are robust to violation of normality. Some such tests have been developed with mixed results in a narrow set of simulation conditions (see, for example, Wilcox, 2016). Nevertheless, it should be noted that the statistical analysis of data must be guided by the research question that a study addresses; replacing Pearson's correlation with an alternative coefficient may not then be appropriate to address the research question.

We must also stress that our study only included data generated to have specific univariate distributions whose higher-order moments are naturally predetermined. Alternatively, non-normal data can be generated to have predefined higher-order moments (skewness and kurtosis) without specification of the form of the distribution. This may seem to capture more realistically the features of psychologically relevant data distributions (see, for example, Cain et al., 2017), which immensely broadens the set of scenarios for an analysis of the robustness of the statistical tests of equality of dependent correlations. In addition, Fairchild et al. (2024) have shown that different algorithms that nominally generate non-normal data with identical levels of skewness and kurtosis produce distributions with distinctly different univariate shapes, with potential implications for the resulting robustness of the tests under study. In consequence, the generalizability of the results of any study on robustness to violations of normality is always uncertain.

Another aspect that must be kept in mind when using these tests in practical applications with (potentially) non-normally distributed variables is that the mere possibility that the null is true is subject to serious structural threats. Consider the hypothetical case in which X_1 has a Lognormal(0, 1) distribution, X_2 has a Beta(2, 5) distribution, and Y has a chi-square distribution with 10 degrees of freedom. In these circumstances, Xiang (2019) showed that the ranges of possible correlations vary greatly across pairs of variables: $\varrho_{1Y} \subset (-.640, .862)$, $\varrho_{2Y} \subset (-.928, .996)$, and $\varrho_{12} \subset (-.646, .815)$. Thus, structurally, there is a very limited range of true nulls on consideration that positive-definiteness is additionally imposed on the triplets of intercorrelations. There may actually be occasions when ϱ_{1Y} and ϱ_{2Y} will both be near their maximum (or minimum) values and, yet, their magnitude will differ. General recommendations may not be given for handling these situations in practice although, in a sense, the utility of a correlation lies in its practical implications and not so much in the extent to which it is free of a structural consequence of the forms of distribution of the variables of concern.

7.3 | Testing for equality of strength of correlation, regardless of sign

What matters in a number of practical situations is whether the *unsigned strength* of the association of two alternative variables with a target variable differs. For instance, for predictive purposes, the sign of correlation is irrelevant and the relative merits of two candidate predictors depend on whether the absolute value of their correlations with the target variable differ. This entails the null hypothesis H_0: $|\rho_{1Y}| = |\rho_{2Y}|$, for which a test does not seem to have been developed. Yet, this null hypothesis branches out into two tractable nulls, namely, H_0: $\rho_{1Y} = \rho_{2Y}$ (for testing that both correlations are equal in value and sign, which is the case considered thus far in this paper) and H_0: $\rho_{1Y} = -\rho_{2Y}$ (for testing that both correlations differ only in sign). As discussed by Boyer et al. (1983) in cases in which r_{1Y} and r_{2Y} differ in sign, this difference in sign can be circumvented by testing instead the null H_0: $\rho_{1Y} = \rho_{-2Y}$, where ρ_{-2Y} stands for the correlation between $-X_2$ and Y. Thus, this test of equality of unsigned strength of correlation is applied identically to the previous one but using instead the correlation between X_1 and $-X_2$ and the correlation between $-X_2$ and Y. An analysis of Type I error rates and power analogous to that presented above was conducted for the same 10 tests considered in this paper and the results were identical. For illustration, Figure 12 shows accuracy results under trivariate normality, which are analogous to those in Figure 6. The only natural difference is that results for negative and positive values of ρ_{12} are swapped because true nulls now lie on the off-diagonal in the panels of Figure 1, where $\rho_{1Y} = -\rho_{2Y}$.

In the absence of a proper test of the null H_0: $|\varrho_{1Y}| = |\varrho_{2Y}|$, it should be stressed that the choice between testing H_0: $\varrho_{1Y} = \varrho_{2Y}$ and testing H_0: $\varrho_{1Y} = -\varrho_{2Y}$ must be based (and justified) on a priori theoretical or practical considerations. The result of using one or the other will generally differ, a natural consequence of the fact that each of the nulls captures a different reality. For instance, consider a situation in which $r_{1Y} = -.22$, $r_{2Y} = .17$, and $r_{12} = -.35$ with a sample size of 100. Use of the standard Williams test for H_0: $\varrho_{1Y} = \varrho_{2Y}$ yields $t_W = -2.4077$ with $p = .018$; in apparent contrast, use of the same test for H_0: $\varrho_{1Y} = -\varrho_{2Y}$ yields $t_W = 0.4437$ with $p = .658$. Although the test comes out significant in the former case and not in the latter, one has to realize that the first test rejects the null that the two correlations have the same magnitude and sign whereas the second test does not reject the null that the two correlations have the same magnitude regardless of sign. It is incumbent on the researcher to decide which of the two tests captures the statistical aspects of the research question that motivated the comparison of correlations. Nevertheless, a pattern of sample correlations similar to that in the preceding example (namely, r_{1Y} and r_{2Y} with different signs along with negative and relatively large r_{12}) is suggestive of a negative relation between X_1 and X_2 that naturally results in correlations between X_1 and Y and between X_2 and Y that differ in sign. In contrast, a different pattern in which

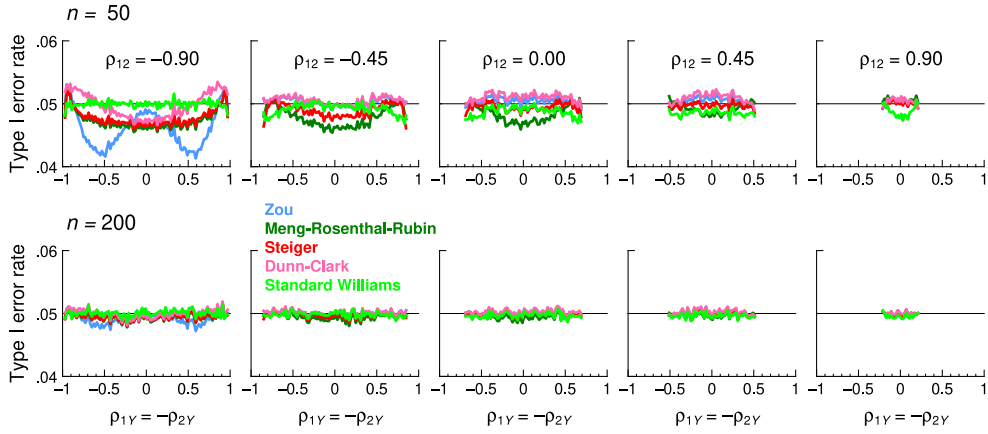

FIGURE 12 Type I error rates of accurate tests (see legend) as a function the actual value of $\rho_{1Y} = -\rho_{2Y}$ (horizontal axis) for selected values of ρ_{12} (columns). Sample sizes are $n = 50$ (top row) and $n = 200$ (bottom row). The nominal $\alpha = .05$ level is indicated by a thin horizontal line (Colour is available online).

r_{12} is relatively large and positive will naturally result in correlations between X_1 and Y and between X_2 and Y that have the same sign.

7.4 | Extensions

A natural extension of the study reported in this paper consists of conducting analogous analyses under the remaining scenarios described in the Introduction, namely, equality of independent correlations and equality of dependent correlations with non-overlapping variables. Indeed, there are a number of alternative tests for each of these scenarios (see Diedenhofen & Musch, 2015) but a study along those lines is beyond the scope of the present paper.

7.5 | Practical recommendations

Our results advise against the use of Pearson–Filon, Olkin, Hotelling, Hendrickson–Stanley–Hills, and Hittner–May–Silver tests of equality of dependent correlations with overlapping variables. The five remaining tests (standard Williams, Dunn–Clark, Steiger, Meng–Rosenthal–Rubin, and Zou) are all dependable and virtually identical as regards their accuracy, power, and applicability under normality, but they are robust to violation of the normality assumption only under some forms of non-normality that are impossible to identify for actual data.

Needless to say, the core recommendation is to compare dependent correlations with overlapping variables by using one of these five dependable tests. Given their uncertain robustness, use of these tests in practical applications should perhaps be accompanied by tests of normality for each of the variables of concern, not with the purpose of validating the tests of equality of correlations themselves but with the purpose of modulating their interpretation. Specifically, the statistical validity of the results of one of the above-mentioned tests of equality of dependent correlations with overlapping variables is less questionable if compatibility with normal distributions is not rejected by normality tests within their own limitations of accuracy and power.

The simple practice of comparing correlations by eye or via the p-values of isolated tests of significance with respect to zero should also be avoided. The prevalence of such practices is hard to investigate because of the absence of an identifiable characteristic of the papers reporting them that could lead to database searches. Yet, reporting practices involving proper use of tests is easier to investigate. The Introduction mentioned a search for papers that cited *cocor* (Diedenhofen & Musch, 2015) and reported the use of tests of equality of dependent correlations, with the results listed in Table A1 in Appendix A. Note that 10 of the papers did not report which test had been used, one reported that all tests gave the same result, and seven reported using the defective Pearson–Filon or Hittner–May–Silver tests. Only nine papers reported using the dependable Dunn–Clark, Steiger, or Meng–Rosenthal–Rubin tests. In the interest of transparency and best reporting practices, the particular test that was chosen in a given study should be reported and the choice should be among the tests that have proven dependable under normality, with a preference for the standard Williams test over all others.

AUTHOR CONTRIBUTIONS

Miguel A. García-Pérez: Conceptualization; methodology; software; data curation; investigation; validation; formal analysis; supervision; resources; project administration; visualization; writing – review and editing; funding acquisition; writing – original draft.

ACKNOWLEDGEMENTS

This research was supported by grant PID2019-11083GB-I00 from the Ministerio de Ciencia e Innovación.

DATA AVAILABILITY STATEMENT

FORTRAN source code and output files for the accuracy and power studies reported in this paper are available at https://osf.io/4pw7r/. The source code calls NAG Library Mark 19 subroutines and it will not run if compiled without access to them. The author did not use generative AI or any other similar tools at any step of this research or for any purpose.

CONFLICTS OF INTEREST STATEMENT

The author declares no conflicts of interest.

ORCID

Miguel A. García-Pérez https://orcid.org/0000-0003-2669-4429

REFERENCES

Bilker, W. B., Brensinger, C., & Gur, R. C. (2004). A two factor ANOVA-like test for correlated correlations: CORANOVA. *Multivariate Behavioral Research*, *39*, 565–594. https://doi.org/10.1207/s15327906mbr3904_1

Blake, A., & Palmisano, S. (2021). Divergent thinking influences the perception of ambiguous visual illusions. *Perception*, *50*, 418–437. https://doi.org/10.1177/03010066211000192

Boyer, J. E., Palachek, A. D., & Schucany, W. R. (1983). An empirical study of related correlation coefficients. *Journal of Educational Statistics*, *8*, 75–86. https://doi.org/10.3102/10769986008001075

Brosnan, M., Gallop, V., Iftikhar, N., & Keogh, E. (2011). Digit ratio (2D:4D), academic performance in computer science and computer-related anxiety. *Personality and Individual Differences*, *51*, 371–375. https://doi.org/10.1016/j.paid.2010.07.009

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*, 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

Choi, S. C. (1977). Tests of equality of dependent correlation coefficients. *Biometrika*, *64*, 645–647. https://doi.org/10.1093/biomet/64.3.645

Cohen, A. (1989). Comparison of correlated correlations. *Statistics in Medicine*, *8*, 1485–1495. https://doi.org/10.1002/sim.4780081208

Ćoso, B., Guasch, M., Bogunović, I., Ferré, P., & Hinojosa, J. A. (2023). CROWD-5e: A Croatian psycholinguistic database of affective norms for five discrete emotions. *Behavior Research Methods*, *55*, 4018–4034. https://doi.org/10.3758/s13428-022-02003-2

De Veaux, D. (1976). *Tight upper and lower bounds for correlation of bivariate distribution arising in air pollution modeling.* Technical report No. 5. Study on statistics and environmental factors in health. Department of Statistics, Stanford University. https://doi.org/10.2172/7124713

Del Giudice, M., & Angeleri, R. (2016). Digit ratio (2D:4D) and attachment styles in middle childhood: Indirect evidence for an organizational effect of sex hormones. *Adaptive Human Behavior and Physiology*, *2*, 1–10. https://doi.org/10.1007/s40750-015-0027-3

Diedenhofen, B., & Musch, J. (2015). `Cocor`: A comprehensive solution for the statistical comparison of correlations. *PLoS One*, *10*, e0121945. https://doi.org/10.1371/journal.pone.0121945

Dukic, V. M., & Marić, N. (2013). Minimum correlation in construction of multivariate distributions. *Physical Review E*, *87*, 032114. https://doi.org/10.1103/PhysRevE.87.032114

Dunn, O. J., & Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, *64*, 366–377. https://doi.org/10.1080/01621459.1969.10500981

Dunn, O. J., & Clark, V. (1971). Comparison of tests of the equality of dependent correlation coefficients. *Journal of the American Statistical Association*, *66*, 904–908. https://doi.org/10.1080/01621459.1971.10482369

Fairchild, A. J., Yin, Y., Baraldi, A. N., Astivia, O. L. O., & Shi, D. (2024). Many nonnormalities, one simulation: Do different data generation algorithms affect study results? *Behavior Research Methods*. https://doi.org/10.3758/s13428-024-02364-w

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fink, B., Manning, J. T., & Neave, N. (2004). Second to fourth digit ratio and the 'big five' personality factors. *Personality and Individual Differences*, *37*, 495–503. https://doi.org/10.1016/j.paid.2003.09.018

García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology*, *3*, 325. https://doi.org/10.3389/fpsyg.2012.00325

García-Pérez, M. A. (2023). Use and misuse of corrections for multiple testing. *Methods in Psychology*, *8*, 100120. https://doi.org/10.1016/j.metip.2023.100120

García-Pérez, M. A., & Núñez-Antón, V. (2013). Correlation between variables subject to an order restriction, with application to scientometric indices. *Journal of Informetrics*, *7*, 542–554. https://doi.org/10.1016/j.joi.2013.01.010

Guo, K., Hare, A., & Liu, C. H. (2022). Impact of face masks and viewers' anxiety on ratings of first impressions from faces. *Perception*, *51*, 37–50. https://doi.org/10.1177/03010066211065230

Hendrickson, G. F., & Collins, J. R. (1970). Note correcting the results in 'Olkin's new formula for the significance of r_{13} vs. r_{23} compared with Hotelling's method'. *American Educational Research Journal*, *7*, 639–641. https://doi.org/10.3102/00028312007004639

Hendrickson, G. F., Stanley, J. C., & Hills, J. R. (1970). Olkin's new formula for significance of r_{13} vs. r_{23} compared with Hotelling's method. *American Educational Research Journal*, *7*, 189–195. https://doi.org/10.3102/00028312007002189

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *Journal of General Psychology*, *130*, 149–168. https://doi.org/10.1080/00221300309601282

Hotelling, H. (1940). The selection of variates for use in prediction with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics*, *11*, 271–283. https://doi.org/10.1214/aoms/1177731867

Maier, M., & Lakens, D. (2022). Justify your alpha: A primer on two practical approaches. *Advances in Methods and Practices in Psychological Science*, *5*, 25152459221080396. https://doi.org/10.1177/25152459221080396

May, K., & Hittner, J. B. (1997a). Tests for comparing dependent correlations revisited: A Monte Carlo study. *Journal of Experimental Education*, *65*, 257–269. https://doi.org/10.1080/00220973.1997.9943458

May, K., & Hittner, J. B. (1997b). A note on statistics for comparing dependent correlations. *Psychological Reports*, *80*, 475–480. https://doi.org/10.2466/pr0.1997.80.2.475

Menculini, G., Gentili, L., Gaetani, L., Mancini, A., Sperandei, S., di Sabatino, E., Chipi, E., Salvadori, N., Tortorella, A., Parnetti, L., & di Filippo, M. (2023). Clinical correlates of state and trait anxiety in multiple sclerosis. *Multiple Sclerosis and Related Disorders*, *69*, 104431. https://doi.org/10.1016/j.msard.2022.104431

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172–175. https://doi.org/10.1037/0033-2909.111.1.172

Neill, J. J., & Dunn, O. J. (1975). Equality of dependent correlation coefficients. *Biometrics*, *31*, 531–543. https://doi.org/10.2307/2529435

Olkin, I. (1967). Correlations revisited. In J. C. Stanley (Ed.), *Improving experimental design and statistical analysis* (pp. 102–128). Rand McNally.

Olkin, I., & Finn, J. D. (1990). Testing correlated correlations. *Psychological Bulletin*, *108*, 330–333. https://doi.org/10.1037/0033-2909.108.2.330

Pearson, K., & Filon, L. N. G. (1898). Mathematical contributions to the theory of evolution. IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London, Series A*, *191*, 229–311. https://doi.org/10.1098/rsta.1898.0007

Rubin, M. (2021). When to adjust alpha during multiple testing: A consideration of disjunction, conjunction, and individual testing. *Synthese*, *199*, 10969–11000. https://doi.org/10.1007/s11229-021-03276-4

Rubin, M. (2024). Inconsistent multiple testing corrections: The fallacy of using family-based error rates to make inferences about individual hypotheses. *Methods in Psychology*, *10*, 100140. https://doi.org/10.1016/j.metip.2024.100140

Silver, N. C., Hittner, J. B., & May, K. (2006). A FORTRAN 77 program for comparing dependent correlations. *Applied Psychological Measurement*, *30*, 152–153. https://doi.org/10.1177/0146621605277132

Silver, N. C., & Merino-Soto, C. (2016). Computer programs for comparing dependent correlations. *Revista Digital de Investigación en Docencia Universitaria*, *10*, 72–82. https://doi.org/10.19083/ridu.10.495

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Thorson, J. A., & Powell, F. C. (1993). Personality, death anxiety, and gender. *Bulletin of the Psychonomic Society*, *31*, 589–590. https://doi.org/10.3758/BF03337363

Uzun, G. B., & Tok, Y. (2023). Investigating the correlation between 2D:4D finger digit ratios and attention gathering skills of 60–72 month-old children. *Early Human Development*, *176*, 105712. https://doi.org/10.1016/j.earlhumdev.2023.105712

Wilcox, R. R. (2016). Comparing dependent robust correlations. *British Journal of Mathematical and Statistical Psychology*, *69*, 215–224. https://doi.org/10.1111/bmsp.12069

Wilcox, R. R., & Tian, T. (2008). Comparing dependent correlations. *Journal of General Psychology*, *135*, 105–112. https://doi.org/10.3200/GENP.135.1.105-112

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, Series B*, *21*, 396–399. https://doi.org/10.1111/j.2517-6161.1959.tb00346.x

Xiang, J. X. (2019). Estimation of minimum and maximum correlation coefficients. *Statistics & Probability Letters*, *145*, 81–88. https://doi.org/10.1016/j.spl.2018.08.010

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399–413. https://doi.org/10.1037/1082-989X.12.4.399

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: García-Pérez, M. A. (2024). Are alternative variables in a set differently associated with a target variable? Statistical tests and practical advice for dealing with dependent correlations. *British Journal of Mathematical and Statistical Psychology*, *00*, 1–29. https://doi.org/10.1111/bmsp.12354

APPENDIX A

Statistical test reportedly used in papers citing Diedenhofen and Musch (2015)

TABLE A1 Statistical tests used in papers that cite *cocor* (Diedenhofen & Musch, 2015) as the tool to test for equality of dependent correlations with overlapping variables.

Paper doi	Test used
https://doi.org/10.1027/1015-5759/a000682	Not reported
https://doi.org/10.1002/jclp.23624	Not reported
https://doi.org/10.1080/1359432X.2023.2193692	Not reported
https://doi.org/10.1037/aca0000539	Not reported
https://doi.org/10.1177/01461672231210465	Not reported
https://doi.org/10.1007/s00426-022-01779-4	Not reported
https://doi.org/10.1037/pas0001249	Not reported
https://doi.org/10.1016/j.jenvp.2022.101919	Not reported
https://doi.org/10.1007/s12144-021-01722-7	Not reported
https://doi.org/10.1186/s40359-023-01111-8	Not reported
https://doi.org/10.1037/emo0001013	All tests
https://doi.org/10.1177/09567976231185127	Pearson–Filon
https://doi.org/10.1027/1618-3169/a000586	Pearson–Filon
https://doi.org/10.1002/ejsp.3019	Hittner–May–Silver
https://doi.org/10.1027/2698-1866/a000037	Hittner–May–Silver
https://doi.org/10.1027/1015-5759/a000717	Hittner–May–Silver
https://doi.org/10.1002/casp.2695	Hittner–May–Silver
https://doi.org/10.1525/collabra.88330	Hittner–May–Silver
https://doi.org/10.1016/j.cogpsych.2023.101550	Dunn–Clark
https://doi.org/10.1037/emo0001229	Steiger
https://doi.org/10.1007/s10869-023-09893-9	Steiger
https://doi.org/10.1007/s12144-023-05183-y	Steiger
https://doi.org/10.3389/fpsyg.2023.1165911	Steiger
https://doi.org/10.1037/aca0000573	Steiger
https://doi.org/10.1080/08870446.2023.2196994	Meng–Rosenthal–Rubin
https://doi.org/10.1186/s40359-023-01463-1	Meng–Rosenthal–Rubin
https://doi.org/10.1080/02134748.2023.2171560	Meng–Rosenthal–Rubin

APPENDIX B

Effects of nonlinear transformations on intercorrelations

This appendix analyses the consequences of nonlinear transformations for the intercorrelations among variables that originally have a trivariate standard normal distribution with correlation matrix

$$P = \begin{pmatrix} 1 & \rho_{X_1X_2} & \rho_{X_1X_3} \\ \rho_{X_1X_2} & 1 & \rho_{X_2X_3} \\ \rho_{X_1X_3} & \rho_{X_2X_3} & 1 \end{pmatrix}.$$

Let

$$\widetilde{X}_i = f_i\left(X_i\right), \quad \text{for} i \in \{1,2,3\}, \tag{B1}$$

where f_i is an arbitrary nonlinear function such that the distribution of $\widetilde{X}_i$ becomes non-normal. For notational convenience, we refer to the third variable here as X_3 instead of Y on the understanding that $X_3 := Y$. We are interested in finding out what the correlation matrix

$$\widetilde{P} = \begin{pmatrix} 1 & \rho_{\widetilde{X}_1\widetilde{X}_2} & \rho_{\widetilde{X}_1\widetilde{X}_3} \\ \rho_{\widetilde{X}_1\widetilde{X}_2} & 1 & \rho_{\widetilde{X}_2\widetilde{X}_3} \\ \rho_{\widetilde{X}_1\widetilde{X}_3} & \rho_{\widetilde{X}_2\widetilde{X}_3} & 1 \end{pmatrix}$$

is for the nonlinearly transformed variables. A general answer to this question does not exist, but some cases have been analysed that lend themselves to finding out what is the minimum correlation attainable for non-normal variables with arbitrary univariate distributions (e.g., Dukic & Marić, 2013; see also Xiang, 2019). For the purpose of this paper, it is useful to start dissecting a sample case that can be easily worked out analytically and where the consequences of nonlinear transformations are readily apparent. Let $\widetilde{X}_i = f_i\left(X_i\right) = \exp\left(X_i\right)$ in Equation (B1) so that each of the original normal variables is transformed into one that has a lognormal distribution. This transformation is mathematically tractable and allows us to obtain.

$$\rho_{\widetilde{X}_i\widetilde{X}_j} = \frac{\exp\left(\rho_{X_iX_j}\right) - 1}{\exp(1) - 1}, \quad \text{for } i,j \in \{1,2,3\} \tag{B2}$$

(De Veaux, 1976). The red curve in the panels of Figure B1 displays this relation, where a nonlinear transformation of correlations is immediately apparent along with the fact that the minimum correlation between exponentially transformed normal variables is $-\exp(-1) \approx -.3679$ (De Veaux, 1976). To further illustrate this effect on correlations, we generated a sample of $n = 10{,}000$ triplets (X_1, X_2, X_3) from a trivariate normal distribution with the applicable correlation matrix under each of the conditions involved in the accuracy study reported in this paper, transformed each triplet (X_1, X_2, X_3) into $\left(\widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3\right) = \left(\exp\left(X_1\right), \exp\left(X_2\right), \exp\left(X_3\right)\right)$, and computed the sample intercorrelations among the original and among the transformed variables. The sample pairwise correlations obtained for the transformed variables are plotted as the ordinate of symbols in the panels of Figure B1, each at the abscissa pertaining to the corresponding pairwise correlation between the originating normal variables. The symbols naturally pack along the theoretical relation in Equation (B2) to within sampling error.

The consequence on estimates of, for example, empirical Type I error rates under violation of trivariate normality when non-normal data are generated in this way is immediately apparent. Results obtained from trivariate normal data in the condition, say, $\rho_{1Y} = \rho_{2Y} = .8$ with $\rho_{12} = .75$ belong, by Equation (B2), to the condition $\rho_{1Y} = \rho_{2Y} = .71$ with $\rho_{12} = .65$ when the original data are transformed by Equation (B1)

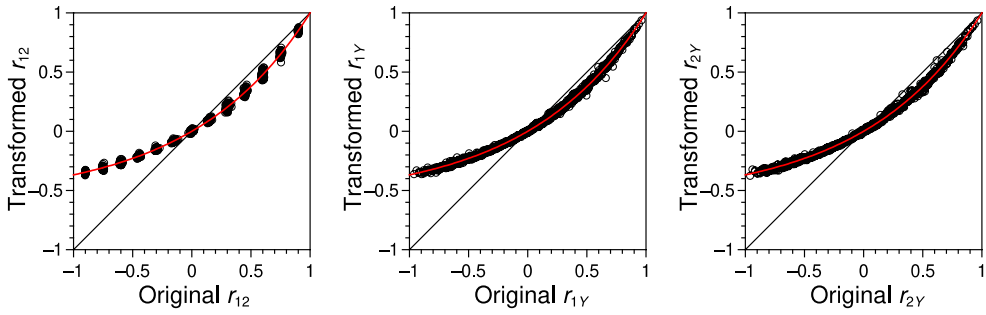

FIGURE B1 Relation between original correlations of normally distributed observations (abscissa) and correlations for their exponentially transformed counterparts (ordinate). Each symbol comes from a different sample of 10,000 triplets with original intercorrelations over the full range of conditions involved in our simulation studies. The red curve in each panel is the theoretical relation in Equation (B2). A diagonal identity line is displayed for reference. (Colour is available online).

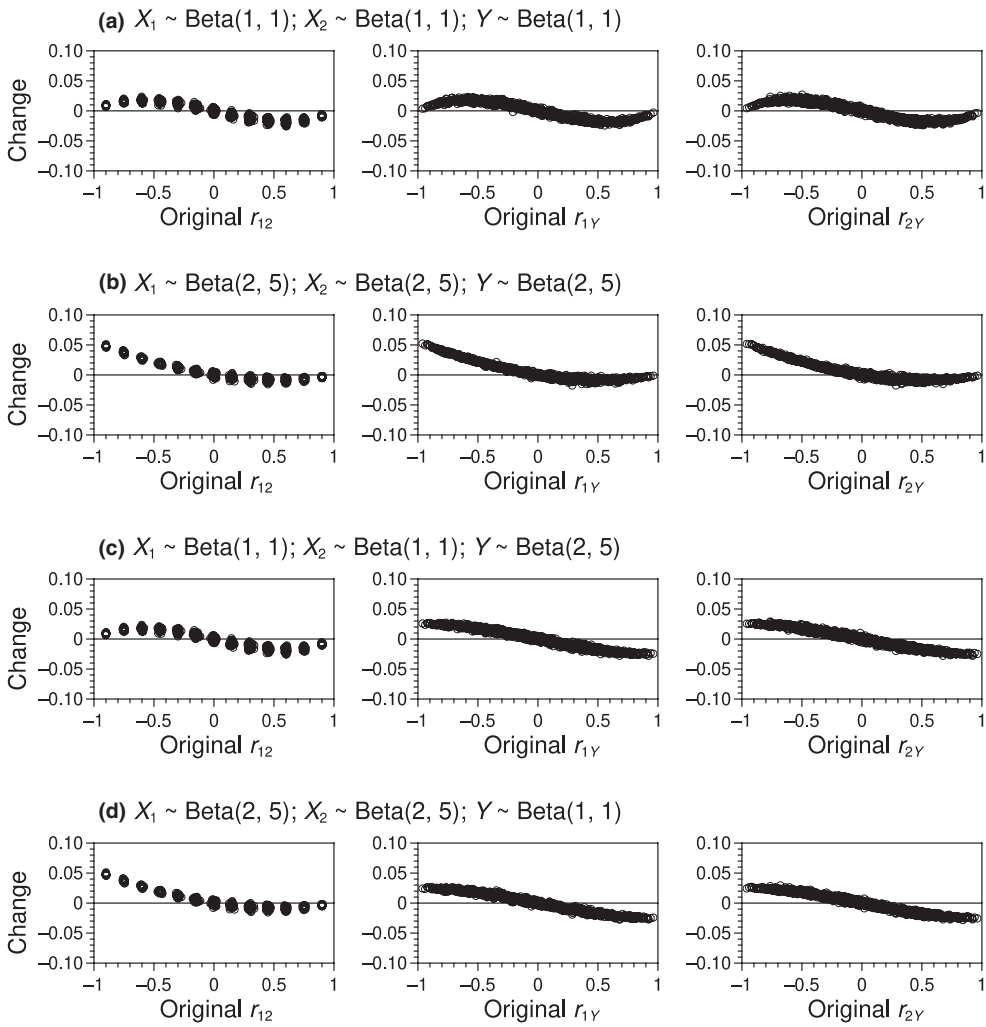

FIGURE B2 Signed change in correlation (ordinate in each panel) produced when trivariate normal variables are differently transformed (rows) to produce variables with univariate beta distributions, as a function of the original value of the correlations (abscissa in each panel). Each symbol comes from a different sample of 10,000 triplets with original intercorrelations over the full range of conditions involved in our simulation studies. A horizontal line at an ordinate of zero is displayed for reference.

to have a lognormal distribution. There is nothing essentially incorrect in this approach, but interpretation of the results must keep in mind that conditions involving different forms of distribution may also bring differences in the values of population correlations to which they apply. Yet, this was only an illustrative example and we are actually interested in the consequences of the transformations used in this study to generate beta distributions.

A closed-form expression analogous to Equation (B2) does not exist to relate the intercorrelations among trivariate normal variables to those of variables transformed to have beta distributions. Yet, the simulation approach illustrated in Figure B1 can be used to investigate the shape of this relation for all of the cases involved in our simulations, which cover different combinations of beta variables. Plots of the results in the form of Figure B1 showed that symbols fall virtually on the diagonal identity line, which hides the minor effects that these transformations have on the resultant intercorrelations. For this reason, the ordinate of the panels in Figure B2 displays the signed difference between transformed and original correlations so that the horizontal line at an ordinate of zero indicates no change in correlation after the transformations. Although all combinations of transformations alter the original correlations, the change is certainly very small. In addition, r_{1Y} and r_{2Y} remain identical (to within sampling error) as long as X_1 and X_2 are identically transformed. This warrants a direct comparison of the performance of the tests with and without violation of normality under otherwise nearly identical conditions.