frontiers | Frontiers in Psychology

BRIEF RESEARCH REPORT
published: 26 May 2022
doi: 10.3389/fpsyg.2022.860213

Testing the Magnitude of Correlations Across Experimental Conditions

*Simone Di Plinio**

Department of Neuroscience Imaging and Clinical Sciences, "G. D'Annunzio" University of Chieti-Pescara, Chieti, Italy

Correlation coefficients are often compared to investigate data across multiple research fields, as they allow investigators to determine different degrees of correlation to independent variables. Even with adequate sample size, such differences may be minor but still scientifically relevant. To date, although much effort has gone into developing methods for estimating differences across correlation coefficients, adequate tools for variable sample sizes and correlational strengths have yet to be tested. The present study evaluated four different methods for detecting the difference between two correlations and tested the adequacy of each method using simulations with multiple data structures. The methods tested were Cohen's q, Fisher's method, linear mixed-effects models (LMEM), and an *ad hoc* developed procedure that integrates bootstrap and effect size estimation. Correlation strengths and sample size was varied across a wide range of simulations to test the power of the methods to reject the null hypothesis (i.e., the two correlations are equal). Results showed that Fisher's method and the LMEM failed to reject the null hypothesis even in the presence of relevant differences between correlations and that Cohen's method was not sensitive to the data structure. Bootstrap followed by effect size estimation resulted in a fair, unbiased compromise for estimating quantitative differences between statistical associations and producing outputs that could be easily compared across studies. This unbiased method is easily implementable in MatLab through the bootes function, which was made available online by the author at MathWorks.

Keywords: correlation, bootstrap, effect size, *p*-value, mixed-effects, sample size, fisher, Cohen

OPEN ACCESS

Edited by:
Laura Badenes-Ribera,
University of Valencia, Spain

Reviewed by:
Angeliki Papana,
University of Macedonia, Greece
Holly Patricia O'Rourke,
Arizona State University, United States

****Correspondence:***
Simone Di Plinio
simone.diplinio@unich.it

Specialty section:
This article was submitted to Quantitative Psychology and Measurement, a section of the journal Frontiers in Psychology

Received: *22 January 2022*
Accepted: *25 March 2022*
Published: *26 May 2022*

Citation:
Di Plinio S (2022) Testing the Magnitude of Correlations Across Experimental Conditions. Front. Psychol. 13:860213. doi: 10.3389/fpsyg.2022.860213

INTRODUCTION

Comparing statistics is a frequent point of contention among researchers. The need to compare correlations is common and requires a specific assay to determine whether a continuous variable, often called covariate, has a different degree of correlation between two sets of data. Examples of fields in need of such an assay include cognitive psychology (e.g., the correlation between the degree of task automation and behavioral performance in extroverted vs. introverted individuals), social psychology (e.g., the correlation between social exclusion and job satisfaction in men vs. women), and cognitive neuroscience (e.g., the correlation between brain activity and behavioral performance under negative vs. positive emotional stimulation). Calculated differences may be minor, even with the recommended sample size, but can still be scientifically relevant (Ellis, 2010). Furthermore, modern science is gradually moving away from p-centric data interpretation toward effect-size-oriented approaches (Kelley and Preacher, 2012; Sullivan and Feinn, 2012). Thus, reporting only

r- and p-values and binarizing result interpretations as either significant or non-significant, depending on an (eventually corrected) threshold of $p < 0.05$, has become outdated (Nichols et al., 2017; Ioannidis, 2019).

The existence of a statistical association cannot be relied upon to evaluate whether the strength of the relation between two variables will always be the same. In a within-group design, the correlation coefficients between the outcome (DV) and the covariate (IV) may be decreased in a specific experimental condition B than in another condition A. Alternatively, in a between-group design, the treatment group may show a weaker correlation between the outcome and the covariate than in the control group, or vice versa. These context-dependent or group-dependent effects on the extent of the correlation between two variables should be investigated using the most suited methods so that scientists from different disciplines can assess the most fitting comparison between the two correlations. Of note, given that null hypotheses are always false when evaluated with large datasets (Cohen, 1990), and that both small samples and large samples can convey useful information (Lindquist et al., 2012; Friston, 2013), the effect of a correlational change should be investigated not only with various correlation values, but also various sample sizes. However, even though much effort has been invested into developing methodologies that can estimate differences across correlation coefficients, studies investigating this problem using a comprehensive approach and variable sample sizes have yet to be published.

Several strategies have been developed to estimate differences between correlations. The simplest method was proposed by Cohen (1988) and estimated an effect size as the difference between two Fisher-transformed correlations. Fisher's method (Fisher, 1921) also accounts for sample size and calculates the probability that two correlations will differ given their strength and the number of samples in the two groups. Whereas Cohen's and Fisher's methods rely exclusively on r-values, ignoring the initial data structure, analysis of covariance (ANCOVA) and linear mixed-effects models (LMEM) retain this information. ANCOVA and LMEMs have been widely adopted for analyzing data in cognitive neuroscience experiments, wherein the parameters observed are affected by multiple factors (Buckner et al., 2008; Garrett et al., 2010; Zilles and Amunts, 2013).

A bootstrap approach followed by calculation of the effect size may also be used to detect changes in correlations between neurophysiological parameters and behavioral performance across experimental conditions (Di Plinio et al., 2018). This method allowed testing the hypothesis that the association between functional connectivity across brain regions and behavioral performance (Hampson et al., 2010) was weakened by negative emotional stimulation. This approach (bootstrap and effect size estimation) is particularly advantageous since it is less impaired by violating normality assumptions (Liu and Popmey, 2020). Finally, structural equation modeling has also been proposed for testing independent or dependent correlational hypotheses (Cheung and Chan, 2004). This method is grounded in confirmatory factor analysis and is useful when data includes both dependent and independent measures. However, this approach may be ill-suited to small sample sizes and is more appropriate for meta-analytic designs (Cheung and Chan, 2005; Cheung, 2014).

This paper examines four different approaches to evaluate their power to detect an effect. The term effect in this study refers to "a change in the correlation between two conditions or two groups." Although p-values are undergoing a theoretical revision by the scientific community, they still provide a universally recognized statistic (Goodman, 2019; Greenland, 2019). As such, both p-values and effect estimates are reported for each method. The methods examined in this study are Cohen's q, Fisher's method, LMEM, and bootstrap with effect size estimation. A series of simulations were implemented to test the four methods in an environment wherein correlational strength and sample size vary from one cycle to the next. Method performances are then discussed.

MATERIALS AND METHODS

Simulation Parameters

In the simulations used in the present study, values of the first correlation coefficient, r_1, occurred in the interval $[-0.5, 0.5]$ in steps of 0.01. For each value of r_1, the second correlation value r_2 occurred in the interval $[(\mathrm{r}_1 - 0.5)\ (\mathrm{r}_1 + 0.5)]$ in steps of 0.01. For each cycle, N samples were simulated for each condition. As such, N will be the number of samples considered. N varied from 10 to 180, in steps of 2. As our focus was on the comparison of repeated measures (within-subject design without missing data), the number of samples for each condition was set to be equal; that is, $N = \mathrm{n}_1 = \mathrm{n}_2$. Correlation strengths and sample sizes were chosen in accordance with cognitive psychology- and neuroscience-like scenarios, but results can be extended to other research fields such as medicine and social psychology. Data were processed using MATLAB 9.2 (The Math Works Inc., Natick, MA). Statistics obtained for each method were averaged across values of r_1 and zero-centered. Each plot shows the average statistics (e.g., p-value, effect size) with respect to increasing values of r_2 and N, and across values of r_1.

For LMEM and bootstrap models, sample data and covariate values were randomly generated and normally distributed, using the MatLab function *randn*. This led to the creation of two distributions for the dependent variable (DV), namely $\mathbf{X_1}$ and $\mathbf{X_2}$, with a mean $\mu_1 = \mu_2 = 0$ and a standard deviation $\sigma_1 = \sigma_2 = 1$. Independent variable (IV) and N covariate values were simulated for each condition and computed to yield the desired value of correlation with the DV. The covariates γ_1 and γ_2 were obtained as follows:

$$\boldsymbol{\gamma_i} = \sigma_\gamma \left(r_i * \boldsymbol{\chi_i} + \sqrt{1 - r_i^2} * \boldsymbol{\psi_i} \right) + \mu_\gamma \tag{1}$$

where μ_γ and σ_γ are the desired mean and standard deviation of the covariate (with $\mu_\gamma = 0$ and $\sigma_\gamma = 1$), respectively, while ψ is a pseudo-random set of N normally distributed values. This procedure was based on the Cholesky decomposition, which is commonly used in Monte-Carlo simulations of multiple correlated variables (Press et al., 1992). To note, since experimental psychology studies the relationship between variables in a sample, the method used is particularly appropriate for the case. In fact, it generates sample-level correlated data, and not population-level correlated data which would be an odd and non-realistic choice for simulations.

Methods for Assessing Correlation Differences

Cohen's q

Cohen proposed a simple method for interpreting the difference between two correlations (Cohen, 1988). Initially, to reduce skewness (asymmetry derived from the definition of r_i in the interval $[-1, 1]$), r-values were transformed to z values *via* the Fisher procedure:

$$z_i = (0.5)\log\left(\frac{1+r_i}{1-r_i}\right) \tag{2}$$

Then, the absolute value of the difference between the two z-values was computed, such that $q = |z_1\text{-}z_2|$. The value q is the estimate of the effect size. The following intervals were proposed by Cohen to interpret these values: $q < 0.1$, *no* effect; $0.1 \leq q < 0.3$, *small* effect; $0.3 \leq q < 0.5$, *medium* effect; $q \geq 0.5$, *large* effect. Since no p-value is associated with Cohen's method, the only statistic reported is the effect size q.

Fisher's Method

Fisher's method (Fisher, 1921) is used to calculate the probability of two correlations being different, given the differences between r-values and the size of the two samples. The null hypothesis is that the correlation between X_1 and γ_1 will be the same as the correlation between X_2 and γ_2 for sample sizes n_1 and n_2. Correlation values r_1 and r_2 are converted to z-values as described in Equation 2. The test statistic t is then calculated:

$$t = \frac{(z_1 - z_2)}{\sqrt{1/\left(n_1 - 3\right) + 1/\left(n_2 - 3\right)}} \tag{3}$$

Finally, using the cumulative distribution function of t in a standard distribution with mean $\mu_t = 0$ and standard deviation $\sigma_t = 1$, the p-value is calculated to assess whether the null hypothesis can be trusted or not. Statistics reported for Fisher's method are the p-value and t-statistic.

Linear Mixed-Effects Model

Linear mixed-effect models (LMEM) were applied in the form [DV $\sim$ IV*condition + (1| subject)] and thus included a fixed effect (the experimental condition), a continuous effect (the covariate IV), and a random intercept at the subject level (1| subject) to account for inter-individual variability. This type of model is applied frequently in psychology and neuroscience. At each cycle, a model was fitted using the MatLab function *fitlme* and the p-value and β statistics for the interaction between the experimental condition and the covariate were extracted.

Among applicable generalized linear models, the choice of the LMEM over, say, ANCOVA is due to the former's increased flexibility and sensitivity (Schneider et al., 2015; Brysbaert and Stevens, 2018). Of note, as the aim of the present work is to investigate the power to predict an *effect*, corrections for multiple comparisons were unnecessary. Furthermore, since within-subject variability at the interaction level (subject: covariate) was not simulated, the introduction of random slopes was not necessary for the purposes of the study.

Bootstrap Method and Effect Size Estimation

The bootstrap method is a resampling technique often used to estimate confidence intervals and allows one to approximate the sampling distribution of a statistic (Efron and Tibshirani, 1986). In this study, we used a univariate, bias-corrected, accelerated bootstrap with replacement (Efron, 1987) to sample the correlation value r_i. A sampling distribution was obtained by resampling the original data k times and obtaining k samples with sizes equal to the starting sample ($N_k = N$); that is, k is the number of bootstrap cycles. A similar sampling approach has been described for correlations (see Olkin and Finn, 1990, 1995); however, the method presented here implements a bias-corrected bootstrap procedure that can accommodate small sample sizes and outliers and includes an effect size estimation.

Each cycle was bootstrapped by estimating the correlation between the DV and covariate for each condition. Individually bootstrapping each correlation allows the estimation of the bootstrapped effect size for each condition, which may be useful for descriptive purposes. After k bootstrap cycles, two distributions of correlations were obtained for each condition and transformed into z-values. Each distribution possessed an associated mean and standard deviation (μ_{z1}, σ_{z1}; μ_{z2}, σ_{z2}). These distributions were then used for analyses. For each cycle (i.e., for each pair of r_2 and N-values), the difference between the two z-distributions was represented as indicated in Equations 4 and 5, with the effect size estimated using Cohen's d (Lipsey and Wilson, 2001; Ellis, 2010):

$$\mathrm{ES}_{12} = d = \frac{\mu_{z1} - \mu_{z2}}{\sigma_{\mathrm{pooled}}} \tag{4}$$

where

$$\sigma_{\mathrm{pooled}} = \sqrt{\frac{\left(n_1 - 1\right)\ \sigma_{z1}^2 + \left(n_2 - 1\right)\sigma_{z2}^2}{n_1 + n_2 - 2}} \tag{5}$$

A Z-test was performed comparing the distribution obtained by subtracting the two bootstrapped correlation distributions against a zero-centered distribution, to prevent biases caused by large sample sizes in estimating a p-value. Effect sizes (ES) can also be interpreted in terms of the percentage of non-overlap of the first group's scores with those of the second group (Cohen, 1988). For example, ESs of 0.0, 0.8, and 1.7 indicate that the distribution of scores for the first group overlapped with the distribution of scores for the second group with 0, 47.45, and 75.4% of non-overlap, respectively.

Extension to Non-normal and Real Data

To generalize findings, the four methods were also applied on non-normally distributed data generated by taking the absolute values of normally distributed data, thus producing right-skewed (positively skewed) distributions. For consistency, the procedure

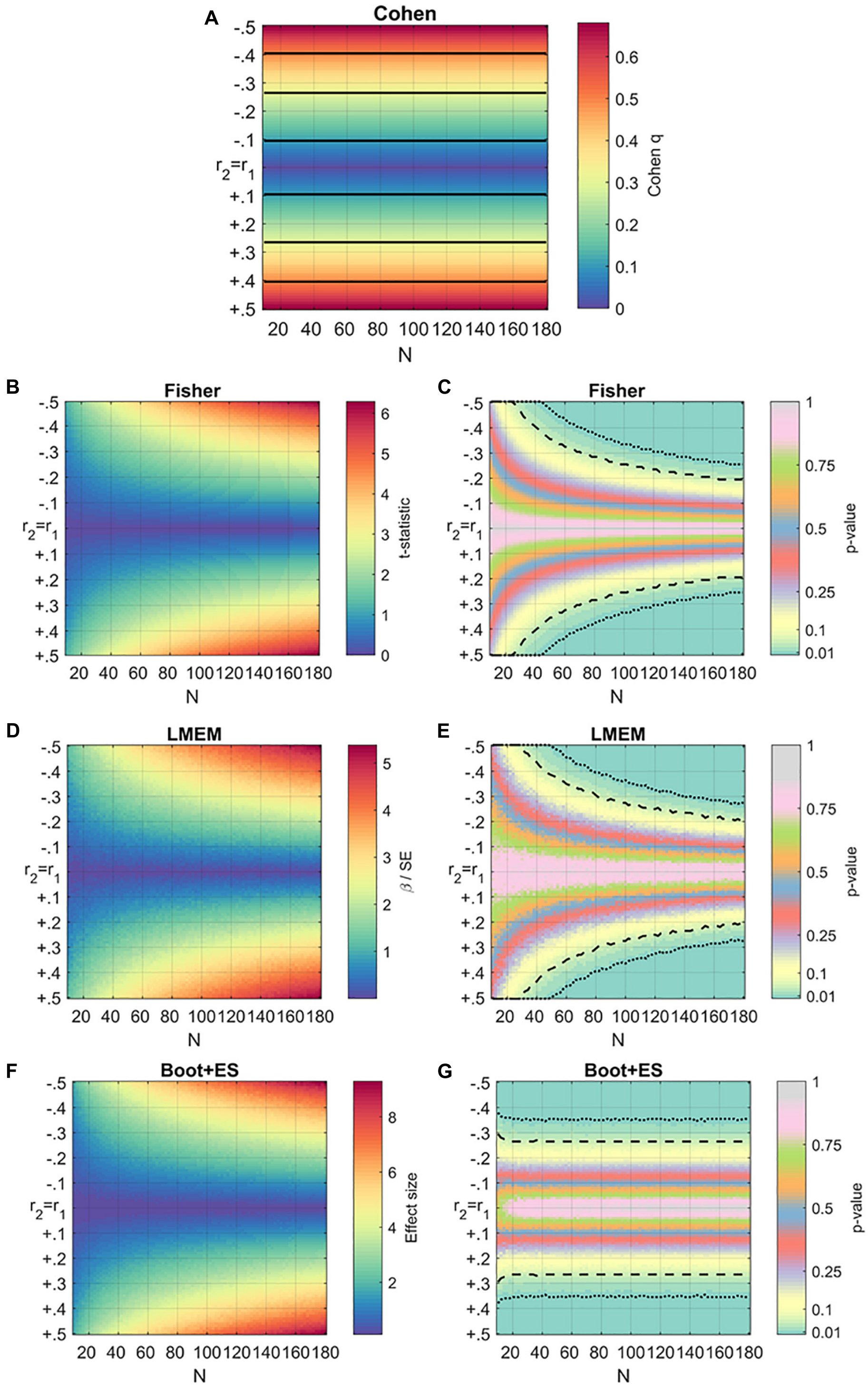

FIGURE 1 | Results of the generalized simulations using the four methods. For all the subfigures, the horizontal axis represents the sample size, while the vertical axis represents values of r2 compared to r1. The third dimension (heat) represents the effect size or the *p*-value of each method **(A)** Results of the simulations using Cohen's q. The effect sizes are averaged across values of r1. Black horizontal lines are used to separate the different levels of effect: small (blue), medium (cyan), large (green). **(B,C)** Results of the simulations using Fisher's method. Statistics are averaged across values of r1. *T*-statistics are reported on the left panel, *p*-values on the right panel. Dashed and dotted lines represent thresholds of $p < 0.05$ and $p < 0.01$, respectively. **(D,E)** Results of the simulations using mixed-effect models (LMEM). Statistics are averaged across values of r1. β values for the interaction condition: covariate divided by their standard errors (SE) are reported on the left panel, corresponding *p*-values are reported on the right panel. **(F,G)** Results of the simulations using bootstrap (Boot + ES). Statistics are averaged across values of r1. Effect sizes (Cohens' d) are reported on the left panel, corresponding *p*-values are reported on the right panel.

adopted for the analyses of non-normally distributed data was the same as the one reported above.

Additionally, the four methods were also applied on behavioral data from the Human Connectome Project (HCP) database.[1] More specifically, the same four methods described above were applied on behavioral data from 100 unrelated subjects which performed a language task (Binder et al., 2011). The task consists of two runs interleaving four blocks of a *story* condition and four blocks of a *math* condition. The story blocks present participants with brief auditory stories followed by a two-alternatives choice question in which participants are questioned about the topic of the story. The math task is adaptive with the aim of maintaining a similar level of difficulty and engagement across participants. The individual reaction time in the two conditions of the language task (within-subject levels: story, math) was used as the first variable. The second variable for correlation was taken from the battery of behavioral and individual difference measures and consisted of the age-adjusted score of *processing speed* as measured using the NIH Toolbox. This test measures the speed of cognitive processing of visually presented pairs of stimuli (Gershon et al., 2010). These measures were selected for the correlational analyses as they bring reliable and stable measurements on the population (Carlozzi et al., 2015; Wilson et al., 2016). Since the total sample size was $N = 100$, a Monte Carlo procedure was employed selecting randomly subsets of participants (from 10 to 100, with intervals of 10). Subgroups associated with each sample size ([10 20 30 40 50 60 70 80 90 100]) were analyzed 100 times, each time varying the randomly chosen subset of subjects.

RESULTS

Results of the simulations are reported in **Figure 1**. In these plots, the x- and y-axes represent sample size and correlation difference, respectively, between r_1 and r_2. Values are averaged across different levels of r_1 tested.

Figure 1A shows the results of Cohen's q method, which returns an estimate of the effect (q) independent from the sample size N, reflecting only the difference between correlations. Empirically, small, medium, and large effects are defined based upon the magnitude of q. The interpretation of the result needs to be contextualized, however; a *large* difference between two correlations in a small sample ($N < 10$) is not systematically trustworthy. Conversely, a *small* effect might still be an important one (Rosnow and Rosenthal, 1989).

Figures 1B,C show p-values and t-statistics obtained using Fisher's method. The test did not return significant values for relatively small sample sizes, even those with large differences between $r1$ and $r2$ ($\Delta r = 0.4$). However, for larger samples ($N = 40$), this method failed to reject the null hypothesis as indicated by $\Delta r = 0.15$. In the formula used for the t-statistic, sample size increases cause the test statistics to increase logarithmically, with the p-values showing a logarithmic decrease.

The mixed model analysis (LMEM) results are reported in **Figures 1D,E**. Like ANCOVA, mixed models are likely to fail to reject the null hypothesis even in the presence of a large difference between $r1$ and $r2$, as it interprets this difference as not being significant. The failure to reject the null hypothesis happens even with adequate sample sizes, such as an N of 40, by neuroscience standards. However, an advantage of LMEMs is that the means and standard deviations of the original data are retained. Results depend upon how the data points are distributed, and graphs tend to be more scattered than those described in previous paragraphs.

The p-value and effect size d obtained with bootstrapping simulations are reported in **Figures 1F,G**. Like in previous methods, N and r_2 vary while r_1 is fixed. The number (k) of random samplings was set to 200 ($k = 200$). Equivalent results were obtained for pilot simulations with $k = 500$ and $k = 1,000$; however, as these simulations included a reduced number of cycles for computational purposes, their results are not included here. As for Cohen's method, the bootstrap approach provides an effect size estimation but not a p-value. Moreover, the procedure is sensitive to variability in the data, as indicated by the smoothness of the colors.

A *post hoc* comparison among p-values gathered using Fisher, LMEM, and bootstrap methods is reported in **Figure 2**. The bootstrap method was relatively unaffected by the sample size, failing to reject more frequently the null hypothesis only with small samples (e.g., $N < 20$). Conversely, Fisher and LMEM failed to reject the null hypothesis even with significant differences between correlations, whereas the inverse bias was observed with large samples.

Since it may be pointed out that the results described until now may be based on averages across values of r1, an example with a specific value of r1 (r1 = 0.30) is illustrated in **Figure 3**. On the one hand, these results confirm the observations made until this point: Cohen's method ignores sample size; Cohen's and Fisher's methods do not account for variability in the data structure; Fisher's method and LMEM tend to fail to reject the null

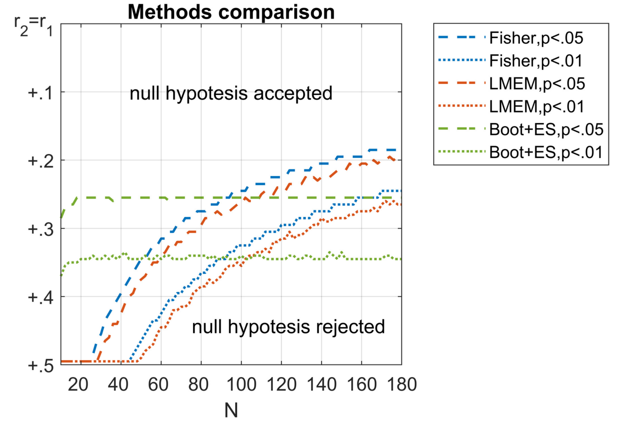

FIGURE 2 | Comparison among Fisher (in blue), LMEM (in red), and bootstrap (in green) p-values. Thresholds of $p = 0.05$ (dashed lines) and $p = 0.01$ (dotted lines) are represented for each method. Only the lower half of the values are reported here for visual clarity.

[1] balsa.wustl.edu

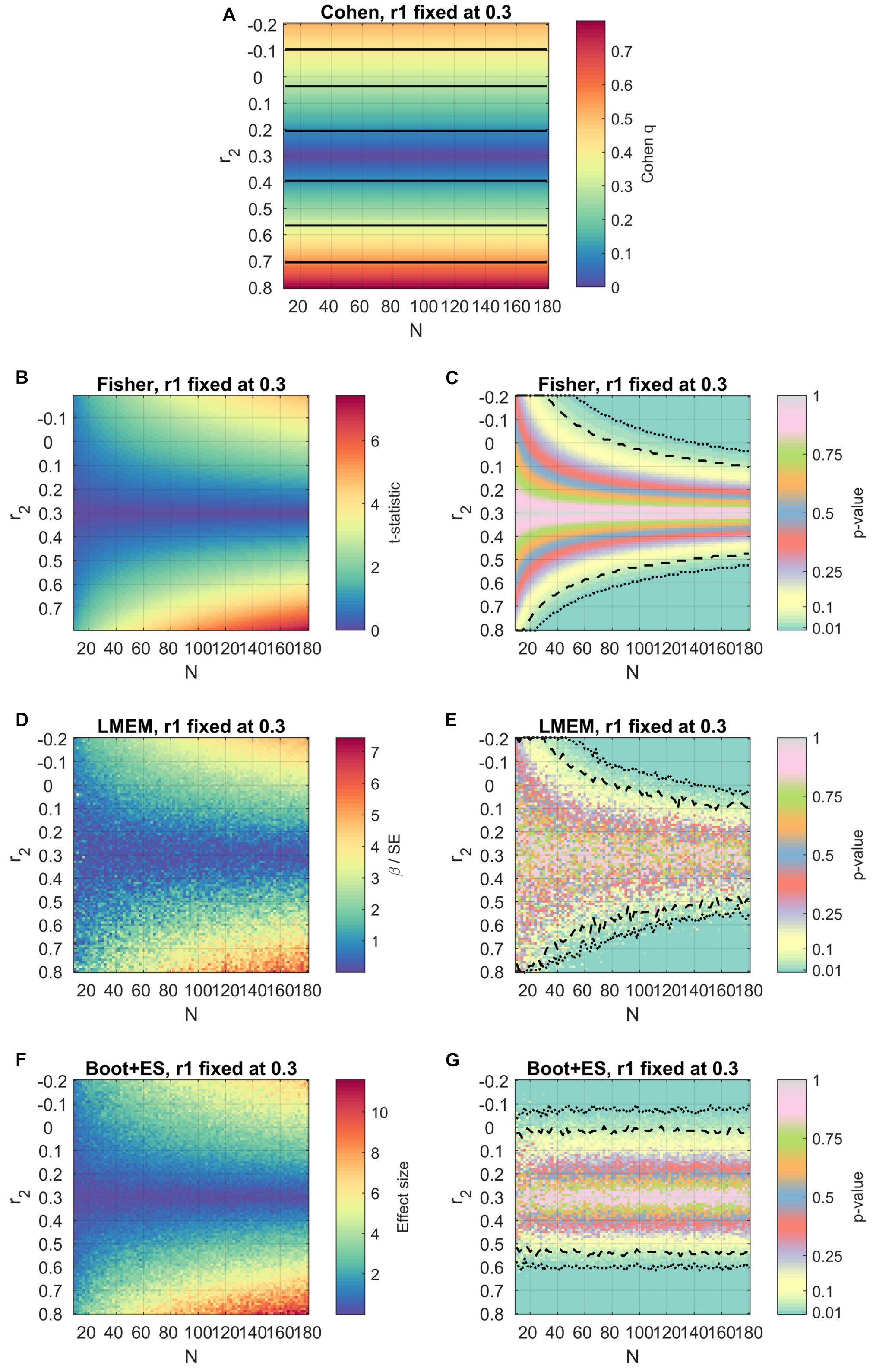

FIGURE 3 | Results of the simulations using the four methods, with the first value of correlation fixed to r_1 = 0.30. The axes, color scheme, and significance lines are the same as used in **Figure 1**. **(A)** Cohen's method. **(B,C)** Fisher's method. **(D,E)** LMEM. **(F,G)** Bootstrap followed by effect size estimation.

hypothesis even with significant differences in the correlational strength. However, the presence of overestimated effect sizes (see for example some points with $N < 20$ in **Figure 3F** for LMEM) shows that the variability in the data structure may endanger the estimation of fixed-effects statistics in mixed-models (Faraway, 2006), probably falling on Simpson's paradoxes (Good and Mittal, 1987). The bootstrap procedure still accounted for variability in the data, but these paradoxical cases were not observed. This eventually happens because the bootstrap procedure limits the pitfalls of classical inferences methods (Killeen, 2005).

Results from the analysis of non-normally distributed data show overlapping results and are reported in **Supplementary Material**.

Finally, results from the analysis of real data confirm a different power of the four methods in detecting a reliable effect. Although the predictable effect of the processing speed on the readiness to respond to math problems, only the bootstrap procedure followed by the effect size estimation allowed to reject the null hypothesis, showing that the reaction time is negatively associated with processing speed only in the math condition of the language task (**Table 1**). Additionally, with this method it is observable a linear increase of the effect size with the sample size, which is a desirable propriety of a statistical method for studying interindividual variability in psychometric and neuroscientific experiments.

DISCUSSION

The present study compared vintage and modern statistical methods used to evaluate differences across correlations in the fields of psychology, medicine, and related disciplines.

A direct comparison between correlation coefficients as provided by Cohen's method can be helpful, given that the calculation only requires the two correlation values (Cohen, 1988). However, Cohen's method estimates the effect magnitude irrespectively of sample size. Conversely, the effect estimated by Fisher's method increases with increasing sample size. Although these two methods can be useful when the original data structure is unavailable, neither Fisher's nor Cohen's method consider potential variability in the data. The effects estimated with these methods should be carefully contextualized and interpreted.

TABLE 1 | Results from the correlational analyses of the real dataset including a task-performance variable (reaction time during language-story and language-math task conditions) and a cognitive "baseline" variable (processing speed as assessed using the NIH Toolbox).

	Cohen	**Fisher**		**LMEM**		**Boot + ES**	
N	***q***	***t*-stat**	***p***	**β**	***p***	***d***	***p***
10	0.33	0.42	0.59	4.63	0.47	0.97	0.065
20	0.29	0.80	0.48	6.54	0.39	1.37	0.103
30	0.23	0.83	0.46	5.52	0.36	1.38	0.094
40	0.23	0.99	0.38	5.93	0.29	1.59	0.052
50	0.23	1.11	0.31	5.81	0.22	**1.81**	**0.024**
60	0.24	1.26	0.25	5.96	0.16	**2.00**	**0.009**
70	0.23	1.31	0.22	5.80	0.14	**2.08**	**0.009**
80	0.24	1.46	0.16	6.06	0.09	**2.31**	**0.002**
90	0.23	1.48	0.15	5.79	0.08	**2.34**	**0.002**
100	0.22	1.56	0.12	5.81	0.06	**2.45**	**<0.001**

Bold values indicate significant effects ($p < 0.05$).

It has been suggested that LMEMs may generate *p*-values that are too small, possibly overestimating the importance of a given effect (Faraway, 2006). The primary purpose of the LMEM is to use data from a continuous variable to estimate differences between levels of a fixed factor, including repeated-measures or longitudinal scenarios (Robinson, 1991). Given the results presented here, LMEM (and other generalized linear models, like the ANCOVA) may be useful for explorative purposes. Still, they may not be the best choice if the aim is to test the difference between the two correlations. The presence of overestimated effect sizes for LMEM suggested possible miscalculation on fixed-effects coefficients in mixed-models (Good and Mittal, 1987; Faraway, 2006).

The bootstrap procedure followed by a test to determine effect size accurately estimated the difference between the strength of the two correlations. Key features of this method are that it considers the sample size and the variability of the initial data, returns a descriptive measure of the difference between the two correlations, and provides a *p*-value not biased by small or large sample sizes and not affected by the pitfalls of classical inferences methods (Killeen, 2005).

Although Fisher, LMEM, and bootstrap methods returned a *p*-value, which is traditionally used to assess the presence or absence of an effect, the debate about how and even if these statistics should be used persists (Goodman, 2019). Due to logistical constraints, certain seminal cognitive neuroimaging studies dealt with relatively small samples (Lindquist et al., 2012; Friston, 2013). Such sample sizes may not be enough to detect significant differences using linear regression (LMEM). Recently, *p*-values have been shown to be misleading measures of the strength of the evidence against the null hypothesis (Berger and Sellke, 1987; Hupé, 2015); also, they do not directly provide an index of effect magnitude (Sullivan and Feinn, 2012). The bootstrap approach provided a descriptive effect size close to the effect estimated by mixed models. The efficiency of the bootstrap approach in this study is in line with the trend toward *p*-independent assays in psychology (Ellis, 2010; Kelley and Preacher, 2012; Tabachnick and Fidell, 2012; Nichols et al., 2017). Furthermore, implementing a *z*-test provided an unbiased, universally used null-hypothesis statistic that may still be useful for *p*-generation researchers (Greenland, 2019; Ioannidis, 2019). Bootstrapping correlations have been discussed extensively in the literature, and researchers have noted that monotonic, transformation invariant procedures like the bootstrap are ill-suited to estimating confidence intervals or testing a null hypothesis (Lunneborg, 1985; Efron and Tibshirani, 1986; Rasmussen, 1987; Efron, 1988; Strube, 1988; Olkin and Finn, 1995). However, the procedure presented here and compared with other methods represents a slightly new approach that combines bootstrap with effect size estimation and null-hypothesis testing.

Importantly, the application to real data confirmed the usefulness of the bootstrap approach. In fact, the Monte-Carlo procedure revealed that the bootstrap followed by the effect size estimation was the only method able to reject the null hypothesis across many sample sizes. Through this method, that the cognitive trait of processing speed significantly predicted reaction time during the "math" condition but not during the "story" condition of a language task. Furthermore, only with the bootstrap method followed by the effect size estimation the strength of the effect regularly increased with increasing sample sizes, reflecting the accumulation of evidence with increasing levels of information.

CONCLUSION

The present study evaluated the efficacy of four different methods for investigating differences in correlations across experimental conditions. Bootstrapping followed by effect size estimation was the most successful, providing a statistic that accounted for both inter-individual and sample size variability in comparing correlation coefficients between experimental conditions. This method is easily implementable in MatLab through the bootes function made available online by the author at MathWorks.

Although these findings have implications for researchers interested in comparing the magnitude of correlations between different experimental conditions, this study has a significant limitation that must be acknowledged. In fact, the bootstrap procedure presented here works well for within-subject analyses and may be applied without complications to between-subjects paradigms but is not yet applicable to mixed experimental designs which include both between- and within-subject factors. Future studies should evaluate these and other methods in such alternative situations to uncover other easily implemented, bias-free tools for researchers in psychology, neuroscience, and medicine.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

SD ideated the study, performed simulations, wrote the manuscript, produced the figures, and wrote the program "bootes."

FUNDING

This work was supported by the "Departments of Excellence 2018-2022" initiative of the Italian Ministry of Education, Research, and University for the Department of Neuroscience, Imaging, and Clinical Sciences of the University of Chieti-Pescara.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.860213/full#supplementary-material

REFERENCES

Berger, J. O., and Sellke, T. (1987). Testing a Point Null Hypothesis: the irreconcilability of p values and evidence. *J. Am. Statist. Assoc.* 82, 112–122.

Binder, J. R., Gross, W. L., Allendorfer, J. B., Bonilha, L., Chapin, J., Edwards, J. C., et al. (2011). Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study. *NeuroImage* 54, 1465–1475. doi: 10.1016/j.neuroimage.2010.09.048

Brysbaert, M., and Stevens, M. (2018). Power analysis and effect size in mixed models: a tutorial. *J. Cogn.* 1:9.

Buckner, R. L., Andrews-Hanna, J. R., and Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann. New York Acad. Sci.* 1124, 1–38. doi: 10.1196/annals.1440.011

Carlozzi, N. E., Beaumont, J. L., Tulsky, D. S., and Gershon, R. C. (2015). The NIH Toolbox Pattern Comparison Processing Speed Test: Normative Data. *Arch. Clin. Neuropsychol.* 30, 359–368. doi: 10.1093/arclin/acv031

Cheung, M. L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: examples and analyses in R. *Behav. Res. Methods* 46, 29–40. doi: 10.3758/s13428-013-0361-y

Cheung, M. L., and Chan, W. (2004). Testing dependent correlation coefficients via structural equation modelling. *Organ. Res. Methods* 7, 206–223. doi: 10.1177/1094428104264024

Cheung, M. L., and Chan, W. (2005). Meta-analytic structural equation modelling: a two-stage approach. *Psychol. Methods* 10, 40–64.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Second Edn. Hillsdale, N.J: Lawrence Erlbaum associates.

Cohen, J. (1990). Things I have learned (so far). *Am. Psychol.* 45, 1304–1312. doi: 10.1037/0003-066x.45.12.1304

Di Plinio, S., Ferri, F., Marzetti, L., Romani, G. L., Northoff, G., and Pizzella, V. (2018). Functional connections between activated and deactivated brain regions mediate emotional interference during externally-directed cognition. *Hum. Brain Mapp.* 39, 3597–3610. doi: 10.1002/hbm.24197

Efron, B. (1987). Better bootstrap confidence intervals. *J. Am. Statist. Assoc.* 82, 171–185.

Efron, B. (1988). Bootstrap confidence intervals: good or bad? *Psychol. Bull.* 104, 293–396. doi: 10.1037/0033-2909.104.2.293

Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statist. Sci.* 1, 54–75.

Ellis, P. D. (2010). *The Essential Guide to Effect Sizes – Statistical power, Meta-analysis, and the Interpretation of Research Results*. Cambridge, UK: Cambridge University press.

Faraway, J. J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. London, UK: Chapman & Hall/CRC Taylor & Francis Group.

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

Friston, K. (2013). Sample size and the fallacies of classical inference. *NeuroImage* 1, 503–504. doi: 10.1016/j.neuroimage.2013.02.057

Garrett, D. D., Kovacevic, N., McIntosh, A. R., and Grady, C. L. (2010). Blood oxygen level-dependent signal variability is more than just noise. *J. Neurosci.* 30, 4914–4921. doi: 10.1523/JNEUROSCI.5166-09.2010

Gershon, R. C., Cella, D., Fox, N. A., Havlik, R. J., Hendrie, H. C., and Wagster, M. V. (2010). Assessment of neurological and behavioural function: The NIH Toolbox. *Lancet Neurol.* 9, 138–139. doi: 10.1016/S1474-4422(09)70335-7

Good, I. J., and Mittal, Y. (1987). The amalgamation and geometry of two-by-two contingency tables. *Ann. Statist.* 15, 694–711.

Goodman, S. N. (2019). Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics. *Am. Statist.* 73, 26–30. doi: 10.1080/00031305.2018.1558111

Greenland, S. (2019). Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values. *Am. Statist.* 73, 106–114. doi: 10.1080/00031305.2018.1529625

Hampson, M., Driesen, N., Roth, J. K., Gore, J. C., and Constable, R. T. (2010). Functional connectivity between task-positive and task-negative brain areas and its relation to working memory performance. *Magn. Reson. Imag.* 28, 1051–1057. doi: 10.1016/j.mri.2010.03.021

Hupé, J. (2015). Statistical Inferences under the null hypothesis: common mistakes and pitfalls in neuroimaging studies. *Front. Neurosci.* 9:18. doi: 10.3389/fnins.2015.00018

Ioannidis, J. P. A. (2019). What have we (not) learnt from millions of scientific papers with p values? *Am. Statist.* 73, 20–25. doi: 10.1080/00031305.2018.1447512

Kelley, K., and Preacher, K. J. (2012). On effect size. *Psychol. Methods* 17, 137–152.

Killeen, P. R. (2005). An Alternative to Null-Hypothesis Significance Tests. *Psychol. Sci.* 16, 345–353. doi: 10.1111/j.0956-7976.2005.01538.x

Lindquist, M. A., Caffo, B., and Crainiceanu, C. (2012). Ironing out the statistical wrinkles in “ten ironic rules”. *NeuroImage* 1, 499–502. doi: 10.1016/j.neuroimage.2013.02.056

Lipsey, M. W., and Wilson, D. B. (2001). *Practical Meta Analysis*. Thousand Oaks, CA: SAGE Publications, 49.

Liu, X. S., and Popmey, K. T. (2020). Bootstrap Estimate of Bias for Intraclass Correlations. *J. Appl. Measure.* 21, 101–108.

Lunneborg, C. E. (1985). Estimating the correlation coefficient: bootstrap and parametric approaches. *Psychol. Bull.* 98, 209–215. doi: 10.1037/0033-2909.98.1.209

Nichols, T. E., Das, S., Eickhoff, S. B., Evans, A. C., Glatard, T., Hanke, M., et al. (2017). Best practices in data analysis and sharing in neuroimaging using MRI. *Nat. Neurosci.* 20, 299–303. doi: 10.1038/nn.4500

Olkin, I., and Finn, J. D. (1990). Testing correlated correlations. *Psychol. Bull.* 108, 330–333. doi: 10.1037/0033-2909.108.2.330

Olkin, I., and Finn, J. D. (1995). Correlations redux. *Psychol. Bull.* 118, 155–164. doi: 10.1037/0033-2909.118.1.155

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes in C. The art of Scientific Computing*, Second Edn. Cambridge: Cambridge University Press, 55.

Rasmussen, J. L. (1987). Estimating the correlation coefficient: bootstrap and parametric approaches. *Psychol. Bull.* 101, 136–139. doi: 10.1037/0033-2909.101.1.136

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statist. Sci.* 6, 15–32. doi: 10-1214/ss/1177011926

Rosnow, R. L., and Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *Am. Psychol.* 44, 1276–1284. doi: 10.1177/0956797620972367

Schneider, B. A., Avivi-Reich, M., and Mozuraitis, M. (2015). A cautionary note on the use of the Analysis of Covariance (ANCOVA) in classification designs with and without within-subject factors. *Front. Psychol.* 6:474. doi: 10.3389/fpsyg.2015.00474

Strube, M. J. (1988). Bootstrap type I error rates for the correlation coefficient: an examination of alternative procedures. *Psychol. Bull.* 104, 290–292. doi: 10.1037/0033-2909.104.2.290

Sullivan, G. M., and Feinn, R. (2012). Using effect size - or Why the P value is not enough. *J. Grad. Med. Educ.* 3, 279–282. doi: 10.4300/JGME-D-12-00156.1

Tabachnick, B., and Fidell, L. S. (2012). *Using Multivariate Statistics*, Sixth Edn. Boston, MA: Pearson.

Wilson, S. M., Bautista, A., Yen, M., Lauderdale, S., and Eriksson, D. K. (2016). Validity and reliability of four language mapping paradigms. *NeuroImage Clin.* 16, 399–408. doi: 10.1016/j.nicl.2016.03.015

Zilles, K., and Amunts, K. (2013). Individual variability is not noise. *Trends Cogn. Sci.* 17, 153–155. doi: 10.1016/j.tics.2013.02.003

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.