



Researcher degrees of freedom in statistical software contribute to unreliable results: A comparison of nonparametric analyses conducted in SPSS, SAS, Stata, and R

Cooper B. Hodges^{1,2,3,4} · Bryant M. Stone⁵ · Paula K. Johnson^{1,6} · James H. Carter III⁷ · Chelsea K. Sawyers⁸ · Patricia R. Roby⁹ · Hannah M. Lindsey^{1,2}

Accepted: 11 July 2022 / Published online: 11 August 2022
© The Psychonomic Society, Inc. 2022

Abstract

Researcher degrees of freedom can affect the results of hypothesis tests and consequently, the conclusions drawn from the data. Previous research has documented variability in accuracy, speed, and documentation of output across various statistical software packages. In the current investigation, we conducted Pearson’s chi-square test of independence, Spearman’s rank-ordered correlation, Kruskal–Wallis one-way analysis of variance, Wilcoxon Mann–Whitney *U* rank-sum tests, and Wilcoxon signed-rank tests, along with estimates of skewness and kurtosis, on large, medium, and small samples of real and simulated data in SPSS, SAS, Stata, and R and compared the results with those obtained through hand calculation using the raw computational formulas. Multiple inconsistencies were found in the results produced between statistical packages due to algorithmic variation, computational error, and statistical output. The most notable inconsistencies were due to algorithmic variations in the computation of Pearson’s chi-square test conducted on 2×2 tables, where differences in *p*-values reported by different software packages ranged from .005 to .162, largely as a function of sample size. We discuss how such inconsistencies may influence the conclusions drawn from the results of statistical analyses depending on the statistical software used, and we urge researchers to analyze their data across multiple packages to check for inconsistencies and report details regarding the statistical procedure used for data analysis.

Keywords Researcher degrees of freedom · Statistical software · Nonparametric procedures · Reproducibility · Statistical conclusion validity

A fundamental component of the scientific method is statistical rigor, which has been defined as the “consistency in conceptual development, epistemological stance, application of analytical tools and transparent reporting of their use, and

subsequent interpretation and reporting of findings” (Köhler et al., 2017, p. 713). Maintaining statistical rigor ensures the appropriate use of statistical analyses that uphold the scientific method and contribute credible results to the scientific literature. Among the most common threats to statistical rigor are the inappropriate use of statistical analyses (Dar et al., 1994; García-Pérez, 2012; Schatz et al., 2005) and the misreporting and misinterpretation of results (Bakker & Wicherts, 2011; Berle & Starcevic, 2007).

The use of inappropriate statistical tests leads to unreliable measurement and threatens statistical conclusion validity, a special form of internal validity that concerns sources of random error and the appropriate use of statistics and statistical tests (Cook & Campbell, 1979; García-Pérez, 2012). Inappropriate test usage often occurs when an analysis is unable to produce a logical or reliable answer to the research question (García-Pérez, 2012). This is commonly seen when traditional, non-robust procedures (e.g., *t*-test, analysis of

A previous version of the present manuscript was uploaded to the PsyArXiv preprint publication database on February 14th, 2020. The preprint publication contains the same material and the same interpretations presented in the current manuscript, but includes a more extensive review of the literature pertaining to the replication crisis and nonparametric inference. Further, the current manuscript used the updated versions of some of the software released after February of 2020 (e.g., SPSS).

Cooper B. Hodges, Bryant M. Stone, and Hannah M. Lindsey contributed equally and are considered co-first authors.

✉ Bryant M. Stone
Bryant.Stone@siu.edu

Extended author information available on the last page of the article

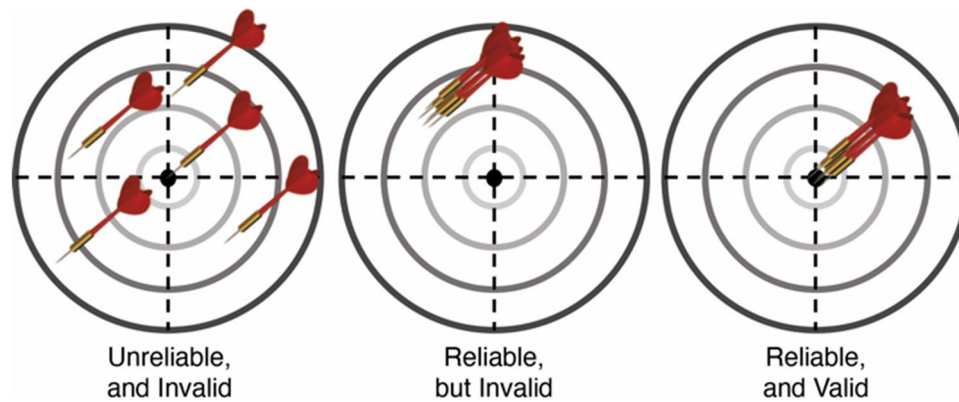


Fig. 1 Reliability is necessary but not sufficient for validity. *Note.* Validity is not present when results are unreliable (left) and may not be present when results are reliable but produced through the same,

improper method (may be the result of direct replication; middle). Validity is only supported by reliable results that are obtained through adequate methods (right).

variance, regression) are used when assumptions, such as normality and homogeneity of variance, are violated (Hoekstra et al., 2012; Keselman et al., 1998; Osborne, 2008). As a result, type I and type II error rates are uncontrolled and effect sizes may be over- or underestimated (Osborne & Waters, 2002), ultimately leading to unreliable results and invalid conclusions. This is a noteworthy problem within the psychological science literature, where any mention of checking parametric assumptions is found in roughly 8% of published studies (Hoekstra et al., 2012; Keselman et al., 1998; Osborne, 2008).

Reliability is necessary, but not sufficient, for validity (Fig. 1). Replication failure is inevitable in the presence of reporting error, which has been shown to occur at relatively high rates in the psychological literature (Bakker & Wicherts, 2011; Berle & Starcevic, 2007; Caperos & Pardo Merino, 2013; Nuijten et al., 2017). While reporting error often results from human error, a lack of knowledge of statistical procedures and a lack of clarity in the output generated by statistical software are also major contributors to the misreporting and misinterpretation of findings. The principles of reliability and validity highlight the importance of reproducing results across various instruments, as one instrument may produce invalid results just as reliably as another produces valid results. This also applies to the reliability of the tools used to conduct statistical analyses, specifically the output of various statistical software packages commonly used within a field of research. Researchers rely on statistical software to produce reliable and accurate results, particularly when large datasets or complex statistical models are used; however, researchers have found differences in computational speed (McCoach et al., 2018), accuracy (McCullough, 2000; McCullough & Heiser, 2008), and quantitative results (Bergmann et al., 2000; Grieder & Steiner, 2020; Keeling & Pavur, 2007) when the same data are analyzed by various statistical software packages. Specifically, inconsistencies

in algorithmic and computational procedures, rounding, and the reporting of test statistics and p -values have been found across various statistical software packages (Bergmann et al., 2000). Such discrepancies may contribute to reporting error (Bakker & Wicherts, 2011; Levine & Atkin, 2004); hence, the statistical software used for data analysis entails a researcher degree of freedom that must be considered when replication attempts are made.

Computational reproducibility can be further complicated by factors such as version control, ambiguous dependencies and programming, and human error in the reporting of results. Although it is becoming increasingly common for scientists to host data, scripts, and results on open-source repositories, problems may still arise in replication. For instance, failure to document file dependencies and lack of clarity in file execution order may make it difficult to replicate results exactly. Even with documentation and code provided, efforts to replicate results can often fail. One recent investigation found that, out of a sample of 2000 studies with accompanying R code, 74% failed to complete without error (Trisovic et al., 2022). Third-party services (such as GitHub and Docker) offer version control and containerization to address these issues; however, researcher knowledge of these services may be limited, as most scientific training does not cover their use (Peikert & Brandmaier, 2021).

In this study, we investigate the reliability of four statistical software programs commonly used in the field of psychology (SPSS, SAS, Stata, and R). To our knowledge, no study to date has examined the reliability of nonparametric statistical output across statistical software programs. We chose the following five nonparametric statistical tests for our primary analysis due to their common use in quantitative research or their equivalence to commonly used parametric procedures (Table 1): Pearson's chi-square test of independence, Spearman's rank-ordered correlation, Kruskal–Wallis one-way analysis of variance, Wilcoxon Mann–Whitney

Table 1 Descriptions of the nonparametric tests included in the present analyses

Nonparametric procedure		Parametric equivalent	
Test	Description	Test	Description
Pearson's chi-square test of independence	A test used to determine whether a relationship exists between discrete variables from two or more independent samples. H_0 : The variables are independent of (i.e., not associated with) one another.	None	–
Spearman's rank-ordered correlation	A rank-based test used to measure the degree of association between two variables measured on at least an ordinal scale. H_0 : The ranked correlation between the two variables for the population is equal to zero ($\rho = 0.00$).	Pearson's product-moment correlation	A test used to measure the degree and direction of a linear relationship between two normally distributed variables that are measured on an interval or ratio scale. H_0 : The linear correlation between the two variables for the population is equal to zero ($\rho = 0.00$).
Kruskal–Wallis one-way analysis of variance	A distribution-free, rank-based test used to compare the medians of three or more samples drawn from independent populations. H_0 : All samples are drawn from symmetrically distributed populations with the same median values ($\theta_1 = \theta_2 = \theta_3$).	One-way analysis of variance	A test used to compare the means of three or more samples drawn from independent populations. H_0 : All samples are drawn from normally distributed populations with the same mean values ($\mu_1 = \mu_2 = \mu_3$).
Wilcoxon Mann–Whitney U rank-sum test	A distribution-free, rank-based test used to compare the medians of two samples drawn from independent populations. H_0 : The two samples are drawn from symmetrically distributed populations with the same median values ($\theta_1 = \theta_2$).	Independent-samples (Student's) t -test	A test used to compare the means of two samples drawn from independent populations. H_0 : The two samples are drawn from normally distributed populations with the same mean values ($\mu_1 = \mu_2$).
Wilcoxon signed-rank test	A distribution-free, rank-based test used to assess the equality of medians between matched pairs of observations drawn from the same population. H_0 : The median of the difference scores for the population is equal to zero ($\theta_D = 0$).	Paired/matched-samples t -test	A test used to assess the equality of means between matched pairs of observations drawn from the same population. H_0 : The mean of the difference scores for the population is equal to zero ($\mu_D = 0$).

U rank-sum test, and Wilcoxon signed-rank tests. We also tested the reliability of estimates of skewness and kurtosis produced by the four software programs, as these estimates are often used by researchers when determining whether nonparametric analysis is appropriate.

The decision to focus on nonparametric procedures was determined through the consensus of all authors for the following reasons: (1) Nonparametric analyses are commonly used in the literature to compare sample characteristics between participant groups (e.g., Pearson's chi-square test of independence), although less commonly for various other analyses. (2) While hand- and computer-based calculations of common parametric analyses are typically covered in undergraduate and graduate statistics courses, nonparametric analyses are rarely or never covered to the same degree, especially in required statistics courses in the behavioral sciences (Alder & Vollick, 2000; Friedrich et al., 2018). This may lead to an increased reliance on statistical software for computing such analyses and a decreased likelihood that errors in computation will be noticed or that variations in default algorithms used by different software platforms will be known. (3) Other researchers have recently published material comparing statistical software programs for common parametric analyses as well as more complex statistical models (Bergmann et al., 2000; Brown et al., 2012; McCoach et al., 2018; Oster & Hilbe, 2008a, 2008b; Wang & Johnson, 2019). (4) The authors felt confident in their own understanding and in the capacity of the software programs used in the present study to perform the chosen nonparametric analyses.

Nonparametric inference

The use of nonparametric procedures provides a way to combat many of the threats to statistical rigor described above. There are several advantages to nonparametric methods, namely that they are distribution-free, and thus require few assumptions about the underlying populations from which the data are obtained. Their relative lack of reliance on assumptions and distribution-free nature protects researchers from making false conclusions that can result from misleading significance values obtained from parametric procedures performed in the presence of violated assumptions (Potvin & Roff, 1993). Disadvantages to using nonparametric methods exist as well, and the greatest of these is that they are geared toward hypothesis testing rather than effect estimation. Although it is possible to obtain nonparametric estimates of effect and associated confidence intervals, the processes by which this is done is generally not straightforward (Whitley & Ball, 2002). For example, while critical value tables for probability distributions used for parametric analyses (e.g., z , t , F) are commonly provided in textbooks and online resources,

such resources are scarce for the majority of nonparametric statistics. Furthermore, statistical software is often limited in its ability to perform nonparametric procedures.

In contrast to parametric analyses, a large number of algorithmic variations exist for the way in which rank-based nonparametric methods can be implemented, and statistical packages are not consistent in their application of such procedures. The algorithms used to compute test statistics for rank- and frequency-based nonparametric tests vary in one or more of the following three ways: (1) whether the exact, null distribution or an asymptotic, large-sample approximation of the normal (z) or χ^2 distribution is used; and in the latter case, whether corrections for (2) continuity and/or (3) tied ranks are applied (Lehmann, 1998; Siegel & Castellan, 1988).

Exact versus asymptotic distributions

Exact p -values and confidence intervals are calculated using the true underlying null distribution, which is discrete in most parametric inference; however, due to computational inefficiency and a lack of null distribution tables for samples of approximately $n > 30$, an approximation of the true distribution is typically used to calculate asymptotic probabilities for large samples. Asymptotic distributions are approximations of the true underlying distribution and rely on the central limit theorem, assuming that the sample is large enough for the test statistic to approach the normal distribution. Parameters that do not assume a normal distribution, such as those derived from rank-ordered nonparametric tests, are often estimated using asymptotic distribution-free methods that evaluate the median, rather than the mean (Huang & Bentler, 2015; Neave & Worthington, 1988). There is some controversy over the appropriateness of asymptotic versus exact p -values for statistical inference with contingency tables (García-Pérez & Núñez-Antón, 2020; Lydersen et al., 2009; Prescott, 2019); however, while exact and asymptotic probabilities obtained from nonparametric analysis are generally very similar when obtained from large samples, they can be quite different when sample sizes are small. In such situations, asymptotic p -values may lead to unreliable and misleading conclusions if used inappropriately. For example, in a systematic examination of differences between asymptotic and exact probabilities extracted from various nonparametric tests performed on small samples, Mundry and Fischer (1997) found that, relative to exact statistics, asymptotic procedures led to an increase in type I error rates when used for Wilcoxon signed-rank and Wilcoxon Mann–Whitney U tests and an increase in type II error rates when used to compute Spearman's rank-ordered correlation coefficient.

Continuity correction

When asymptotic distributions are continuous (e.g., normal, χ^2), a continuity correction (the addition or subtraction of 0.5 to a discrete x -value) may be applied to improve approximations of the underlying discrete null distribution of nonparametric procedures (Gibbons & Chakraborti, 2011). The appropriateness of continuity correction procedures is also controversial, however, as some argue that they are only appropriate for one-sided tests (Haber, 1982; Mantel, 1976). Some commonly used continuity correction procedures, such as Yates' correction for the chi-square test (Yates, 1934), are systematically conservative, resulting in overcorrected (larger) p -values when applied to two-sided tests (Maxwell, 1976; Stefanescu et al., 2005). Arguments against the use of a continuity correction for chi-square tests of independence for 2×2 tables date back to 1947, where Pearson stated that "... it becomes clear that in the case of small samples, at any rate, this method of introducing the normal approximation gives such an overestimate of the true chances of falling beyond a contour [of $p = .05$ or $.01$] as to be almost valueless" (p. 155). In support of this, simulation studies have demonstrated that the application of Yates' continuity correction to Pearson's chi-square tests performed on medium to small samples overcorrects the probability of an outcome, resulting in below nominal type I error rates and a substantial loss of power (Campbell, 2007; Garside & Mack, 1976; Grizzle, 1967; Richardson, 1990). Despite this, current textbooks and software packages are inconsistent in their recommendation and default use of continuity correction procedures (Hitchcock, 2009).

Correction for ties

Rank-ordered statistics are typically assumed to be drawn from a continuous population, of which the probability of any two observations being equal in magnitude is zero (Gibbons & Chakraborti, 2011; Siegel, 1957). In practice, two or more observations of the same magnitude commonly occur due to measurement imprecision or because the population distribution is actually discrete. Such observations are considered to be *tied*, and some method of assigning unique ranks to tied values must be applied so that the test statistics that depend on relative magnitudes of observations (i.e., rank-order statistics) can be computed (Gibbons & Chakraborti, 2011). Most commonly, individual observations within a group of tied observations are assigned the average of the group's ranked value. Although this method does not affect the mean ranked value, it reduces the variance in the ranks. This affects the underlying null distribution; thus, a correction for ties is typically applied to the calculation of the test statistic.

The appropriateness of applying any of the algorithmic variations described above depends on the characteristics of the data. Typically, the default algorithms used within a given statistical software package depend on the sample size and the presence or absence of tied ranks. As a result, a single software package may apply different procedures to different datasets without user intervention. In addition, the criteria used to determine when exact or asymptotic distributions are used and whether corrections are applied differ between software packages. To further complicate this, multiple algorithms exist for determining exact statistics, and various continuity corrections have been suggested for a given nonparametric test; often, the algorithms and correction procedures used are simply the choice of the software developer rather than that which is most appropriate for the data. When the default algorithmic specifications are not made clear in statistical output, non-statisticians are unlikely to know whether correction procedures were applied to asymptotic p -values or whether the probability statistics reported are exact. As a consequence, inappropriate procedures may be used, and statistical errors or misleading results may be reported. Furthermore, because default algorithms vary across statistical packages, the results reported by researchers who are naïve to the specific algorithms used to compute them might have differed if another statistical package had been used; thus, the conclusions made from statistical analyses conducted under such conditions may be mere reflections of the software used rather than the actual data analyzed.

Transparency in research

To address recent concerns over replication (Open Science Collaboration, 2015; 2012), the scientific community is calling for transparency in research through open science, which may aid in both reducing the number of false positives published in the literature and increasing reproducibility and replicability by encouraging ethical research practices (Ioannidis, 2014). Proponents of open science advise researchers to document and share a detailed account of their study procedures (Borghi & Van Gulick, 2018), make their data available to the public, and preregister all studies. Some have also argued that researchers should document all researcher degrees of freedom (Wicherts et al., 2016), or choices that researchers make when designing a study and collecting, analyzing, and reporting the data (Simmons et al., 2011); this should include a report of the statistical software (and version) used and documentation of the command syntax used for statistical analysis. Researcher degrees of freedom may increase type I error rates, inflate effect sizes, misrepresent the scientific process by undermining the hypothetico-deductive model of the scientific method (Fig. 2), and lead to

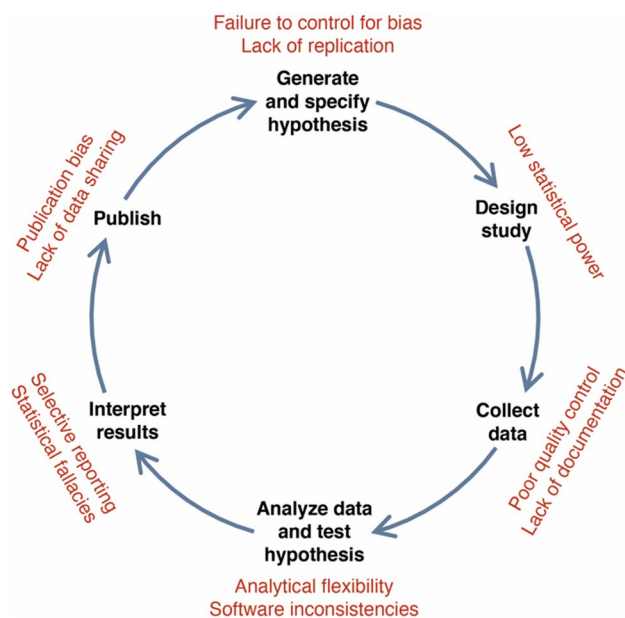


Fig. 2 Researcher degrees of freedom compromise the hypothetico-deductive model of the scientific method. *Note.* There are several opportunities for researcher degrees of freedom (outside, in red) to impede the scientific method, including a failure to control for bias, lack of replication, low statistical power, poor quality control, lack of documentation, analytical flexibility, inconsistencies in algorithms used across statistical software (particularly relevant to the present study), selective reporting, statistical fallacies, publication bias, and lack of data sharing, among others. These factors work together to undermine the robustness of published research. Adapted from *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*, by C. Chambers, 2019, Princeton University Press, p. 17 and “A Manifesto for Reproducible Science,” by Munafò et al., 2017

statistical findings that are difficult to reproduce in the same data or replicate in new data (Asendorpf et al., 2013; Ioannidis, 2005; Wicherts et al., 2016). Documenting researcher degrees of freedom may make it possible to distinguish between true and artifactual replication failures by showing that significant findings occur only when another researcher makes the exact same methodological decisions (Epskamp, 2019; Wicherts et al., 2016).

Rationale and objective

We see a gap in the current literature, as there are few published studies that methodically document the differences produced between statistical software programs. Although some previous work has demonstrated the formulaic and computational differences in the calculation of certain parametric and nonparametric methods (Bergmann et al., 2000; Brown et al., 2012; McCoach et al., 2018; Oster & Hilbe, 2008a, 2008b; Wang & Johnson, 2019), this work is largely

geared toward mathematicians and statisticians, resulting in considerable difficulty in generalization to various subfields within the social and behavioral sciences. We believe that the existence of such statistical discrepancies would constitute a larger problem in the field that may be currently overlooked; hence, the present methodological investigation seeks to systematically document the reliability of statistical output generated from five nonparametric analyses and two measures of normality across four statistical software packages commonly used in the social and behavioral sciences: SPSS, SAS, Stata, and R.

We conducted this investigation using a large sample ($n = 1000$) drawn from a public dataset that is widely used in the social psychology literature. All analyses were conducted using the default parameters of each software package and as few additional specifications/options as possible, so as to reduce the number of researcher degrees of freedom involved in our analysis. We hypothesized that the results of all analyses would be largely consistent across the statistical platforms, although minor discrepancies were expected at the decimal point level due to rounding error. If any larger discrepancies in the results were found, they were expected to occur due to differences in the default algorithms used across the software packages.

Additional investigations

Upon seeing the results of our primary analysis, we questioned whether the same findings would be observed in smaller samples that more accurately reflect those seen in the majority of the psychological science literature. We performed an informal systematic review of the recent literature to determine the appropriate sample sizes to use in these post hoc investigations. The PsycINFO database was searched on November 9, 2020, for peer-reviewed, English-language articles published in scientific journals within the last 5 years that employed one or more of the nonparametric analyses investigated herein using the following terms (search criteria were applied within the full text of the article): “Chi-square” OR “Spearman” OR “Kruskal–Wallis” OR “Wilcoxon” OR “Mann–Whitney.” Published study protocols, conference abstracts, meta-analyses, review articles, and mathematical models were excluded. The search results included a total of 792 publications with linked full text available to PsycINFO subscribers. The abstract and methods sections were examined, and the full-text article was retrieved for the first 100 studies (sorted by relevance) that met our inclusion criteria. Data pertaining to sample size, nonparametric test(s) performed, and statistical software package used were extracted from each article (Supplemental Table S1).

The most recent 100 published studies within the psychological science literature that used nonparametric analyses

reported results for samples sizes ranging from 9 to 2991 ($M = 377.64$, $SD = 636.92$, median = 165.50). Based on these data, our large- N sample size of $n = 1000$ was as large as or larger than approximately 90% of the samples represented in these 100 studies. To generalize the results to more commonly reported sample sizes within the field, we replicated our analysis on medium ($n = 200$) and small ($n = 100$) sample sizes of nonoverlapping data taken from the same public dataset used for the large- N analysis. The medium and small sample sizes chosen approximately reflect the 60th and 30th percentiles, respectively, of those represented in the recent literature. It is important to note that, because the decision to conduct the medium- N and small- N analyses was made after the results of the initial, large- N analysis were known, these post hoc investigations were conducted for exploratory purposes only. Specifically, we explored the extent to which the inconsistencies found in the results of the large- N analysis were present in the results of medium- N and small- N analyses.

Finally, to determine the extent to which any discrepancies found occurred consistently across various samples of data, we performed our analyses on thirty simulated datasets based on the properties of our original small-, medium-, and large- N datasets for confirmatory purposes. The original and simulated datasets are publicly available and can be found in our Open Science Framework (OSF) storage directory at <https://osf.io/35umb/>.

Method

Dataset

The open-source Race Implicit Association Test (IAT) 2018 data (Xu et al., 2014) were used for all of the present analyses.¹ The Race IAT was designed to assess for underlying attitudes toward white and black people that participants may be unwilling or unable to identify within themselves (Lane et al., 2007). Project Implicit hosts the Race IAT data for public use (<https://osf.io/gwofk/>) and allows researchers free access to the data. We chose this dataset due to its widespread use within the social and behavioral sciences (e.g.,

14,200 results were returned for a single Google Scholar search for “Implicit Association Test” AND “Race” on November 8, 2020).

We selected eight variables (Table 2) from the Race IAT Public 2018 dataset based on the relevance of their scale of measurement and number of levels for nonparametric analyses, the fewest number of missing data points, and the likelihood of being used in the experimental research published in the literature. It is important to emphasize that the variables used in this work were not chosen for their underlying measured constructs, and thus the results of our analyses are not meant for conceptual interpretation; rather, the variables used in the present study are solely meant for the *quantitative* comparison of results obtained across statistical software packages.

The raw Race IAT Public 2018 dataset consists of 460 variables and 859,470 observations. Data cleaning was performed using Stata version 16.1 (StataCorp LLC, College Station, TX), and this initially involved dropping all observations from respondents who did not complete the Race IAT (session_status values != “C”), in addition to 1290 observations, which were dropped for respondents who failed to complete all 120 trials across the four combined-task blocks (N_{3467} values < 120). Finally, Excel was used to identify 59,530 of the remaining 414,167 observations, which were removed due to missing data. From the remaining 354,637 observations, the small-, medium-, and large- N datasets were generated by randomly selecting 100, 200, and 1000 nonoverlapping observations, respectively, using Excel’s RAND function. Observations randomly assigned numbers 1–1000, 1001–1200, and 1201–1300 were respectively assigned to the large-, medium-, and small- N datasets.

Simulations

After obtaining preliminary results from the primary and exploratory analyses, the decision was made to conduct a simulation study to estimate the relative consistency of computational errors and algorithmic variations that were uncovered. Ten simulated datasets were generated for each of the three original datasets, resulting in a total of thirty simulations. Data were generated in MS Excel for Mac version 16.4 using the random number generator function and based on the properties of the original datasets (Supplemental Table S2). Five of the eight variables (sex, race, black, white, order) follow a discrete distribution, and the remaining three variables (latency, discrim1, discrim2) follow a normal distribution, for which a seed of 1234 was used. The simulated data mirror the descriptive properties of the full dataset. All simulated datasets and tables with results for each dataset are publicly available in our OSF repository (<https://osf.io/35umb/>).

¹ The original Race IAT Public 2018 dataset was downloaded on June 6, 2019, from the data archives hosted on the Project Implicit website (<https://projectimplicit.net>); however, the Project Implicit data archives have since been relocated to <https://osf.io/z4bd2/>, where the raw Race IAT Public dataset and corresponding codebook can be accessed directly. It is unknown whether any data loss/corruption may have occurred during this transition; thus, the raw dataset we downloaded prior to the relocation of the Project Implicit data archives, which was used in the present analyses, can also be accessed from our OSF storage directory.

Table 2 Variables selected from the Race IAT Public 2018 dataset for use in the present analyses

Variable name (<i>IAT name</i>)	Description	Measurement scale	Levels
id (<i>session_ID</i>)	Unique session identification number generated when one begins the IAT	String	–
sex (<i>birthsex</i>)	Biological sex assigned at birth	Nominal	2
race (<i>raceomb_002</i>)	Racial association	Nominal	8
black (<i>tblack_0to10</i>)	Feelings of warmth or coldness toward Black people	Ordinal	11
white (<i>twhite_0to10</i>)	Feelings of warmth or coldness toward White people	Ordinal	11
order (<i>Order</i>)	Task presentation order	Nominal	2
latency (<i>Mn_Rt_all_3467</i>)	Mean reaction time across all task blocks	Ratio	–
discrim1 (<i>D_biep.White_Good_36</i>)	Discriminability score on blocks 3 and 6	Interval	–
discrim2 (<i>D_biep.White_Good_47</i>)	Discriminability score on blocks 4 and 7	Interval	–

The full codebook for the variables used in the present study can be accessed in our OSF storage directory at <https://osf.io/35umb/>

Present analyses

We compared the results of five commonly used nonparametric tests, including Pearson’s chi-square test of independence, Spearman’s rank-ordered correlation (ρ), the Kruskal–Wallis one-way analysis of variance, the Wilcoxon–Mann–Whitney U rank-sum test, and the Wilcoxon signed-rank test. Each of these nonparametric tests are described in detail in Table 1. Because it is often necessary to evaluate normality when deciding whether parametric tests are inappropriate, we also compared calculations of skewness (the degree to which a set of data are symmetrically or asymmetrically distributed around the mean) and kurtosis (the extent to which the peakedness of a probability distribution deviates from the shape of a normal distribution) across software platforms.

Although we selected these nonparametric analyses to test in the current study, we recommend that any analyses conducted within MS Excel, SPSS, SAS, Stata, or R be scrutinized and checked across two or more software packages. Researchers who use one software package exclusively for their analyses may find it beneficial to replicate their findings in different software packages, especially when the underlying algorithmic procedures vary across packages.

Documentation of procedure

The statistical software packages used include the following: SPSS 27 (IBM Corporation, LLC, Armonk, NY, 2020), SAS JMP Pro 15.0 (running SAS v 9.4; SAS Institute Inc., Cary, NC, 2019), Stata 16.1 (StataCorp LLC, College Station, TX, 2020), and R 4.0.3 (The R Foundation, 2020). Although R provides many libraries that we could have utilized in our investigation, the purpose of this project was to evaluate the reliability of nonparametric tests using the fewest researcher degrees of freedom; thus, for the current study, we only used the “stats” package, which is 1 of 14 pre-installed packages in R (v 4.0.3). It is possible and likely that the results of the same analyses produced by different R packages would result in meaningful differences; however, that examination is beyond the scope of the current paper. All code used to generate results in command-line-based programs (Stata, R) were documented, and in the case of programs that rely primarily on a graphical user interface (SPSS, SAS JMP Pro), the underlying command syntax generated from the point-and-click commands was also documented. The primary and simulated datasets, the scripts or command syntax used for all analyses (including the MS Excel spreadsheets used for hand calculations, which are described below), all relevant help files/documentation provided by each software package,

and logs of the full statistical output generated from each software platform can be found in our OSF storage directory at <https://osf.io/35umb/>.

Hand calculation procedures

Prior to comparing the results across packages, we calculated each test statistic and asymptotic p -value with and without corrections for continuity and ties (when applicable) using the raw, computational formulas (see Supplemental Material), and these hand-calculated results were considered the ground truth by which all results generated from statistical packages were compared. Due to the large size of the dataset used, literal hand calculations were not possible. Rather, we performed these calculations in MS Excel for Mac version 16.4 (Microsoft Corporation, Redmond, WA, 2020) and henceforth refer to these results as those derived by hand calculation. All hand calculations were checked for accuracy independently by two or more researchers with adequate knowledge of the relevant statistical procedures and extensive experience using the formula builder function in MS Excel.

Hand calculations are useful to perform if the results of the same analyses are conflicted across two or more statistical software packages, as hand calculations may provide a standard by which to compare the discrepant results produced by the statistics software. The benefit of hand calculation is that researchers may more easily determine where the statistical software programs diverge, and which one is providing the output intended by the researcher. This method is best when documentation is ambiguous or missing (e.g., in some R libraries). If documentation is present and thorough across programs, checking the differences in documentation may be easier than hand calculations.

The following procedures were involved in the hand calculations for each statistical test (refer to the Supplemental Material for additional detail):

- *Pearson's chi-square test of independence*: To obtain asymptotic test statistics, 2×2 observed and expected frequency tables were generated, and the resulting values were used to obtain the degrees of freedom, χ^2 test statistic, and a right-tailed, asymptotic p -value with and without Yate's continuity correction.
- *Spearman's rank-ordered correlation*: Spearman's rank correlation coefficient (ρ) and a two-tailed, asymptotic p -value were generated after applying a correction for ties. A table of tied ranks was also generated, where unique values and the corresponding number of ties are provided for each variable.
- *Kruskal–Wallis one-way analysis of variance*: To obtain asymptotic test statistics, a frequency table including the rank sums and average squared rank sums was generated.

Additionally, the degrees of freedom, Kruskal–Wallis H test statistics (following the χ^2 distribution), and right-tailed, asymptotic p -values with and without a correction for ties were calculated. To determine the tie correction value (C_H), a second table was generated, which includes a list of all unique values with the corresponding number of ties, as well as a range of the number of tied values present in each set of tied values.

- *Wilcoxon Mann–Whitney U rank-sum test*: To obtain asymptotic test statistics, a frequency table was generated, which reports the number of total observations and rank sums for each level of the grouping variable, as well as the mean and variance of the Wilcoxon W statistic and the mean and standard deviation of the rank sums for the smaller of the two groups. Additionally, a second table was generated to aid in the tie correction, which includes a list of all unique values with the corresponding number of ties. Finally, the following test statistics were calculated after correcting for ties, both with and without application of a continuity correction: Wilcoxon W , Mann–Whitney U , z -statistic, and a two-tailed, asymptotic p -value.
- *Wilcoxon signed-rank test*: To obtain asymptotic test statistics, we generated a frequency table including the number of observations and rank sums for the positive and negative ranks, as well as the number of zero differences, the number of ties, and the mean and standard deviation of the Wilcoxon T statistic (based on the positive ranks). Due to a lack of tied ranks, no table of tied values was produced for this analysis. Finally, the following test statistics were calculated both with and without application of a continuity correction: Wilcoxon T , z -statistic, and a two-tailed, asymptotic p -value.
- *Skewness and kurtosis*: Computational formulas were used to determine the population skewness (Cramér, 1946), sample skewness (Bliss, 1967), population kurtosis, where the mean kurtosis of the normal distribution is equal to three (Bock, 1975), and excess sample kurtosis, where the kurtosis of the normal distribution is corrected to have a mean equal to zero (Cramér, 1946).

Criteria for meaningful differences in results

We used the following criteria to determine what would be considered “meaningful” differences in results produced across packages: inconsistencies due to (1) algorithmic variation (e.g., no correction, adjustment for ties, continuity correction), (2) computational error (i.e., results differ from those obtained through hand calculation despite the use of the same computational procedures described in the software documentation), or (3) statistical output (e.g., reporting of exact versus asymptotic p -value; reporting a p -value without the test statistic or vice versa). The latter

criteria were included due to the assumption that the same statistics should be reported across software programs, and the use of any variations in the calculation of a test statistic or p -value should be made explicitly clear in the output and provided along with traditional (i.e., uncorrected) calculations. Finally, we selected these criteria because they have the potential to affect statistical conclusion validity. We recommend that researchers who choose to examine their results across statistical software platforms assess for these criteria as the cause for these meaningful differences.

Results

The default procedures and algorithms used by SPSS, SAS, Stata, and R to conduct Pearson's chi-square test of independence, Spearman's rank correlation coefficient (ρ), Kruskal–Wallis one-way analysis of variance, Wilcoxon Mann–Whitney U rank-sum test, Wilcoxon signed-rank test, skewness, and kurtosis are reported in Table 3. Additionally, Table 3 includes details regarding the hypothesis tested, according to the respective package's help files/documentation, the statistical output generated from each analysis, and the basic command syntax with additional options or arguments that must be specified by the user (i.e., are not applied by default) in order to generate specific results (e.g., test statistic, exact p -values) or to use alternative algorithms (e.g., apply a continuity correction). Greater detail regarding the default procedures and algorithms used by each package can be found in the Supplemental Material and in the help files/documentation for each software package; we have extracted the relevant documentation for the present analyses, and it is included in our OSF storage repository at <https://osf.io/35umb/>.

The numerical results of all large-, medium-, and small- N analyses are reported in Tables 4, 5, and 6, respectively, for comparison across all statistical packages with the hand-calculated results. The results in each table are organized according to the default algorithms used by a given package (i.e., asymptotic test statistics with or without correction for ties and/or continuity). Exact p -values were not generated by hand for any analysis.

Algorithmic variation

The default algorithms used to produce the results of all analyses differ widely across the four packages we compared.

Pearson's chi-square

When conducting Pearson's chi-square test, SAS and Stata apply no continuity correction, R applies Yates' continuity

correction, and SPSS generates results both with and without Yates' continuity correction. As a result, the only p -value produced by R is larger than the only p -values produced by SAS and Stata by values of .049, .073, and .154 for the large-, medium-, and small- N analyses, respectively. Additionally, Fisher's exact one- and two-tailed p -values are provided by SPSS for all data and by SAS for 2×2 contingency tables, by default. With the exception of minor rounding discrepancies, the two-tailed p -values provided were equal between SPSS and SAS across the large-, medium-, and small- N analyses.

Spearman's rho

All four packages adjust for ties, by default, when computing Spearman's rho; however, if no ties are present and the sample size is less than 1290, R will produce an exact p -value in place of the asymptotic p -value.

Kruskal–Wallis

Although no ties were present in the variables assessed by the Kruskal–Wallis test, R does not adjust for ties by default, whereas results are adjusted for ties, by default, in SPSS and SAS, and Stata produces results both with and without an adjustment for ties.

Wilcoxon Mann–Whitney U

When performing the Wilcoxon Mann–Whitney U test, all packages adjust for ties; however, SAS and R also apply a continuity correction, by default, resulting in p -values that are larger by values of .001 and .002 than those which are produced by SPSS and Stata for the medium- and small- N analyses, respectively. Additionally, the formulas used to calculate the Wilcoxon Mann–Whitney U test statistics are not consistent across packages. In Stata and R, Wilcoxon's W is calculated according to the method originally defined by Wilcoxon (1945), where W is the sum of ranks for the first sample (determined by the values that define each group in ascending order), whereas Wilcoxon's W is defined by SPSS as the sum of ranks for the second sample and by SAS as the sum of ranks for the sample of smaller size (if sample sizes are equal, the sum of ranks for the second sample is used). As a result, SPSS produced different values for W and U , along with an inverted z -statistic across the large-, medium-, and small- N analyses, and SAS produced a different value for W and an inverted z -statistic for the large- and small- N analyses. Finally, Stata also produces an exact p -value for samples less than or equal to 200 using a recursive algorithm defined by Hill and Peto (1971). No other packages produced exact p -values for the Wilcoxon Mann–Whitney U ; however,

Table 3 Default statistical procedures employed by each software package

Package	Null Hypothesis (H_0)	Algorithm(s)	Statistical Output	Command Syntax
SPSS	The rows and columns in a two-way table are independent	No correction; Continuity-corrected; Exact ^a	<p>Pearson's Chi-square</p> <ul style="list-style-type: none"> - n_{valid}, $n_{missing}$, n_{total} - $\%_{valid}$, $\%_{missing}$, $\%_{total}$ - Contingency table with f_o (cell, margin, and grand total) - Pearson's χ^2 test statistic <ul style="list-style-type: none"> - <i>df</i> - Two-tailed, asymptotic <i>p</i>-value - Continuity-corrected Pearson's χ^2 test statistic <ul style="list-style-type: none"> - <i>df</i> - Two-tailed, asymptotic <i>p</i>-value - Likelihood ratio χ^2 test statistic <ul style="list-style-type: none"> - <i>df</i> - Two-tailed, asymptotic <i>p</i>-value - Fisher's exact one- and two-tailed <i>p</i>-values (McNemar, 1947; 2×2 tables) - Mantel-Haenszel's Linear-by-Linear Association χ^2 test statistic <ul style="list-style-type: none"> - <i>df</i> - Two-tailed, asymptotic <i>p</i>-value - Mosaic plot - Contingency table with f_o, %, and χ^2 for each cell - n_{total} - <i>df</i> - Negative log-likelihood - RSquare (U) - Likelihood ratio χ^2 test statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic <i>p</i>-value - Pearson's χ^2 test statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic <i>p</i>-value - Fisher's Exact one- and two-tailed <i>p</i>-values (2×2 tables) - Contingency table with f_o - Pearson's χ^2 test statistic <ul style="list-style-type: none"> - <i>df</i> - Two-tailed, asymptotic <i>p</i>-value 	<p>CROSSTABS /TABLES=v1 BY v2 /STATISTICS=CHISQ.</p> <p><i>Additional options</i> /CELLS=EXPECTED: Report expected frequencies in contingency table /CELLS=ROW COLUMN TOTAL: Report row, column, and total percentages in contingency table /METHOD=EXACT: Compute the exact significance level for all statistics in addition to the asymptotic results</p>
SAS	The distribution of one categorical variable is equal across each level of the other variable	No correction; Exact ^a	<ul style="list-style-type: none"> - Two-tailed, asymptotic <i>p</i>-value - Mosaic plot - Contingency table with f_o, %, and χ^2 for each cell - n_{total} - <i>df</i> - Negative log-likelihood - RSquare (U) - Likelihood ratio χ^2 test statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic <i>p</i>-value - Pearson's χ^2 test statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic <i>p</i>-value - Fisher's Exact one- and two-tailed <i>p</i>-values (2×2 tables) - Contingency table with f_o - Pearson's χ^2 test statistic <ul style="list-style-type: none"> - <i>df</i> - Two-tailed, asymptotic <i>p</i>-value 	<p>Contingency (Y (:v1) , X (:v2) , Contingency Table (Cell Chi Square (1))) ;</p>
Stata	The rows and columns in a two-way table are independent	No correction	<ul style="list-style-type: none"> - Two-tailed, asymptotic <i>p</i>-value 	<p>tabulate v1 v2, chi2</p> <p><i>Additional options</i> cell: Report relative frequency of each cell cchi2: Report Pearson's χ^2 in each cell expected: Report expected frequencies exact: Compute Fisher's exact one- and two-tailed <i>p</i>-values chisq.test (v1, v2)</p> <p><i>Additional options</i> correct=FALSE: Do not apply continuity correction</p>
R	The joint distribution of the cell counts in a two-way contingency table is the product of the column and row marginal values	Continuity-corrected	<ul style="list-style-type: none"> - Continuity-corrected Pearson's χ^2 test statistic - <i>df</i> - Two-tailed, asymptotic <i>p</i>-value 	<p>chisq.test (v1, v2)</p> <p><i>Additional options</i> correct=FALSE: Do not apply continuity correction</p>

Table 3 (continued)

SPSS	The correlation between the two rank-ordered variables is equal to zero	Adjusted for ties	Spearman's Rho - Spearman's rho correlation coefficient - Two-tailed, asymptotic <i>p</i> -value - <i>n</i> _{Total}	NONPAR CORR /VARIABLES=v1 v2. <i>Additional options</i> /PRINT=ONETAIL: Report one-tailed significance value Multivariate (Y (:v1, :v2) , Spearman's rho (1)); spearman v1 v2 cor.test (v1, v2, method="spearman") <i>Additional options</i> continuity=TRUE: Apply continuity correction
SAS	The correlation between the two rank-ordered variables is equal to zero	Adjusted for ties	- Pearson's correlation coefficient - <i>n</i> _{Missing} - Spearman's rho correlation coefficient - Two-tailed, asymptotic <i>p</i> -value - Color-mapped scatterplot matrix	
Stata	The two variables are independent of each other	Adjusted for Ties	- <i>n</i> _{Total} - Spearman's rho correlation coefficient - Two-tailed, asymptotic <i>p</i> -value - S test statistic ^b	
R	The rank-based association between paired samples is equal to zero	Adjusted for ties OR Exact ^a	- Spearman's rho correlation coefficient - Two-tailed, asymptotic <i>p</i> -value (if <i>n</i> > 1290 or ties are present) - Exact <i>p</i> -value (Best & Roberts, 1975; if <i>n</i> < 1290 and no ties are present)	
SPSS	Two or more samples are from the same rank-based population distribution	Adjusted for ties	Kruskal-Wallis - <i>n</i> and <i>M</i> _{<i>R</i>} for each level of grouping variable - <i>n</i> _{Total} - Kruskal-Wallis χ^2 test statistic - <i>df</i> - One-tailed, asymptotic <i>p</i> -value - One-way plot of responses across all levels of the grouping variable - <i>n</i> , ΣR , expected ΣR , <i>M</i> _{<i>R</i>} , and <i>z</i> _{<i>R</i>} for each level of the grouping variable - Kruskal-Wallis χ^2 test statistic - <i>df</i> - One-tailed, asymptotic <i>p</i> -value	NEPAR TESTS /K-W=v1 BY cv (# #) . Oneway (Y (:v1) , X (:cv) , Wilcoxon Test (1)); kwallis v1, by (cv)
SAS	The group means or medians are in the same location across all levels of the grouping variable	Adjusted for ties	- One-tailed, asymptotic <i>p</i> -value - <i>n</i> _{Total} - ΣR for each level of the grouping variable - Kruskal-Wallis χ^2 test statistic - <i>df</i> - One-tailed, asymptotic <i>p</i> -value - Kruskal-Wallis χ^2 test statistic, adjusted for ties	
Stata	Several samples are drawn from the same rank-based population distribution	No correction; Adjusted for Ties	- One-tailed, asymptotic <i>p</i> -value - <i>n</i> _{Total} - ΣR for each level of the grouping variable - Kruskal-Wallis χ^2 test statistic - <i>df</i> - One-tailed, asymptotic <i>p</i> -value - Kruskal-Wallis χ^2 test statistic, adjusted for ties	kruskal.test (v1, cv)
R	The location parameters of the distribution are the same in each sample	No correction	- One-tailed, asymptotic <i>p</i> -value - <i>n</i> _{Total} - ΣR for each level of grouping variable - Wilcoxon's Mann-Whitney U - <i>n</i> , <i>M</i> _{<i>R</i>} , and ΣR for each level of grouping variable - <i>n</i> _{Total} - Mann-Whitney U statistic - Wilcoxon W statistic - <i>z</i> -statistic - Two-tailed, asymptotic <i>p</i> -value - Exact <i>p</i> -value (2*one-tailed <i>p</i> -value, if <i>n</i> is "not too large")	
SPSS	The two samples are from populations with the same distribution function	Adjusted for ties; Exact (not adjusted for ties) ^a	- Two-tailed, asymptotic <i>p</i> -value - Exact <i>p</i> -value (2*one-tailed <i>p</i> -value, if <i>n</i> is "not too large")	NEPAR TESTS /M-W=v1 BY cv (# #) .

Table 3 (continued)

SAS	Means or medians of ranked responses are equal across both levels of the grouping variable; the location of the two distributions are the same	Continuity-corrected	<ul style="list-style-type: none"> - One-way plot of responses across all levels of the grouping variable - n, ΣR, expected ΣR, M_R, and z_R for each level of the grouping variable - Wilcoxon W test statistic^c (labeled as "S") - Continuity-corrected z-statistic - Two-tailed, asymptotic p-value - One-way χ^2 test statistic <ul style="list-style-type: none"> - df - One-tailed, asymptotic p-value 	<pre> One-way (Y (: v1) , X (: cv) , Wilcoxon Test (1)); </pre> <p><i>Note.</i> Exact statistics can be computed in JMP Pro using the Nonparametric > Exact Test > Wilcoxon Exact Test function ranksum v1, by (cv)</p> <p><i>Additional options</i> exact: Compute exact p-value in addition to asymptotic p-value for $n \leq 1000$</p> <pre> wilcox.test (v1 , cv) </pre> <p><i>Additional options</i> correct=FALSE: Do not apply continuity correction</p>
Stata	The two independent samples are drawn from populations with the same distribution	Adjusted for ties; Exact ^a	<ul style="list-style-type: none"> - n, ΣR, and expected ΣR for each level of grouping variable - Unadjusted variance - Adjustment for ties - Adjusted variance - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Exact p-value (Hill & Peto, 1972; if $n \leq 200$) - Mann-Whitney U statistic (labeled as "W") - Continuity-corrected, two-tailed p-value (if $n \geq 50$ or ties are present) - Exact p-value (Bauer, 1972; if $n < 50$ and no ties are present) 	<pre> NPAR TESTS /WILCOXON=v1 WITH v2 (PAIRED) . </pre> <p>Matched Pairs (Y (: v1 , : v2) , Wilcoxon Signed Rank (1));</p> <pre> signrank v2 = v1 </pre> <p><i>Additional options</i> exact: Compute exact p-value in addition to asymptotic p-values (for $n \leq 2000$ only)</p>
R	The two sample distributions are the same; the true location shift is equal to 0	Adjusted for ties and continuity-corrected OR Exact ^a	<ul style="list-style-type: none"> - n, M_R, and ΣR for positive and negative differences - The number of zero differences (incorrectly labeled as "ties") - n_{total} - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Plot of paired differences - M of each pair - M, SE, and 95% confidence limits for paired difference - n_{total} - Pearson's correlation coefficient - Paired samples t-statistic <ul style="list-style-type: none"> - df - One- and two-tailed p-values - Wilcoxon T statistic (labeled as "S")^d <ul style="list-style-type: none"> - One- and two-tailed, asymptotic p-values - Exact p-value (if $n \leq 20$) - n, ΣR, and expected ΣR for positive, negative, zero difference, and total pairs - Unadjusted variance - Adjustment for ties - Adjustment zero differences - Adjusted variance - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Exact p-value (Baker & Tilbury, 1993; if $n \leq 200$) 	<pre> wilcox.test (v1 , v2 , paired = TRUE , exact = TRUE , conf.level = 0.95 , plot = TRUE , main = "Paired Differences") </pre>
SPSS	The median of the distribution of the matched pairs of observations is equal to zero	Adjusted for Ties and Zero differences	<ul style="list-style-type: none"> - n, M_R, and ΣR for positive and negative differences - The number of zero differences (incorrectly labeled as "ties") - n_{total} - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Plot of paired differences - M of each pair - M, SE, and 95% confidence limits for paired difference - n_{total} - Pearson's correlation coefficient - Paired samples t-statistic <ul style="list-style-type: none"> - df - One- and two-tailed p-values - Wilcoxon T statistic (labeled as "S")^d <ul style="list-style-type: none"> - One- and two-tailed, asymptotic p-values - Exact p-value (if $n \leq 20$) - n, ΣR, and expected ΣR for positive, negative, zero difference, and total pairs - Unadjusted variance - Adjustment for ties - Adjustment zero differences - Adjusted variance - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Exact p-value (Baker & Tilbury, 1993; if $n \leq 200$) 	<pre> NPAR TESTS /WILCOXON=v1 WITH v2 (PAIRED) . </pre> <p>Matched Pairs (Y (: v1 , : v2) , Wilcoxon Signed Rank (1));</p> <pre> signrank v2 = v1 </pre> <p><i>Additional options</i> exact: Compute exact p-value in addition to asymptotic p-values (for $n \leq 2000$ only)</p>
SAS	The distribution of paired differences is symmetric around zero	Adjusted for ties and zero differences; Exact ^a	<ul style="list-style-type: none"> - n, M_R, and ΣR for positive and negative differences - The number of zero differences (incorrectly labeled as "ties") - n_{total} - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Plot of paired differences - M of each pair - M, SE, and 95% confidence limits for paired difference - n_{total} - Pearson's correlation coefficient - Paired samples t-statistic <ul style="list-style-type: none"> - df - One- and two-tailed p-values - Wilcoxon T statistic (labeled as "S")^d <ul style="list-style-type: none"> - One- and two-tailed, asymptotic p-values - Exact p-value (if $n \leq 20$) - n, ΣR, and expected ΣR for positive, negative, zero difference, and total pairs - Unadjusted variance - Adjustment for ties - Adjustment zero differences - Adjusted variance - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Exact p-value (Baker & Tilbury, 1993; if $n \leq 200$) 	<pre> wilcox.test (v1 , v2 , paired = TRUE , exact = TRUE , conf.level = 0.95 , plot = TRUE , main = "Paired Differences") </pre>
Stata	The distributions of matched pairs of observations are the same; the distribution of matched pair differences is symmetrical and has a median equal to zero	Adjusted for ties and zero differences; Exact ^a	<ul style="list-style-type: none"> - n, M_R, and ΣR for positive and negative differences - The number of zero differences (incorrectly labeled as "ties") - n_{total} - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Plot of paired differences - M of each pair - M, SE, and 95% confidence limits for paired difference - n_{total} - Pearson's correlation coefficient - Paired samples t-statistic <ul style="list-style-type: none"> - df - One- and two-tailed p-values - Wilcoxon T statistic (labeled as "S")^d <ul style="list-style-type: none"> - One- and two-tailed, asymptotic p-values - Exact p-value (if $n \leq 20$) - n, ΣR, and expected ΣR for positive, negative, zero difference, and total pairs - Unadjusted variance - Adjustment for ties - Adjustment zero differences - Adjusted variance - z-statistic <ul style="list-style-type: none"> - Two-tailed, asymptotic p-value - Exact p-value (Baker & Tilbury, 1993; if $n \leq 200$) 	<pre> wilcox.test (v1 , v2 , paired = TRUE , exact = TRUE , conf.level = 0.95 , plot = TRUE , main = "Paired Differences") </pre>

Table 3 (continued)

R	The distribution of paired differences is symmetric around μ ; the true location shift is equal to 0	Adjusted for ties and continuity-corrected OR Exact ^a	<ul style="list-style-type: none"> - Wilcoxon T statistic (labeled as "V") - Continuity-corrected, two-tailed p-value (if $n \geq 50$ or ties are present) - Exact p-value (Bauer, 1972; if $n < 50$ and no ties are present) 	<p>wilcox.test(v1, v2, paired=TRUE)</p> <p><i>Additional options</i> correct=FALSE; Do not apply continuity correction</p>
SPSS	Sample estimate	Skewness and Kurtosis	<ul style="list-style-type: none"> - n - Skewness (Bliss, 1967) - Kurtosis (Cramér, 1946) - Frequency table with row and cumulative % - Histogram - Quantile report - n, M, SD, SEM, and 95% limits for variable - Skewness (Bliss, 1967) - Kurtosis (Cramér, 1946) 	<pre>FREQUENCIES /VARIABLES=v1 /STATISTICS=SKEWNESS KURTOSIS. Distribution (Continuous Distribution (Column (:v1), Customize Summary Statistics (Skewness (1), Kurtosis (1),)); summarize v1, detail</pre>
SAS	Sample estimate			
Stata	Population estimate	<ul style="list-style-type: none"> - Percentile report - Four smallest and largest values - n, Σn_{rs}, M, SD, and variance of the variable - Skewness (Cook & Campbell, 1979; Cramér, 1946) - Kurtosis (Bock, 1975) 	<p><i>Note.</i> Users may download a free, web-based Stata package (moments2) to estimate skewness or kurtosis using user-specified formulas from Joanes and Gill (1998). <i>No built-in functionality</i></p>	
R				<p><i>Note.</i> Users may download free, web-based packages (e1071 or moments) to estimate skewness or kurtosis using user-specified formulas from Joanes and Gill (1998).</p>

f_0 = observed frequency; df = degrees of freedom; $v1$ = variable 1; $v2$ = variable 2; M_R = mean rank; cv = categorical variable; ΣR = sum of ranks; z_R = standardized rank

^aExact statistics are only computed by default if certain criteria are met (see description in statistical output column)

^bDefined as $(n^3-n)(1-r)/6$, where n is the number of observations and r is Pearson's correlation coefficient

^cDefined as the sum of ranks for the level with fewer observations, or for the second level, if observations are equal across levels

^dBased on the sum of the absolute value of the signed ranks

Table 4 Results obtained across packages compared with hand calculations for the large-*N* analyses

Test	Algorithm	Statistic	Hand calculation	SPSS	SAS	Stata	R
Pearson’s chi-square		<i>df</i>	1	1	1	1	1
	Asymptotic	χ^2	0.168	0.168	0.168	0.168	–
		<i>p</i>	0.682	0.682	0.682	0.682	–
	Continuity	χ^2	0.119	0.119	–	–	0.119
		<i>p</i>	0.731	0.731	–	–	0.731
	Exact	<i>p</i>	–	0.696	0.697	–	–
Spearman’s rho	Ties	r_s	0.591	0.591	0.591	0.591	0.591
		<i>p</i>	0.000	0.000	0.000	0.000	0.000
Kruskal–Wallis		<i>df</i>	7	7	7	7	7
	Asymptotic	χ^2	9.508	–	–	9.508	9.509
		<i>p</i>	0.218	–	–	0.218	0.218
	Ties	χ^2	9.508	9.508	9.509	9.508	–
<i>p</i>		0.218	0.218	0.218	0.218	–	
Wilcoxon Mann–Whitney <i>U</i>		<i>W</i>	256,248	244,252	244252^a	–	–
		<i>U</i>	129,492	120,499	–	–	129492 ^b
	Ties	<i>z</i>	1.001	–1.001	–	1.001	–
		<i>p</i>	0.317	0.317	–	0.317	–
	Continuity	<i>z</i>	1.001	–	–1.001	–	–
		<i>p</i>	0.317	–	0.317	–	0.317
	<i>T</i>	211,006	–	–39244^a	–	211006 ^c	
Wilcoxon signed-rank	Ties	<i>z</i>	–4.296	–4.296	–	–4.296	–
		<i>p</i>	0.000	0.000	0.000	0.000	–
	Continuity	<i>z</i>	–4.296	–	–	–	–
		<i>p</i>	0.000	–	–	–	0.000
Skewness	Population		7.463	–	–	7.463	–
	Sample		7.474	7.474	7.474	–	–
Kurtosis	Population		127.313	–	–	127.313	–
	Sample		124.943	124.943	124.943	–	–

n = 1000. Results are organized according to the default algorithms used by each program, as noted in the respective output and/or help files/documentation. Bold values indicate results that are inconsistent with the hand-calculated results, or those that are inconsistent between packages, in the case of exact *p*-values. The values reported from “asymptotic” algorithms are those that were computed with no correction for ties or continuity applied. Dashes indicate values that were not provided in the default output produced by the respective statistical package or those that were not able to be produced by hand (i.e., exact *p*-values)

^aTest statistic labeled as “*S*”

^bTest statistic labeled as “*W*”

^cTest statistic labeled as “*V*”

it is of note that for sample sizes less than 50, R will produce an exact *p*-value, calculated using the algorithm described by Bauer (1972) with a Hodges and Lehmann (1963) estimation, in place of the asymptotic *p*-value.

Wilcoxon signed-rank

Similar to the procedures applied to the Wilcoxon Mann–Whitney *U*, while all packages adjust for ties when performing the Wilcoxon signed-rank test, R is the only package to also apply a continuity correction, by default; however, this did not affect the *p*-values produced by R to a meaningful amount, as the only

discrepancy was observed for the *p*-value derived from the small-*N* analysis, which was smaller by a value of .001 in R, relative to the values produced by SPSS, SAS, and Stata. Additionally, the algorithm used to calculate the Wilcoxon *T* statistic in SAS differs from the generally accepted definition provided by Wilcoxon (1945), where the test statistic is the smaller of either the sum of ranks with positive differences or the sum of ranks with negative differences. This definition is used by SPSS, Stata, and R; however, the test statistic produced by SAS is based on the sum of the absolute value of signed ranks.

Table 5 Results obtained across packages compared with hand calculations for the medium-*N* analyses

Test	Algorithm	Statistic	Hand calculation	SPSS	SAS	Stata	R
Pearson's chi-square		<i>df</i>	1	1	1	1	1
	Asymptotic	χ^2	1.040	1.040	1.040	1.040	–
		<i>p</i>	0.308	0.308	0.308	0.308	–
	Continuity	χ^2	0.767	0.767	–	–	0.767
		<i>p</i>	0.381	0.381	–	–	0.381
	Exact	<i>p</i>	–	–	0.318	0.318	–
Spearman's rho	Ties	r_s	0.686	0.686	0.686	0.686	0.686
		<i>p</i>	0.000	0.000	0.000	0.000	0.000
Kruskal–Wallis		<i>df</i>	7	7	7	7	7
	Asymptotic	χ^2	11.344	–	–	11.344	11.344
		<i>p</i>	0.124	–	–	0.124	0.124
	Ties ^a	χ^2	11.344	11.344	11.344	11.344	–
		<i>p</i>	0.124	0.124	0.124	0.124	–
	Wilcoxon Mann–Whitney <i>U</i>		<i>W</i>	9685.5	10,414.5	9685.5 ^b	–
		<i>U</i>	5220.5	4743.5	–	–	5220.5 ^c
Ties ^a		<i>z</i>	0.598	–0.598	–	0.598	–
		<i>p</i>	0.550	0.550	–	0.550	–
Continuity		<i>z</i>	0.597	–	0.597	–	–
		<i>p</i>	0.551	–	0.551	–	0.551
Exact	<i>p</i>	–	–	–	0.551	–	
Wilcoxon signed-rank		<i>T</i>	8484	–	–1566^b	–	8484 ^d
	Ties ^a	<i>z</i>	–1.911	–1.911	–	–1.911	–
		<i>p</i>	0.056	0.056	0.056	0.056	–
	Continuity	<i>z</i>	–1.911	–	–	–	–
		<i>p</i>	0.056	–	–	–	0.056
	Exact	<i>p</i>	–	–	–	0.056	–
Skewness	Population		2.964	–	–	2.964	–
	Sample		2.986	2.986	2.986	–	–
Kurtosis	Population		17.670	–	–	17.670	–
	Sample		15.075	15.075	15.075	–	–

n = 200. Results are organized according to the default algorithms used by each program, as noted in the respective output and/or help files/documentation. Bold values indicate results that are inconsistent with the hand-calculated results, or those that are inconsistent between packages, in the case of exact *p*-values. The values reported from “asymptotic” algorithms are those that were computed with no correction for ties or continuity applied. Dashes indicate values that were not provided in the default output produced by the respective statistical package or those that were not able to be produced by hand (i.e., exact *p*-values)

^aNo ties present

^bTest statistic labeled as “*S*”

^cTest statistic labeled as “*W*”

^dTest statistic labeled as “*V*”

Skewness and kurtosis

Finally, unbiased calculations of sample skewness and excess sample kurtosis (where the expected value for kurtosis of a sample with a normal distribution is 0) are only used by SPSS and SAS. Despite the bias inherent in the use of population estimates, Stata does not provide any other built-in option for the way in which these measures of normality are estimated for a sample of data. Additionally, it

is important to note that R has no built-in functionality for calculating skewness or kurtosis.

Simulations

The results of the simulation studies shed further light on the impact of algorithmic variations on the results of nonparametric analyses between software packages (tables with the simulations results can be found in our

Table 6 Results obtained across packages compared with hand calculations for the small-*N* analyses

Test	Algorithm	Statistic	Hand calculation	SPSS	SAS	Stata	R
Pearson’s chi-square		<i>df</i>	1	1	1	1	1
	Asymptotic	χ^2	0.164	0.164	0.164	0.164	–
		<i>p</i>	0.685	0.685	0.685	0.685	–
	Continuity	χ^2	0.041	0.041	–	–	0.041
		<i>p</i>	0.839	0.839	–	–	0.839
	Exact	<i>p</i>	–	0.840	0.840	–	–
Spearman’s rho	Ties	r_s	0.523	0.523	0.523	0.523	0.523
		<i>p</i>	0.000	0.000	0.000	0.000	0.000
Kruskal–Wallis		<i>df</i>	5	5	5	5	5
	Asymptotic	χ^2	5.139	–	–	5.139	5.139
		<i>p</i>	0.399	–	–	0.399	0.399
	Ties ^a	χ^2	5.139	5.139	5.139	5.139	–
<i>p</i>		0.399	0.399	0.399	0.399	–	
Wilcoxon Mann–Whitney <i>U</i>		<i>W</i>	2648	2402	2402^b	–	–
		<i>U</i>	1373	1127	–	–	1373 ^c
	Ties ^a	<i>z</i>	0.861	–0.861	–	0.861	–
		<i>p</i>	0.389	0.389	–	0.389	–
	Continuity	<i>z</i>	0.858	–	–0.858	–	–
		<i>p</i>	0.391	–	0.391	–	0.391
Exact	<i>p</i>	–	–	–	0.391	–	
Wilcoxon signed-rank		<i>T</i>	2100	–	–425^b	–	2100 ^d
	Ties ^a	<i>z</i>	–1.461	–1.461	–	–1.461	–
		<i>p</i>	0.144	0.144	0.145	0.144	–
	Continuity	<i>z</i>	–1.463	–	–	–	–
		<i>p</i>	0.143	–	–	–	0.144
	Exact	<i>p</i>	–	–	–	0.145	–
Skewness	Population		0.696	–	–	0.696	–
	Sample		0.707	0.707	0.707	–	–
Kurtosis	Population		3.171	–	–	3.171	–
	Sample		0.242	0.242	0.242	–	–

n = 100. Results are organized according to the default algorithms used by each program, as noted in the respective output and/or help files/documentation. Bold values indicate results that are inconsistent with the hand-calculated results, or those that are inconsistent between packages, in the case of exact *p*-values. The values reported from “asymptotic” algorithms are those that were computed with no correction for ties or continuity applied. Dashes indicate values that were not provided in the default output produced by the respective statistical package or those that were not able to be produced by hand (i.e., exact *p*-values)

^aNo ties present

^bTest statistic labeled as “*S*”

^cTest statistic labeled as “*W*”

^dTest statistic labeled as “*V*”

OSF storage at <https://osf.io/35umb/>). Of particular importance is the extent to which the *p*-values differ between uncorrected and continuity-corrected *p*-values produced for Pearson’s chi-square. Across all 30 simulations, the difference in continuity-corrected versus uncorrected *p*-values ranged from .005 to .162 (*M* = .082, *SD* = .050). These differences were more extensive for the smaller samples (Fig. 3), where the results for the small-*N* simulations differed by an average of .123 (*SD* = .048), and

those of the large-*N* simulations differed by an average of .033 (*SD* = .018). While continuity-corrected *p*-values were the same as uncorrected *p*-values produced for the Wilcoxon tests in the large-*N* simulations, continuity-corrected *p*-values for the Wilcoxon Mann–Whitney *U* test were greater by a value of .001 on 7 of the 10 medium-*N* simulations, and by an average value of .002 (*SD* = .001) for all of the small-*N* simulations.

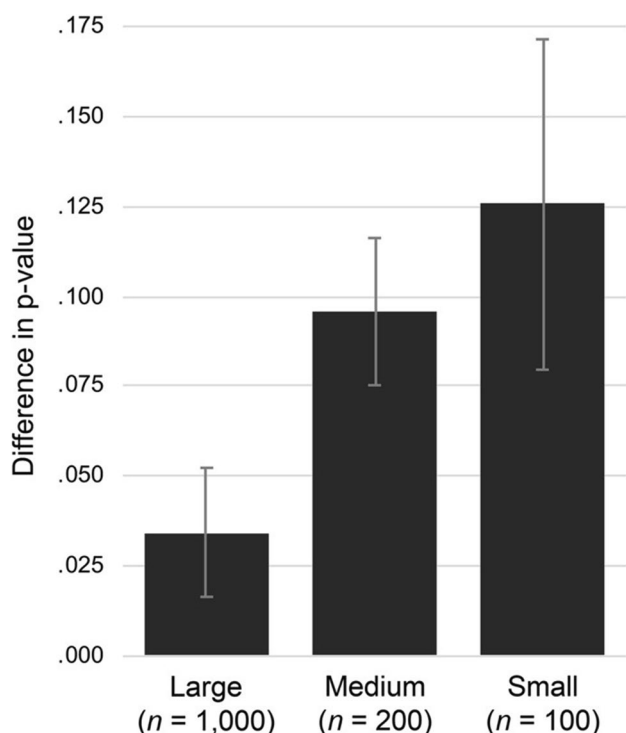


Fig. 3 Average difference in Yates' continuity-corrected versus uncorrected asymptotic p -values for Pearson's chi-square test across large, medium, and small sample sizes. *Note.* Large $n=1000$; medium $n=200$; small $n=100$; average values are based on differences (continuity-corrected – uncorrected) in two-tailed p -values across primary and simulated datasets, with the standard deviation of the differences reflected in the error bars.

Computational error

Few inconsistencies due to computational error were observed between the hand-calculated results and those produced by the statistical packages, and all appeared in the results from the simulated datasets. These inconsistencies could not be reduced to rounding error when the same computational procedures were used for calculation (according to the documentation provided by the software) and were therefore determined to be produced by the statistical package in error. The most prevalent error occurred in several of the results provided by SPSS for the Wilcoxon Mann–Whitney U test across the large-, medium-, and small- N simulations, where Wilcoxon's W was computed as the sum of ranks for the first sample. The documentation clearly states that the sum of ranks for the second sample is used to calculate Wilcoxon's W , and there is no apparent pattern as to when this deviation from the documented procedure would occur. Although the results produced matched those obtained through hand calculation, we consider them to be due to computational error, as the method used to obtain these findings is not consistent with that which is described

in the package's documentation. Additionally, SPSS and R failed to compute the test statistic and continuity-corrected p -value for Pearson's chi-square test on two of the simulated medium- N datasets and one of the simulated small- N datasets due to the presence of one cell with zero observations in the 2×2 contingency table; the exact p -values produced by SPSS and SAS for these three simulated datasets were also affected by this (i.e., $p = 1.000$).

Statistical output

Extensive differences were observed in the statistical output produced for each nonparametric test across the four statistical packages, and the output was consistent within each package across the primary and simulated analyses. Each package produces a variety of statistical output in addition to the specific test statistics that are typically required to be reported and interpreted in the literature (Table 3); see the Supplemental Material for details of the statistical output provided across packages. With regard to the inferential statistics that are necessary for one to make basic conclusions about the hypothesis that was tested (i.e., the test statistics, degrees of freedom, and p -values), several inconsistencies continued to be seen across SPSS, SAS, Stata, and R.

Pearson's chi-square and Kruskal–Wallis

While test statistics, degrees of freedom, and p -values were reported by all packages for Pearson's chi-square and Kruskal–Wallis tests, only SPSS provided these both with and without continuity correction for Pearson's chi-square, and only Stata provided these both with and without an adjustment for ties for the Kruskal–Wallis test. Fisher's exact one- and two-tailed p -values are reported with Pearson's chi-square results by SPSS and SAS, by default, for 2×2 contingency tables. Stata will also produce Fisher's exact p -values for any data if the option, `exact`, is included in the command. Across the primary and simulated datasets, exact two-tailed p -values reported by SPSS and SAS differed from uncorrected and continuity-corrected asymptotic p -values as a function of sample size (Fig. 4). On average, exact two-tailed p -values were greater than uncorrected asymptotic p -values by .021 ($SD = .015$), .049 ($SD = .031$), and .055 ($SD = .046$) for the large-, medium-, and small- N analyses, respectively. In contrast, exact two-tailed p -values were smaller than continuity-corrected asymptotic p -values obtained from the large-, medium-, and small- N datasets by values of $-.013$ ($SD = .010$), $-.047$ ($SD = .032$), and $-.071$ ($SD = .049$), respectively.

Wilcoxon Mann–Whitney U

In the output produced for the Wilcoxon Mann–Whitney U test, SPSS alone generated Wilcoxon's W , Mann–Whitney's

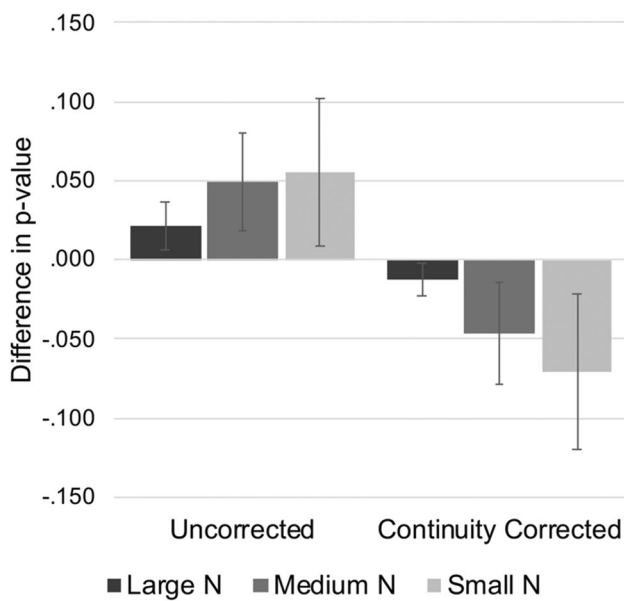


Fig. 4 Average differences in Fisher’s exact (two-tailed) versus uncorrected and continuity-corrected asymptotic p -values for Pearson’s chi-square test across large, medium, and small sample sizes. *Note.* Large $n = 1000$; medium $n = 200$; small $n = 100$; average values are based on differences (exact – asymptotic) in two-tailed p -values across primary and simulated datasets, with the standard deviation of the differences reflected in the error bars.

U , the asymptotic test statistic (z -statistic), and the asymptotic p -value. In contrast, the output from SAS and Stata did not include the Mann–Whitney U statistic, and while Wilcoxon’s W technically appears in the rank-sum frequency table, it is neither clearly indicated nor differentiated from the other rank-sum values reported in Stata’s output. Additionally, only the Mann–Whitney U statistic and the continuity-corrected p -value are provided in the output generated in R; however, the test statistic is incorrectly labeled as “ W .” Exact two-tailed p -values for the Wilcoxon Mann–Whitney U test are also reported by Stata for samples of $n \leq 200$ and therefore appear in our medium- and small- N results; exact p -values were very similar to asymptotic p -values across all medium- and small- N datasets. No other packages produced exact statistics for the Wilcoxon Mann–Whitney U test performed on our data; however, an exact two-tailed p -value is reported by R in place of the continuity-corrected p -value when $n < 50$ and no ties are present, and according to the documentation, SPSS reports an exact 2*one-tailed p -value along with the asymptotic p -value when n is “not too large.”

Wilcoxon signed-rank

Finally, none of the four packages reported all of the fundamental inferential statistics in their output for the Wilcoxon signed-rank test. While both SPSS and Stata report the

asymptotic z -statistic and p -value, neither reports the Wilcoxon T statistic (although it appears without a label in the rank-sum frequency table). In contrast, while the Wilcoxon T statistic and p -values are reported in the output of both SAS and R, neither of these packages reports the asymptotic z -statistic, and R incorrectly labels the test statistic as “ V .” An exact two-tailed p -value was generated by Stata for the medium- and small- N analyses, and this occurs by default for samples of $n \leq 200$; exact p -values were very similar to asymptotic p -values across all medium- and small- N datasets. No other exact statistics were generated for the Wilcoxon signed-rank test across packages; however, SAS will report an exact p -value when $n \leq 20$ and R will replace the asymptotic p -value with an exact p -value when $n < 50$ and no ties are present. It is important to note that SAS labels all test statistics as “ S ,” even though several test statistics are commonly defined by specific labels.

Discussion

In the current era of psychological science, much of the field’s attention has been directed at methodological practices and reproducible science. However, we have seen little evidence that this increased attention has brought much awareness to inconsistencies between statistical software packages. The present study investigated whether the same results would be generated from the same nonparametric statistical procedures when performed on the same dataset by different statistical software packages. Although results were largely consistent across software packages, our findings bring into question the extent to which reporting error and/or misinterpreted results are present in the psychological science literature. Our most notable findings were related to the extent of algorithmic variation that exists across packages for a given nonparametric analysis. The algorithms implemented are often native to each software package and do not require user specification, and they are not consistently applied within a single package for samples with varying characteristics. These adjustments included performing automatic corrections or transformations that, while useful in certain situations, are not always readily apparent to and may not be desirable by the user conducting the tests.

In our data, the algorithmic variations generally resulted in minor test statistic or p -value discrepancies that may not have substantial impacts on one’s statistical conclusions. The exception to this lies in the continuity-corrected versus uncorrected p -values generated for Pearson’s chi-square analysis (Fig. 3). By default, uncorrected p -values are generated by SAS and Stata, and Yates’ continuity correction is applied to the calculation of the chi-square statistic in R, resulting in a larger p -value. The only p -value produced by SAS and Stata was smaller than the only p -value produced

by R by an average value of .083 across our primary and simulated datasets. The differences in continuity-corrected and uncorrected p -values ranged from .005 to .162, largely as a function of sample size. In addition, Fisher's exact p -values are provided with these results by some packages, depending on the characteristics of the data. When provided with the results of our analyses, exact p -values were larger than uncorrected asymptotic p -values but smaller than continuity-corrected asymptotic p -values, and these differences also varied as a function of sample size (Fig. 4). The discrepancies seen here demonstrate the potential for statistical conclusions to be influenced by the software package used, especially when users are unaware of the underlying algorithm used or naïve as to when/if a continuity correction or exact statistic is appropriate. This has important implications, given that Pearson's chi-square is among the most commonly used nonparametric tests in the psychological science literature. Furthermore, in cases where multiple p -values are provided for a given test, a naïve user may be inclined to report the value that supports their hypothesis, especially if they are not sure which p -value is the most appropriate for their data.

Perhaps the most intriguing finding, however, is the computational error we encountered from SPSS, where the program would unpredictably calculate the Wilcoxon Mann–Whitney U test statistic based on the rank sum of the first sample, when the documentation clearly states that the second sample is used for calculation. Although this has no effect on the p -value and should not impact conclusions of statistical significance, the inverted z -statistic may result in confusion over the direction of an effect.

Overall, our findings suggest that the results of several common nonparametric tests and measures of normality differ when performed on the same data across SPSS, SAS, Stata, and R. These differences were driven primarily by the inconsistent application of default algorithmic procedures across software packages. Furthermore, our medium- and small- N analyses revealed that, as we approach sample sizes more commonly used in the current literature, discrepancies in statistical output are increased, due to the application of continuity corrections or calculation of exact statistics.

We also found that some statistical output generated from these nonparametric tests included unnecessary information. We believe that programs producing unnecessary and ambiguous output may increase confusion and/or selective reporting of misleading results in the literature, especially when there is no clear indication or description of the values generated. In particular, we argue that confusion is more likely when a student, psychological science researcher, or other non-statistician investigator attempts to interpret the results of a single test when multiple p -values are present in the generated output. This may be especially relevant when algorithms used to generate the different p -values are not

clearly indicated in the output or defined in the documentation. At worst, when multiple p -values are provided for a single test, we fear that publication bias and the current “publish or perish” mentality in scientific research may lead to selective reporting of whichever p -value most closely aligns with the hypothesis. Taken together, these factors and discrepancies may contribute to the replication failure in the social and behavioral science literature.

Overall, although we found that there were no discrepancies across statistical packages that would result in an incorrect conclusion from our data, the findings do suggest that within statistics packages, poor documentation and labeling of output may lead to an incorrect conclusion. For example, we found large differences in p -values between continuity-corrected and uncorrected p -values. It is possible that a researcher may select the smaller of the two p -values, either purposefully, due to unclear labeling of the output, or because it is the only p -value a given statistical package produces. As can be seen in Figs. 3 and 4, differences in p -values that occur due to algorithmic variations can be large, especially in smaller datasets.

Limitations and future directions

The current investigation has noteworthy limitations. Although we found differences in the results generated across software packages, we want to encourage caution when generalizing these findings to other programs and other versions of the software. Software developers (especially those working on open-access software, like R) are continually updating their programs and correcting bugs in the software. Many of the errors we found may be due to bugs in the software that have not yet been discovered or due to differences in the output selected by the program; thus, the findings discussed herein are specific to the programs we compared and the versions of these programs we used for the present analyses. We encourage readers to consider the context in which the present discrepancies across results were found and to use the present findings as motivation to more closely consider their own statistical results. It is also important to note that the inconsistent findings considered to be due to computational error may not be the result of miscalculation *per se*; these results were simply referred to as computational error because they did not reflect the true result when the computational procedures described in the packages' documentation were followed. It is likely that the procedures were incorrectly described in the documentation, and in such cases, these inconsistencies would be better described as misspecifications than error.

Our results may not generalize to other, more commonly used statistical procedures, such as parametric analyses. A large contributor to the heterogeneity we found in results across statistical packages was due to algorithmic variation,

particularly those related to continuity corrections and the calculation of exact p -values. These procedures are not relevant in parametric analyses; thus, there may be fewer ways in which the results of parametric statistics could differ across software packages. It is also important that our findings are replicated using other meaningful data so as to determine the extent to which the discrepancies seen here may impact results from data of more practical or clinical importance. Further, we strongly encourage future work to leverage systematic and meta-analytic methods to examine the existing literature in terms of statistical rigor, researcher degrees of freedom, and methodology involved when conducting nonparametric analysis, as such work would greatly benefit ongoing replication efforts. Additionally, a similar consideration of the most frequently used statistical methods in the general psychological science literature is needed. While we conducted a cursory review of nonparametric methods used in the recent psychological science literature, a meta-analysis or systematic review documenting researcher degrees of freedom in methods used when performing nonparametric analysis would provide a clearer understanding of the current state of the problem in our field.

Bootstrapping is a technique used to replicate the results of an analysis a large number of times across simulated datasets given a chosen distribution. The technique is growing in popularity and is frequently being included in the latest statistical software. We did not conduct our simulations by pulling the formulas from each statistical software platform and testing them within the program itself to examine how they perform comparatively, and the lack of such simulation procedures conducted in this manner may be a limitation. Still, we did not include this simulation as we argue that it is beyond the scope of the current paper (i.e., we can address the question of whether there are differences in output across software packages without conducting a large simulation in R). Further, although using bootstrapping may address some limitations of relying on a single analysis in a single statistical software package (e.g., by producing simulated confidence intervals), we caution using bootstrapping as a “fix” for discrepancies across statistical software. Bootstrapping analyses may be prone to the same variations in computations, formulas, labeling, and output across statistical software that the analyses examined in the current paper produced. Thus, our recommendation for cross-examining results applies similarly to bootstrapping and other simulation techniques.

Finally, we would like to note the open-source nature of R that separates it from the other statistical packages evaluated herein. Any user can create libraries that can be downloaded and used by others for their analyses. User-written libraries do not undergo any rigorous testing or vetting processes; thus, there is a risk in utilizing libraries outside of the 14 that are built into the program, as they may be prone

to error. Furthermore, the documentation for these user-written libraries ranges from comprehensive to none. Given these limitations, if a researcher only uses R with third-party libraries to analyze their data, cross-examination between R and other established statistical software platforms may be especially important.

Implications and recommendations

The present results raise questions about previous replication efforts. It is well known within the social and behavioral sciences that efforts to replicate many landmark studies have failed (Open Science Collaboration, 2015). It is possible that algorithmic variation or other differences in the computation or reporting of results between statistical software packages may be contributing to this problem, and future replication efforts should consider this possibility and focus on determining whether failures to replicate may be due, at least in part, to differences between the software used. The heterogeneity we have demonstrated in our results speaks to the importance of open-access science; without knowing the underlying procedures used to generate statistical results, it may be difficult to directly replicate previous studies. Replication efforts within the social and behavioral sciences have made significant strides toward promoting open-access science, but work remains to be done.

The results of the current study suggest that researchers may benefit from cross-examining their results across statistical software platforms. We recommend making a consistent practice of analyzing the results from one statistical software platform to at least one other platform. We argue that this practice may reduce the chances of drawing inaccurate conclusions due to inconsistencies or errors in the output, labeling, or underlying algorithms used by a given software package. At the time of this publication, Stata appeared to be the most consistent with our hand calculations. However, the differences described in the current paper are likely to be resolved or altered through software updates, so this may not always be the case. Furthermore, consistency with hand calculation does not necessarily mean one platform produced more accurate results than another, but that the algorithms used to compute the results were more consistent with our expectations, based on original formulas provided in textbooks and other literature. Finally, if inconsistencies arise in a cross-examination, we recommend checking documentation and conducting hand calculations to best determine the root of the discrepancies. Thus, we argue that the choice to cross-examine one’s results across software platforms and to review the formulas provided in the software’s documentation are more important than the choice of software. Once researchers have chosen which software to use, they should provide a justification for their choice in their manuscript.

We advocate for fully open-access science, including public hosting of datasets, command syntax used to generate results, and all study materials used in the final analysis. New organizations, such as the Open Science Framework (OSF), have made considerable progress in the promotion of open-access science by emphasizing the preregistration of scientific studies, where researchers may report all study procedures prior to beginning data collection and publish the results upon study completion. Additionally, OSF has data-hosting options that allow researchers to post their datasets and analysis pipelines for future replication efforts. This level of transparency in science remains promising, as more publishers require the documentation of study procedures and as academia begins to encourage preregistration and sharing of all study data and materials. We suggest that future studies continue this trend of transparency.

We strongly encourage all researchers, regardless of career stage, to carefully examine the documentation in software packages, and to ensure that the package selected is performing the test as they expect it to. Furthermore, users should be aware of the computational idiosyncrasies of the software used to conduct the statistical analyses, and it is recommended that the results are compared across platforms to avoid reporting errors or misleading results that are due to algorithmic variation, computational error, or other output-related characteristics of a given program. Of utmost importance is the necessity for adequate reporting of the statistical procedures used when analyzing data and full transparency when publishing the results of such analyses.

We also suggest that future work consider the use of outside services to address issues, such as version control, containerization, and ambiguous documentation of code/dependencies, which may cause future replications to fail. GitHub is a popular third-party service that allows for the hosting of version-controlled code. Similarly, Docker is another service that constructs a virtual environment (referred to as a “container”) that allows individuals to share not only their code, but also all of the dependencies and software needed to run the code. The use of containers eliminates any ambiguity that may arise concerning file execution order or the dependencies required to replicate an analysis. Further, we strongly encourage researchers to consider clearly annotating any code, scripts, or other documentation shared publicly. Commenting at each step of a script not only increases clarity and transparency, but will assist future replication efforts in understanding the steps of the analysis that were taken. An excellent example of the level of transparency required is provided by Peikert and colleagues’ (2021) tutorial on accomplishing these tasks in R.

Some of the burden of open-access science lies on the shoulders of the software developers. Several issues were present in the help files and documentation provided by the software packages, where the information provided for

the various analyses we conducted was often lacking sufficient information about a given test (e.g., R) or about the algorithms used to produce the results (e.g., Stata), and in some cases, the documentation was not easily accessible (e.g., SAS) or was difficult to interpret (e.g., SPSS). Ambiguity in documentation could play a role in the inconsistent replication of scientific experiments that has been noted in recent years, and it is essential that the software developers are made aware of the importance of clear, exhaustive, and easily accessible documentation for these packages.

Conclusion

The results of our study demonstrate the unreliability of results produced for nonparametric tests and measures of normality across SPSS, SAS JMP Pro, Stata, and R. The inconsistent results were primarily due to differences in the default algorithms used; however, computational error and differences in statistical output also contributed to the unreliability of results. These discrepancies, along with unnecessary and/or unclearly defined statistical output generated for a given test, may contribute to confusion, selective reporting or reporting error, and ultimately, replication failure. We urge researchers to refer to documentation when interpreting results (when it is available), compare their statistical results across software platforms, and describe all statistical procedures, including the statistical software package used, when publishing results, to aid in the success of future replication efforts.

Acknowledgments The authors would like to thank Logan Kowallis, PhD (Department of Psychology, Brigham Young University) and Elizabeth Hovenden (Department of Neurology, University of Utah) for their contribution to the preparation of this manuscript.

References

- Alder, A. G., & Vollick, D. (2000). Undergraduate statistics in psychology: A survey of Canadian institutions. *Canadian Psychology/Psychologie Canadienne*, *41*(3), 149–151.
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., & Nosek, B. A. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119.
- Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavioral Research Methods*, *43*(3), 666–678. <https://doi.org/10.3758/s13428-011-0089-5>
- Bauer, D. F. (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, *67*, 687–690. <https://doi.org/10.1080/01621459.1972.10481279>
- Bergmann, R., Ludbrook, J., & Spooren, W. P. J. M. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. *The American Statistician*, *54*(1), 72–77. <https://doi.org/10.1080/00031305.2000.10474513>

- Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p -values in two psychiatry journals. *International Journal of Methods in Psychiatric Research*, 16(4), 202–207. <https://doi.org/10.1002/mpr.225>
- Bliss, C. I. (1967). *Statistics in biology, 1*, McGraw-Hill.
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. McGraw-Hill.
- Borghi, J. A., & Van Gulick, A. E. (2018). Data management and sharing in neuroimaging: Practices and perceptions of MRI researchers. *PLoS One*, 13(7), Article e0200562. <https://doi.org/10.1371/journal.pone.0200562>
- Brown, B. L., Hendrix, S. B., Hedges, D. W., & Smith, T. B. (2012). *Multivariate analysis for the biobehavioral and social sciences: A graphical approach*. John Wiley & Sons.
- Campbell, I. (2007). Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, 26, 3661–3675. <https://doi.org/10.1002/sim.2832>
- Caperos, J. M., & Pardo Merino, A. (2013). Consistency errors in p -values reported in Spanish psychology journals. *Psicothema*, 25(3), 408–414. <https://doi.org/10.7334/psicothema2012.207>
- Chambers, C. (2019). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin.
- Cramér, H. (1946). *Mathematical methods of statistics*. Princeton University Press.
- Dar, R., Serlin, R. C., & Omer, H. (1994). Misuse of statistical test in three decades of psychotherapy research. *Journal of Consulting and Clinical Psychology*, 62(1), 75–82. <https://doi.org/10.1037//0022-006x.62.1.75>
- Epskamp, S. (2019). Reproducibility and replicability in a fast-paced methodological world. *Advances in Methods and Practices in Psychological Science*, 2(2), 145–155.
- Friedrich, J., Childress, J., & Cheng, D. (2018). Replicating a National Survey on statistical training in undergraduate psychology programs: Are there “new statistics” in the new millennium? *Teaching of Psychology*, 45(4), 312–323. <https://doi.org/10.1177/0098628318796414>
- 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., & Núñez-Antón, V. (2020). Asymptotic versus exact methods in the analysis of contingency tables: Evidence-based practical recommendations. *Statistical Methods in Medical Research*, 29(9), 2569–2582. <https://doi.org/10.1177/0962280220902480>
- Garside, G. R., & Mack, C. (1976). Actual type I error probabilities for various tests in the homogeneity case of the 2×2 contingency table. *The American Statistician*, 30, 18–21.
- Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference* (5th ed.). Taylor & Francis Group.
- Grieder, S., & Steiner, M. (2020). *Algorithmic jingle jungle: A comparison of implementations of Principal Axis Factoring and pro-max rotation in R and SPSS*. PsyArXiv. <https://doi.org/10.31234/osf.io/7hwrw>
- Grizzle, J. E. (1967). Continuity correction in the χ^2 -test for 2×2 tables. *The American Statistician*, 21(4), 28–32.
- Haber, M. (1982). The continuity correction and statistical testing. *International Statistical Review*, 50, 135–144.
- Hill, I. D., & Peto, R. (1971). Algorithm AS 35: Probabilities derived from finite populations. *Applied Statistics*, 20, 99–105.
- Hitchcock, D. B. (2009). Yates and contingency tables: 75 years later. *Electronic Journal for History of Probability and Statistics*, 5, 1–14.
- Hodges, J. L., & Lehmann, E. L. (1963). Estimation of location based on ranks. *Annals of Mathematical Statistics*, 34(2), 598–611. <https://doi.org/10.1214/aoms/1177704172>
- Hoekstra, R., Kiers, H. A., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, 3, 137. <https://doi.org/10.3389/fpsyg.2012.00137>
- Huang, Y., & Bentler, P. M. (2015). Behavior of asymptotically distribution free test statistics in covariance versus correlation structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4), 489–503. <https://doi.org/10.1080/10705511.2014.954078>
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), 0696–0701. <https://doi.org/10.1371/journal.pmed.0020124>
- Ioannidis, J. P. (2014). How to make more published research true. *PLoS Medicine*, 11(10), Article e1001747. <https://doi.org/10.1371/journal.pmed.1001747>
- Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *The Statistician*, 47, 183–189.
- Keeling, K. B., & Pavur, R. J. (2007). A comparative study of the reliability of nine statistical software packages. *Computational Statistics & Data Analysis*, 51(8), 3811–3831.
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, 68(3), 350–386.
- Köhler, T., Landis, R. S., & Cortina, J. M. (2017). From the editors: Establishing methodological rigor in quantitative management learning and education research: The role of design, statistical methods, and reporting standards. *Academy of Management Learning and Education*, 16(2), 173–192. <https://doi.org/10.5465/amle.2017.0079>
- Lane, K. A., Banaji, M. B., Nosek, B. A., & Greenwald, A. G. (2007). Understanding and using the implicit association test: What we know (so far) about the method. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 59–102). Guilford Press.
- Lehmann, E. L. (1998). *Nonparametrics: Statistical methods based on ranks* (revised 1st ed.). Prentice Hall.
- Levine, T. R., & Atkin, C. (2004). The accurate reporting of software-generated p -values: A cautionary research note. *Communication Research Reports*, 21(3), 324–327. <https://doi.org/10.1080/08824090409359995>
- Lydersen, S., Fagerland, M. W., & Laake, P. (2009). Recommended tests for association in 2×2 tables. *Statistics in Medicine*, 28, 1159–1175.
- Mantel, N. (1976). The continuity correction. *The American Statistician*, 30, 103–104.
- Maxwell, E. A. (1976). Analysis of contingency tables and further reasons for not using Yates correction in 2×2 tables. *Canadian Journal of Statistics*, 4, 277–290.
- McCoach, D. B., Rifken, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A. J., & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, 43(5), 594–627.
- McCullough, B. (2000). Is it safe to assume that software is accurate? *International Journal of Forecasting*, 16(3), 349–357.
- McCullough, B. D., & Heiser, D. A. (2008). On the accuracy of statistical procedures in Microsoft excel 2007. *Computational Statistics & Data Analysis*, 52(10), 4570–4578.








- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153–157.
- Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), Article 0021. <https://doi.org/10.1038/s41562-016-0021>
- Mundry, R., & Fischer, J. (1997). Use of statistical programs for nonparametric tests of small samples often leads to incorrect *p* values: Examples from animal behaviour. *Animal Behaviour*, *56*, 256–259.
- Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests*. Unwin Hyman Ltd..
- Nuijten, M. B., Borghuis, J., Veldkamp, C. L., Dominguez-Alvarez, L., Van Assen, M. A., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra. Psychology*, *3*(1).
- Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657–660.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. <https://doi.org/10.1126/science.aac4716>
- Osborne, J. (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational Psychology*, *28*, 151–160.
- Osborne, J., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, & Evaluation*, *8*, Article 2.
- Oster, R. A., & Hilbe, J. M. (2008a). An examination of statistical software packages for parametric and nonparametric data analyses using exact methods. *The American Statistician*, *62*(1), 74–84. <https://doi.org/10.1198/000313008X268955>
- Oster, R. A., & Hilbe, J. M. (2008b). Rejoinder to “an examination of statistical software packages for parametric and nonparametric data analyses using exact methods”. *The American Statistician*, *62*(2), 173–176. <https://doi.org/10.1198/000313008X306853>
- Pearson, E. S. (1947). The choice of statistical test illustrated on the interpretation of data classed in a 2 x 2 table. *Biometrika*, *34*, 139–167.
- Peikert, A., & Brandmaier, A. M. (2021). A reproducible data analysis workflow with R Markdown, Git, Make, and Docker. *Quantitative and Computational Methods in Behavioral Sciences*, Article e3763. <https://doi.org/10.5964/qcumb.3763>
- Potvin, C., & Roff, D. A. (1993). Distribution-free and robust statistical methods: Viable alternative to parametric statistics? *Ecology*, *74*(6), 1617–1628.
- Prescott, R. J. (2019). Two-tailed significance tests for 2 × 2 contingency tables: What is the alternative? *Statistics in Medicine*, *38*, 4264–4269.
- Richardson, J. T. E. (1990). Variants of chi-square for 2 × 2 contingency tables. *British Journal of Mathematical and Statistical Psychology*, *43*, 309–326.
- Schatz, P., Jay, K. A., McComb, J., & McLaughlin, J. R. (2005). Misuse of statistical tests in archives of clinical neuropsychology publications. *Archives of Clinical Neuropsychology*, *20*(8), 1053–1059. <https://doi.org/10.1016/j.acn.2005.06.006>
- Siegel, S. (1957). Nonparametric statistics. *The American Statistician*, *11*(3), 13–19.
- Siegel, S., & Castellan, N. J. (1988). *Non-parametric statistics for the behavioural sciences* (2nd ed.). McGraw-Hill.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Stefanescu, C., Berger, V. W., & Hershberger, S. (2005). Yates’s continuity correction. In B. S. Everit & D. Howell (Eds.), *Book Yates’s continuity correction* (Vol. 4, pp. 2127–2129). John Wiley & Sons.
- Trisovic, A., Lau, M. K., Pasquier, T., & Crosas, M. (2022). A large-scale study on research code quality and execution. *Scientific Data*, *9*(60), 1–16. <https://doi.org/10.1038/s41597-022-01143-6>
- Wang, J., & Johnson, D. E. (2019). An examination of discrepancies in multiple imputation procedures between SAS® and SPSS®. *The American Statistician*, *73*(1), 80–88. <https://doi.org/10.1080/00031305.2018.1437078>
- Whitley, E., & Ball, J. (2002). Statistics review 6: Nonparametric methods. *Critical Care*, *6*, 509–513. <https://doi.org/10.1186/cc1820>
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking [review]. *Frontiers in Psychology*, *7*(1832). <https://doi.org/10.3389/fpsyg.2016.01832>
- Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.
- Xu, K., Nosek, B., & Greenwald, A. (2014). Psychology data from the race implicit association test on the project implicit demo website. *Journal of Open Psychology Data*, *2*(1), e3. <https://doi.org/10.5334/jopd.ac>
- Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, *1*(2), 217–235. <https://doi.org/10.2307/2983604>

Open Practices Statement

All datasets used in the present study, the scripts or syntax used for all analyses, all relevant help files/documentation provided by each software package, and logs of the full statistical output generated from each software platform can be found in our OSF storage directory at <https://osf.io/35umb/>. This study was not preregistered.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Cooper B. Hodges^{1,2,3,4}  · Bryant M. Stone⁵  · Paula K. Johnson^{1,6}  · James H. Carter III⁷  · Chelsea K. Sawyers⁸  · Patricia R. Roby⁹  · Hannah M. Lindsey^{1,2} 

¹ Department of Neurology, University of Utah School of Medicine, Salt Lake City, UT, USA

² Department of Psychology, Brigham Young University, Provo, UT, USA

³ Department of Physical Medicine and Rehabilitation, Virginia Commonwealth University, Richmond, VA, USA

⁴ School of Social and Behavioral Sciences, Andrews University, Berrien Springs, MI, USA

⁵ Department of Psychology, Southern Illinois University, Carbondale, 1125 Lincoln Drive, Carbondale, IL 62901, USA

⁶ Neuroscience Center, Brigham Young University, Provo, UT, USA

⁷ Department of Psychology, Stanford University, Stanford, CA, USA

⁸ Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, USA

⁹ Center for Injury Research and Prevention, Children's Hospital of Philadelphia, Philadelphia, PA, USA