Pitfalls of Statistical Methods in Traffic Psychology

de Winter, J.C.F.; Dodou, D.

DOI
10.1016/B978-0-08-102671-7.10665-7

Publication date
2021

Document Version
Final published version

Published in
International Encyclopedia of Transportation

Pitfalls of Statistical Methods in Traffic Psychology

J.C.F. de Winter, D. Dodou, Delft University of Technology, Delft, The Netherlands

Introduction

In traffic psychology, researchers typically conduct an experiment, the data of which are subsequently analyzed. For example, a researcher may conduct a driving simulator experiment with different time-budget conditions to evaluate how drivers reclaim control from an SAE Level 3 automated vehicle. For each driver and time-budget condition, scores on measures are obtained, such as the maximum absolute steering wheel angle, self-reported criticality of the situation, and take-over time. In other cases, the traffic psychologist may investigate individual differences, for example, by analyzing accident records or data obtained from Internet-based questionnaires.

The available experimental data are invariably subjected to statistical tests. We analyzed the papers of 129 experiments on automation-to-manual take-overs (list of papers available in a meta-analysis by Zhang et al., 2019) and found that *P*-values were used in as much as 95% of the 129 studies. ANOVAs (65% of studies), usually combined with post hoc (Bonferroni-corrected) *t*-tests, as well as paired/unpaired *t*-test presented separately (46% of studies), were among the most common statistical procedures. Nonparametric test variants (e.g., Wilcoxon test) were common as well. In areas of traffic psychology that are concerned with individual differences, on the other hand, correlation coefficients and data reduction techniques, such as exploratory factor analysis, are typically used (**Fig. 1**).

When we were asked to contribute an article about statistics in traffic psychology, we realized that there are already many articles and books that detail how statistical tests should be conducted, sometimes accompanied by an extensive step-by-step plan tailored to a specific software package (e.g., Field, 2013). It does not appear scientifically valuable to repeat such information. What appears valuable, however, is to report on common pitfalls and misconceptions regarding statistics in the field of traffic psychology.

This article was written based on the first author's subjective experiences as a researcher and peer reviewer in the field. The pitfalls were selected based on estimates of risk, where risk is a function of how often these pitfalls occur and how severe their consequences are. There are, of course, many other pitfalls of statistical methods, but the list below is regarded to be of direct relevance to the traffic psychology researcher.

The *P*-Value is Not the Same as the Probability That the Null Hypothesis is True

In a large proportion of papers in traffic psychology, statistical tests are performed. These tests yield *P*-values, based on which statements are made about hypotheses. Formally, the *P*-value indicates how extreme one's observations are under the assumption of a null hypothesis. A critical thing that tends to be overlooked in null-hypothesis significance testing is that *P*-values do not allow for making direct claims about whether the null hypothesis is true or false.

A thought experiment can prove the point. Suppose one investigates psychic abilities, such as whether participants can predict the outcome of a random number generator (see Bem, 2011). This hypothesis is certainly intriguing and spectacular if true. However, it is impossible to be correct, given the current knowledge of physics, and the fact that, as far as we know, no person in the world can systematically outplay the roulette table at casinos. Any significant *P*-value ($P < 0.05$) must therefore be a false positive (see Wagenmakers et al., 2011 for a similar argumentation).

The above message is illustrated in **Fig. 2**. Let us assume that, in a given subfield of traffic psychology, as much as 95% of the hypotheses are false. That is, there is only a 5% probability that the effect one hopes to find exists in reality. In such a case, even if assuming a large sample size and corresponding statistical power of 80%, then still only 46% of the statistically significant effects reflect a true effect. In other words, it is more likely that a significant effect ($P < 0.05$) is a false positive than a true positive.

This type of Bayesian reasoning corresponds to the argumentation of John Ioannidis in his well-known work "Why most published research findings are false" (Ioannidis, 2005). **Fig. 2** provides an example where statistical power is reasonably high

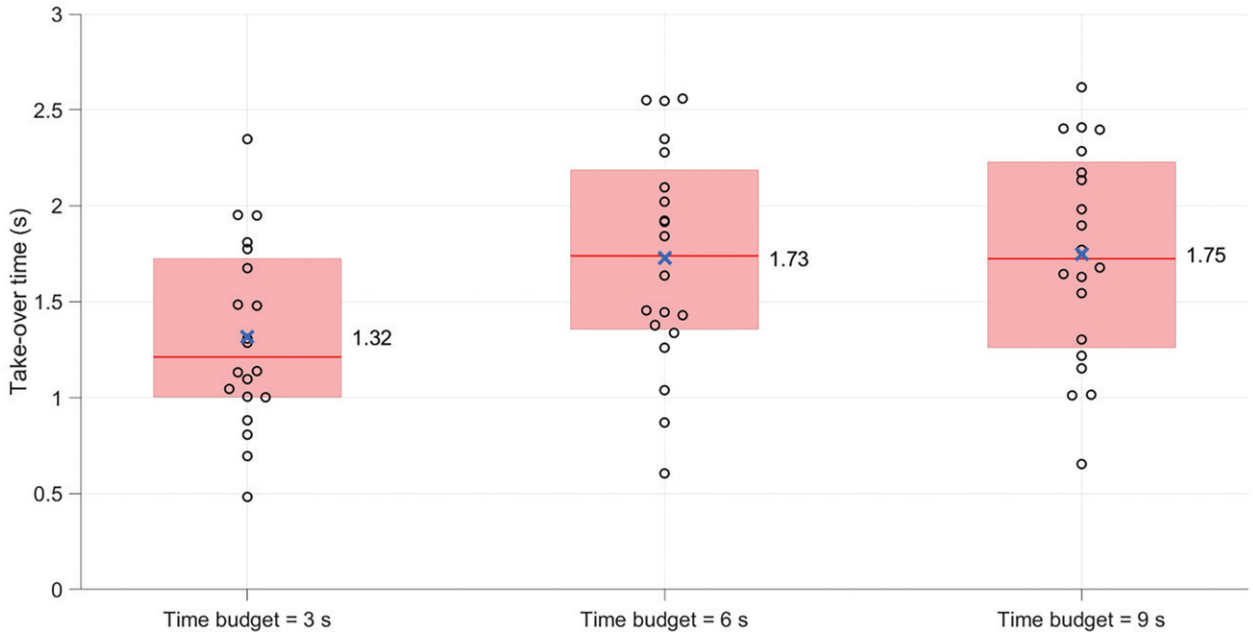

Figure 1 Example (simulated) data for an experiment on automation-to-manual take-overs, with three time-budget conditions. A sample size of 20 per condition was assumed. The blue cross and the number next to each bar indicate the group mean.

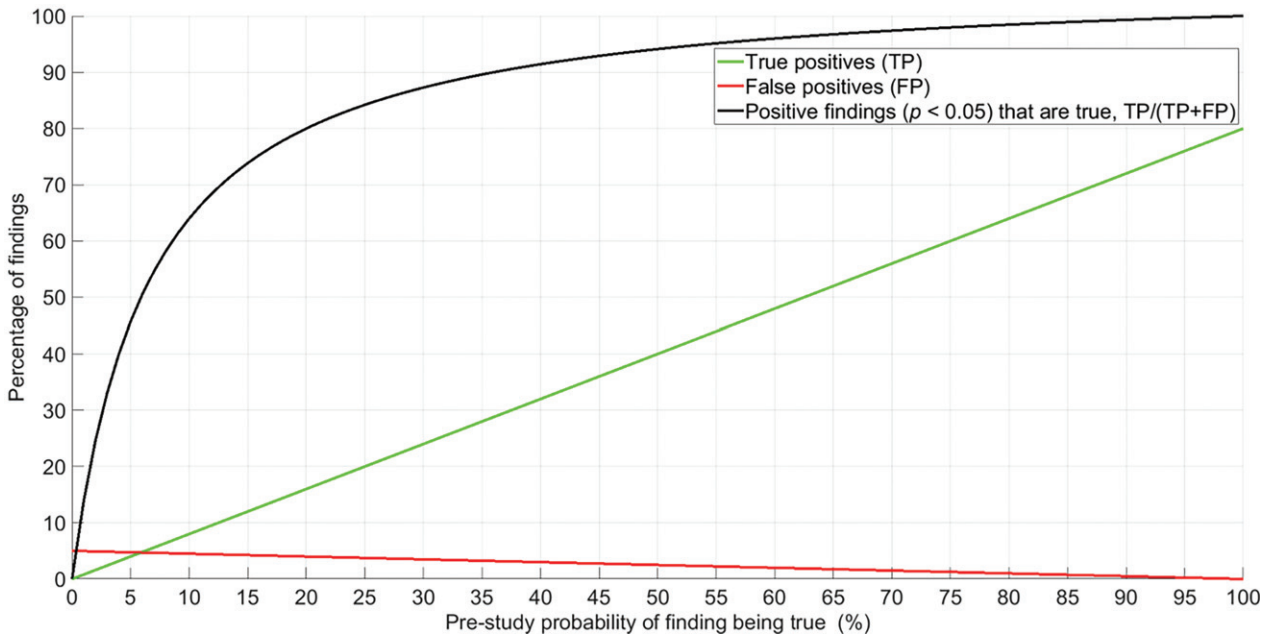

Figure 2 Illustration of the effect of the pre-study probability that a research finding is true for statistical power ($1 - \beta$) = 80%. For a significance level (α) of 5% and a pre-study probability of 5%, only 46% of the detected statistically significant effects are true positives.

(80%). Using the same figure, it can be understood why low statistical power (e.g., small expected effects or small sample sizes) reduces the probability that a research finding is true. For example, if power were 60% instead of 80%, then there would be fewer true positives, and, as a result, only 39% of the positive (i.e., statistically significant) results would be true positives. In other words, low statistical power does reduce not only the probability of detecting an effect but also the likelihood that an observed significant effect is a true effect (Button et al., 2013). Note that the example in Fig. 2 does not even consider the phenomenon of bias, where researchers may unconsciously or consciously distort research findings because researchers generally like to find significant effects (Ware and Munafò, 2015). Bias is discussed in the third section of this article.

In traffic psychology, effects are often well established and replicable, because experimental conditions are clearly operationalized (e.g., human–machine interface conditions, traffic conditions in a driving simulator), and the driver's inputs (e.g., steering

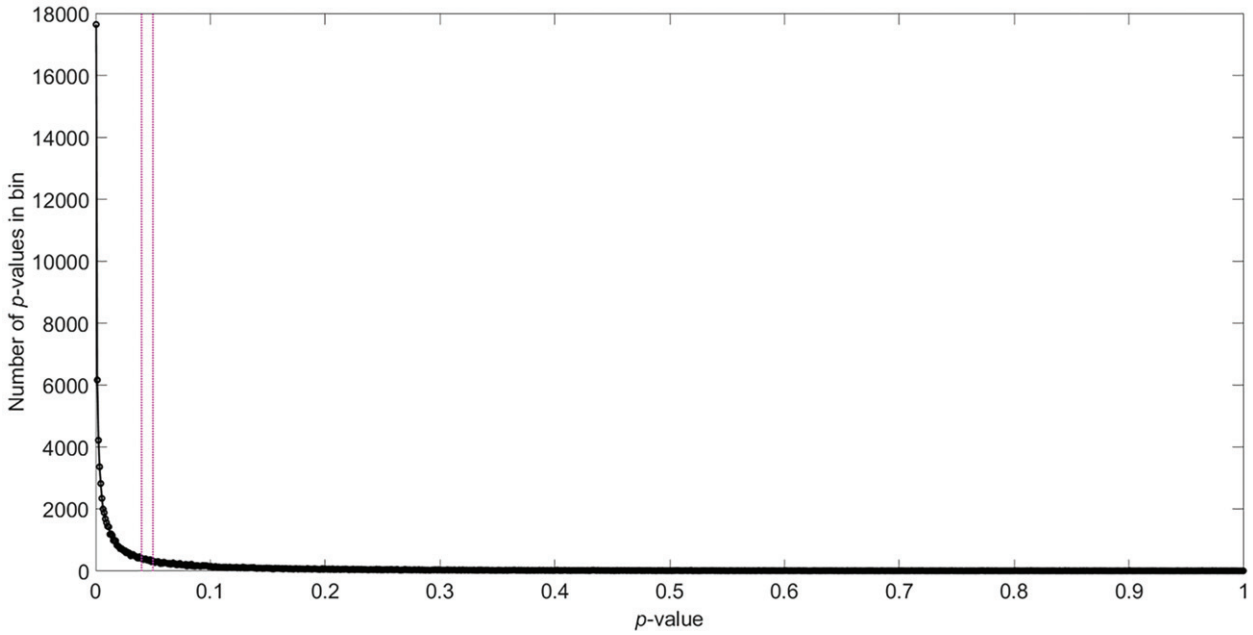

Figure 3 Distribution of *P*-values (so-called *P*-curve) when submitting two samples ($n = 20$ each), with Cohen's $d = 0.8$, to an independent-samples *t*-test. The vertical dashed lines are drawn at $P = 0.04$ and $P = 0.05$, respectively. The bins in this figure are 0.001 wide.

wheel angle, throttle and brake inputs) are easily measurable and causally related to the task conditions. For example, the effect of time budget and take-over request modality on take-over times are known to be strong and reproduced many times (see Zhang et al., 2019, for a meta-analysis). However, in subfields of traffic psychology, researchers may sometimes make rather spectacular, surprising, or intriguing claims. Examples may be the effects of emotional priming, mindfulness, or pleasant smells in the car on driving behavior. While the existence of such effects cannot be ruled out, the underlying mechanisms of how these would affect driving behavior are less clear, and hence the credibility of the findings is low. When these types of findings are obtained via small samples and supported by a barely significant effect (e.g., $P = 0.04$), scepticism is appropriate. Conversely, if an effect is supported by theory or the *P*-value is small (e.g., $P < 0.001$), then the finding is more likely to be replicable.

Incidentally, it should be noted that *P*-values below 0.05 are more prevalent in the literature than they should be. It is well established that significant findings ($P < 0.05$) are more likely to appear in papers than nonsignificant findings ($P > 0.05$), a phenomenon known as publication bias. Research also shows that the use of *P*-values is becoming increasingly common in the abstracts of papers (Chavalarias et al., 2016; De Winter and Dodou, 2015), and there is some evidence for "convenient" downwards rounding of *P*-values to be able to claim "$P \leq 0.05$" (Hartgerink et al., 2016).

The expected rarity of *P*-values in the $P = 0.04$–0.05 region can be illustrated using a computer simulation. Suppose one wishes to compare the means of two samples. The first sample is drawn from a population with a mean of 0 and a standard deviation of 1; the second sample is drawn from a population with a mean of 0.8 and also a standard deviation of 1. In other words, the true effect size is large (Cohen's $d = 0.8$). If one draws random samples of $n = 20$ each and performs 100,000 independent-samples *t*-tests, then most *P*-values will turn out to be very small. In this specific case, only 3.6% of the *P*-values are between 0.04 and 0.05 (see **Fig. 3** for the results of the simulation).

In summary, a *P*-value slightly below 0.05 does not imply that the alternative hypothesis is true; one should consider the plausibility of the hypothesis (Ioannidis, 2005). Furthermore, if there truly is a strong effect in the population, *P*-values just below 0.05 are rare, and small *P*-values (e.g., $P < 0.001$) are more likely.

Possible solutions to the above-mentioned problems are the use of larger sample sizes (Button et al., 2013) and the use of lower alpha values (e.g., 0.005 instead of 0.05; Benjamin et al., 2018; but see Lakens et al., 2018, for counterarguments). Some researchers in traffic psychology have used Bayesian statistics to counteract the problem of making wrong inferences based on null hypothesis significance testing (see Körber et al., 2016 for a study on Bayesian statistics in automation-to-manual take-overs). Although Bayesian statistics are popular in several areas of transportation research, its use in traffic psychology and other human-related fields is still limited (see Chavalarias et al., 2016 for a survey of biomedical papers).

Everything is Correlated

In traffic psychology, researchers sometimes have access to extremely large databases. For example, it has been found in a study of over 10,000 drivers that self-reported violations are predictive of self-reported accidents (Wells et al., 2008) and that demographic variables are predictive of the willingness to use a driverless shuttle (Nordhoff et al., 2018). Similarly, accident statistics of hundreds

of thousands of drivers have revealed interesting relationships between crash involvement, traffic tickets, and individual characteristics (Factor, 2014). For such large sample sizes, however, even extremely small effects are statistically significantly different from 0 (Nunnally, 1960; Standing et al., 1991), an inconvenience which Meehl (1990) referred to as the "crud factor" (p. 108).

The bottom line is that a statistically significant effect does not imply that this effect is important. For example, the correlation between self-reported violations and self-reported accidents is about 0.15 (De Winter et al., 2015), which may be interesting for testing theories about personal predispositions but probably not very useful for identifying accident-prone drivers. Colleague Evans (2009) once explained that, in principle, wearing a second t-shirt contributes to a reduction of road traffic deaths, as this extra layer of clothing has a protective effect in absorbing energy during a crash. However, as he pointed out, this, of course, does not mean that researchers should advise drivers to put on an extra t-shirt before entering their cars.

The above concerns about the meaninglessness of null hypothesis significance testing do not only apply to small effects; they also apply to strong effects, in particular in the context of human performance and behavior. A common phenomenon in the analysis of human performance, including driving, is that performance measures are almost always correlated. For example, a driver who is skilled in lane-keeping will probably also be skilled in car following, and a positive statistically significant correlation coefficient between performance measures (e.g., standard deviation of lateral position and standard deviation of headway) is virtually guaranteed. Horn and McArdle (2007) noted: "*practically every variable that in any sense indicates the good things in life—health, money, education, and so forth, and athletic and artistic abilities, as well as cognitive abilities—correlate positively with every other such thing that everything is significantly correlated*" (p. 219).

In questionnaire research as well, strong positive correlations between variables are the rule rather than the exception, due to method effects. This problem, which has been vigorously highlighted by a few authors in traffic psychology before (in particular Af Wåhlberg, 2010, 2017), seems to have fallen on deaf ears in a large portion of researchers.

The problem of methods effects in questionnaire research can be illustrated using a small computer simulation (see also De Winter et al., 2019). First, let us assume two constructs (e.g., driving errors and driving violations), which are uncorrelated in the population. Let us also assume that these constructs can be measured using a five-point Likert scale (e.g., 1 = never, 2 = hardly ever, 3 = occasionally, 4 = quite often, 5 = frequently). The distribution of the correlation coefficients, assuming a sample size of 300 respondents, is shown in **Fig. 4** (this distribution is based on 100,000 repetitions) and averages at $r = 0.00$, as expected. Next, one-third of the respondents (100 of 300) are simulated to have a different notion of quantity and the meaning of words such as "occasionally". Let us assume that these respondents report on average 2 points lower than a bias-free sample (i.e., they report "never" instead of "hardly ever" and "occasionally", "hardly ever" instead of "quite often", and "occasionally" instead of "frequently"). Additionally, one-third of the respondents report on average 2 points higher than a bias-free sample (i.e., more toward "frequently"). The distribution of the overall sample, yielding an average correlation between the two variables of 0.55, is provided in **Fig. 4**. What this simulation shows is that method effects alone, in this case, response style ("nay-saying", "yea-saying"), can cause strong correlations (see also De Winter and Dodou, 2017).

The problem illustrated here seems to be strikingly common in traffic psychology. One group of unnamed authors, for example, evaluated the acceptance of automated vehicles using the Unified Theory of Acceptance and Use of Technology (UTAUT). They

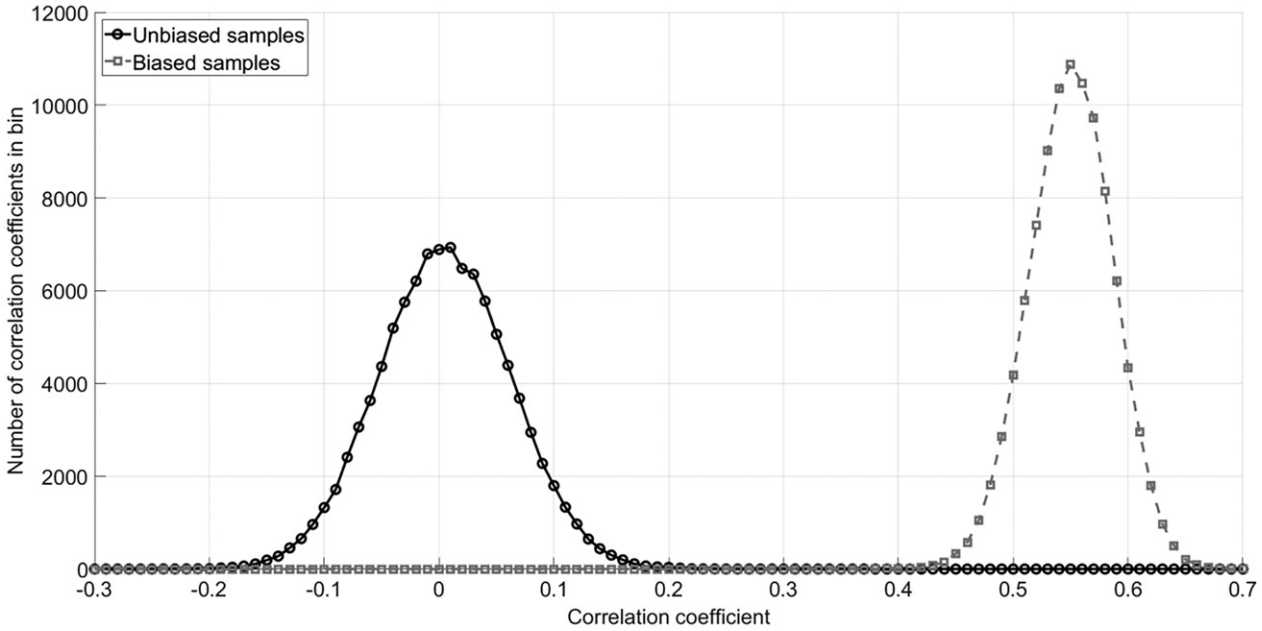

Figure 4 Distribution of Pearson product-moment correlation coefficients between two variables for a sample size of 300 when the correlation in the population is 0 (unbiased samples) and when the population correlation is 0 but response style contaminates the results (biased samples). The bins in this figure are 0.01 wide.

found positive correlations ($r = 0.5$ to 0.6, all P-values < 0.001) between behavioral intentions to use an automated vehicle, on the one hand, and effort expectancy, social influence, facilitating conditions, and hedonic motivation, on the other. These findings, although correct from the viewpoint of null hypothesis significance testing, do not imply that one can predict real-life behavioral intentions that strongly. The positive correlations probably reflect a method effect or a form of social desirability or self-esteem which translates into response style.

The bottom line is that effect sizes should be reported and interpreted. Researchers should not focus solely on P-values: even trivially small effects can yield highly significant P-values, and strong effects are, of course, statistically significantly different from 0, but may not reflect the true state of the world. Possible remedies to common method biases are to measure predictor and criterion variables at different points in time, the use of forced-choice survey items (Bartram, 2007; Brown and Maydeu-Olivares, 2011; Nederhof, 1985), and the use of a criterion variable that is recorded rather than self-reported (Af Wåhlberg, 2017).

Letting the Statistical Test Depend on a Statistical Test

In traffic psychology, it often happens that researchers pretest their data. For example, a researcher may test for normality before deciding whether to conduct a parametric t-test or ANOVA or a nonparametric Wilcoxon rank-sum test, Kruskal–Wallis test, or Friedman test. A problem with pretesting is that the "numbers don't remember where they came from" (Lord, 1953, p. 21), and that such tests may not have sufficient power for detecting violations of the assumptions of the follow-up tests (Rasch et al., 2011; Zimmerman, 2011).

Suppose a researcher is faced with two samples of data which he would like to statistically compare. The researcher decides to conduct a t-test or a Wilcoxon rank-sum test based on the result of a Shapiro–Wilk test of normality. The issue here is that, if the sample size is higher, the statistical power of the Shapiro–Wilk test will tend to be higher as well, and so the outcome of the Shapiro–Wilk test depends on sample size as well as the idiosyncratic nature of the dataset. It is much more efficient to decide beforehand which statistical test to use. For example, it is already well known that Driver Behaviour Questionnaire (DBQ) item data are heavily tailed (e.g., Mattsson, 2012). It may therefore be better to use tests which are robust to tailed distributions (e.g., Wilcoxon rank-sum tests, signed-rank tests, and Spearman's rank-order correlations). No pretest is needed if prior research already reveals which type of test is most appropriate.

A computer simulation could prove the point. Suppose one wishes to test whether the central tendency of two distributions is different. The first distribution has a mean of 0 and a standard deviation of 1, and the second distribution has a mean of 1 and a standard deviation of 1 (i.e., Cohen's $d = 1$). Furthermore, the sample size is 20 per group, and the first group features an outlier. More specifically, the value for the 20th participant in the first group equals 4 (i.e., the mean + 4 standard deviations).

Table 1shows the results of a simulation study of a two-stage process, with 100,000 repetitions. That is, 100,000 times, 20 samples were drawn from the first distribution, and 20 samples were drawn from the second distribution, and the two samples were compared using an independent-samples t-test and a Wilcoxon rank-sum test, as well as a conditional test. In the conditional test, an independent-samples t-test was conducted when a Shapiro–Wilk test used on the first sample yielded $P \geq 0.05$, and a Wilcoxon rank-sum test was conducted when the same Shapiro–Wilk test yielded $P < 0.05$.

Table 1 reveals that the statistical power is higher for the Wilcoxon rank-sum test than for the t-test, which can be explained by the outlier. After all, it is well established that the t-test is not particularly robust to outliers. The conditional approach improves statistical power substantially compared to the independent-samples t-test but is still less powerful than the Wilcoxon rank-sum test.

Using an insurance policy analogy (Anscombe, 1960), it can be wise to choose a nonparametric test such as a Wilcoxon rank-sum test or a Welch's t-test (Lakens, 2015a) and not use a pretesting approach. Such a test yields a slight loss of power compared to the regular t-test if the normality assumption is met. However, the insurance premium is only small for the high protection it offers against outliers or unequal variances.

More damaging outcomes can occur when the researcher chooses the statistical test based on statistical significance. For example, a researcher may decide to remove outliers and subsequently perform the statistical test again, or may first perform a regular t-test, and, if the result is not significant, perform a Wilcoxon-rank-sum test. Dropping or merging of experimental conditions, optional stopping, or adding covariates/moderator variables (e.g., age, gender) are similar examples. These types of so-called questionable research practices, which can dramatically increase false-positive rates, have been previously identified in psychological research (Bakker and Wicherts, 2014; Nelson et al., 2018), and may well be exacerbated in traffic psychology, an area that has been progressing in parallel to mainstream psychology and engineering.

Table 1 Results of a simulation study for detecting a significant difference between two groups in the presence of an outlier in one of the two groups

	Statistical power (% of 100,000 repetitions yielding $P < 0.05$)
Independent-samples t-test	56.7%
Wilcoxon rank-sum test	72.3%
Conditional test	66.9%

Note. The Shapiro–Wilk test yielded a significant effect in only 60.5% of the 100,000 repetitions.

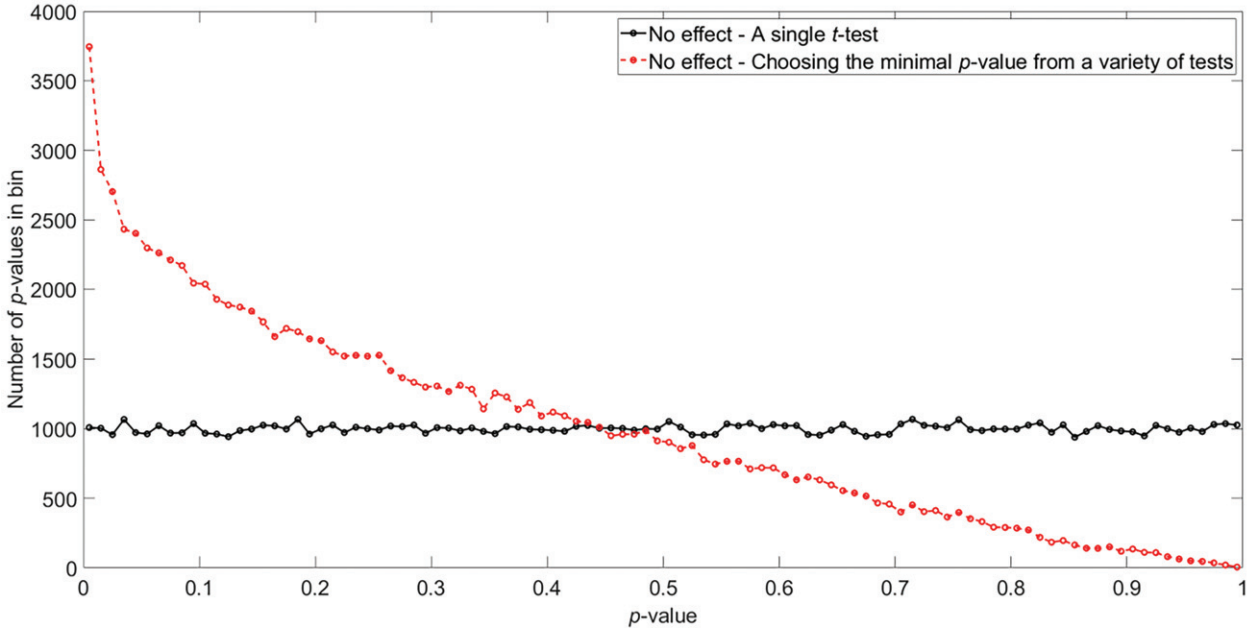

Figure 5 *P*-value distribution (*P*-curve) in case there is no effect in the population. By "trying out" various statistical test and reporting the "best" effect, the false positive rate may increase substantially (from 5.0% to 14.1% in this simulation). The bins in this figure are 0.01 wide.

Fig. 5 illustrates the result of independent-samples *t*-tests in black. In this case, two samples with equal means and standard deviations were compared, $n = 20$ per group. The results of this simulation, performed with 100,000 repetitions, show that the distribution of the *P*-value is uniform (see also Lakens, 2015b).

The red line in Fig. 5 shows the results of statistical tests for the same data, but here we simulated researchers who engaged in questionable research practices. More specifically, the simulated researchers all performed an independent-samples *t*-test, a paired-samples *t*-test, an independent-samples *t*-test after automated outlier removal, and an independent samples *t*-test for subgroups. The latter could be done, for example, if the samples consist of 10 males and 10 females, and the researchers test for a difference in means for males and females separately. If the researchers use the minimum *P*-value among all five options, the *P*-values become lower than they should be. In this case, the false-positive rate has become 14.1%, that is, 14.1% of the 100,000 cases yielded $P < 0.05$.

In traffic psychology, a popular method is to perform a stepwise regression analysis. A stepwise analysis works by including statistically significant predictors only and removing variables when they are not significant, and suffers from similar problems as the problems as illustrated in Fig. 5. Further critique of stepwise regression analysis is provided by Antonakis and Dietz (2011) and Thompson (1989), amongst others.

In summary, the statistical test should not depend on another statistical test: Two-stage approaches can yield suboptimal statistical power and, more worryingly, peeking at the data and adjusting the statistical test accordingly increases the prevalence of false positives. Methods such as preregistration of one's hypotheses (Nosek et al., 2018) can help prevent the issues mentioned in this section.

Violation of Independence

It often happens that the traffic psychologist computes correlation coefficients to explore relationships between variables collected in the experiment. For example, the researcher may be interested in the relationship between take-over time and the maximum absolute steering wheel angle. Fig. 6 shows a scatter plot of take-over time and maximum absolute steering wheel angle, as could have been obtained from a simulator-based experiment. In this case, the association between the two variables is moderate and not significantly different from 0 ($r = 0.27$, $P = 0.247$).

Typically, experiments contain multiple repetitions of the same scenario. For example, there may have been five identical take-over trials per time budget condition. In traffic psychology, the first author has seen it happening on numerous occasions that researchers inadvertently treat those repeated measures as independent. For the sake of argument, let us assume that participants are consistent with respect to themselves. This is a reasonable assumption in behavioral research, where in the long run, participants exhibit stable behavioral patterns (excluding learning, mood swings, etc.). Thus, it is assumed that the 20 participants output practically the same take-over time and the same maximum absolute steering wheel angle for each of the identical five take-over trials.

The corresponding scatter plot is drawn in Fig. 7; it contains the same data as in Fig. 6, except that there are now 100 data points because the repeated measures for the same persons are plotted separately, whereas the markers in Fig. 6 were assumed to reflect the

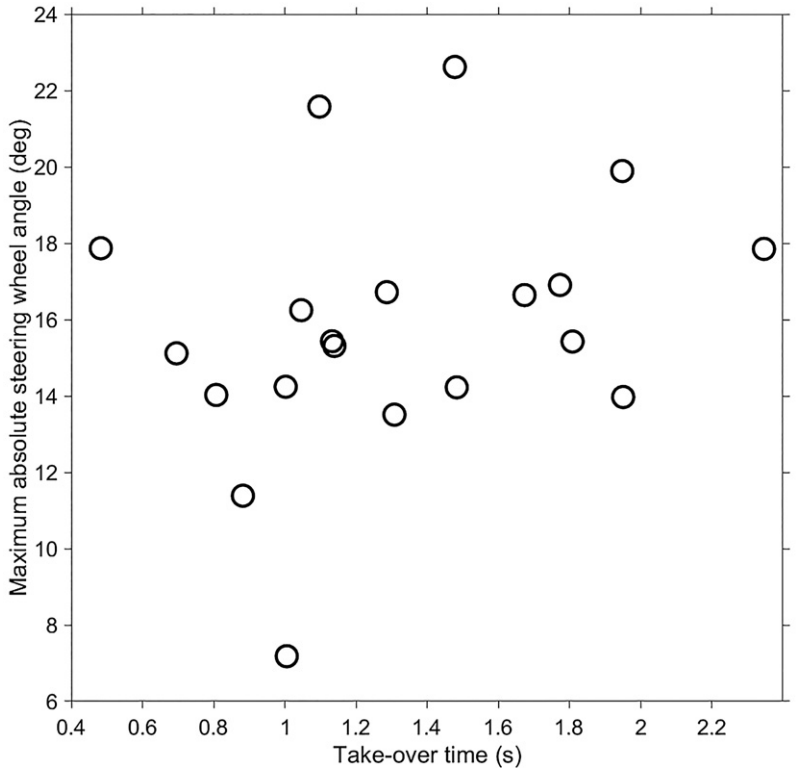

Figure 6 Example (simulated) data for an experiment on automation take-overs. The horizontal axis shows the take-over time (same data as the data in **Fig. 1** for the 3 s time budget condition), and the vertical axis shows the maximum absolute steering wheel angle.

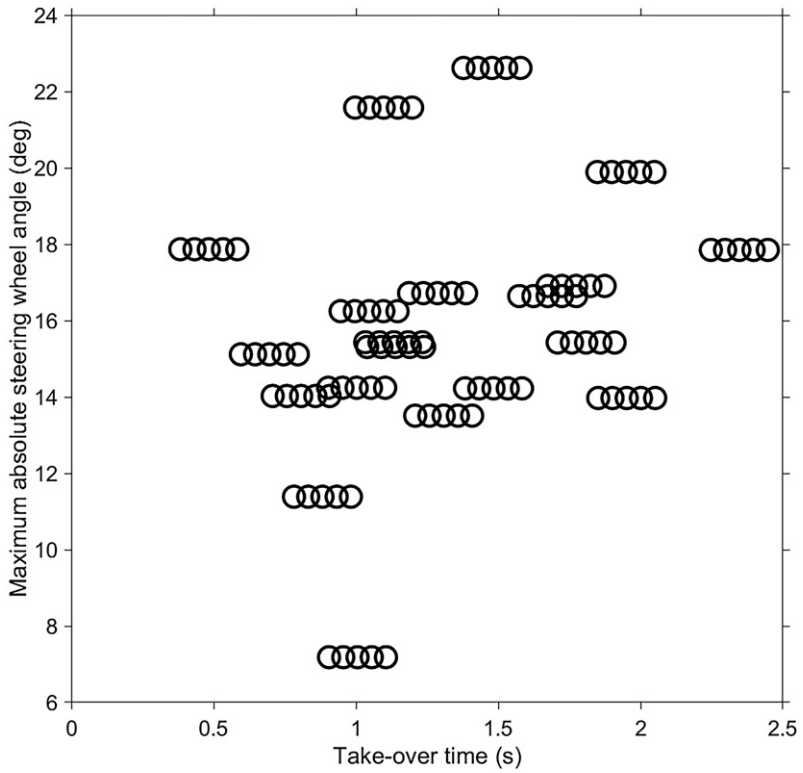

Figure 7 Example (simulated) data for an experiment on automation take-overs. The horizontal axis shows the take-over time, and the vertical axis shows the maximum absolute steering wheel angle. The data are the same as in **Fig. 6**, except that each value is now repeated five times. A small variation in take-over time was added (within the range of −0.1 s to 0.1 s) to assure that the markers are not overlapping.

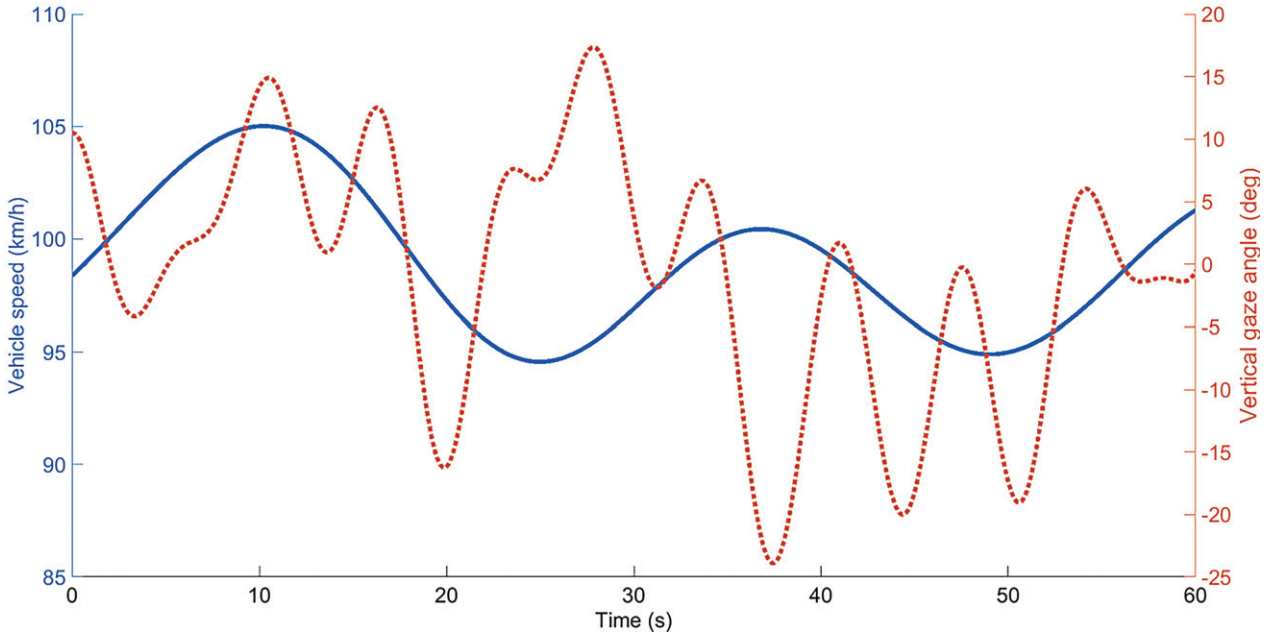

Figure 8 Example (simulated) data for vehicle speed and the driver's vertical eye movements. These data were generated using a multi-sine approach and a random number generator.

average of the five trials. The correlation coefficient, as depicted in **Fig. 7**, is the same as before ($r = 0.27$), but the *P*-value is clearly lower ($P = 0.007$). The reason for the reduced *P*-value is that repeated samples are incorrectly treated as independent. In other words, the researcher has artificially multiplied the sample size by a factor 5.

Violation of independence can come in different guises. It occurs not only in the computation of *P*-values for correlation coefficients but also in other types of statistical tests (e.g., ANOVAs) and other data types, such as time series. Assume that a researcher computes the relationship between two signals. For example, the traffic psychologist may be interested in whether drivers are more likely to look at the road instead of the dashboard when the vehicle speed is higher. Accordingly, the researcher computes the correlation coefficient between vehicle speed in km/h and the vertical gaze angle in degrees, as measured with an eye-tracker in the car. In the example shown in **Fig. 8**, which involves simulated data, a correlation coefficient of 0.22 is observed. It often happens that the researcher claims that such a correlation is statistically significant from 0, by using all sampling instances as inputs. For example, the data shown in **Fig. 8** have been recorded at 50 Hz, so 1 min of data involves 3000 samples, and the corresponding *P*-value is 8.37×10^{-34}. However, this *P*-value is clearly misleading and incorrect, as the adjacent sample points are correlated. In fact, the two signals shown in **Fig. 8** were generated using a random number generator and are independent.

In summary, researchers in traffic psychology should ensure that the participants are the unit of analysis, not the repeated measurements of the same participants. It is also possible to treat multiple independent trials as the unit of analysis, an approach which tends to be taken in psychophysics (Smith and Little, 2018).

Conclusion

This article highlighted a number of common pitfalls in the use of statistics in traffic psychology. The take-home message of this work is that a single statistically significant *P*-value cannot be used to disprove one's null hypothesis, especially when the *P*-value is barely significant (e.g., $P = 0.04$) or when it is a surprising discovery. Furthermore, the article pointed out that effect sizes should always be considered and that decision-making should not be based on *P*-values only: sometimes even extremely small and practically insignificant effects are statistically significantly different from zero, or strong effects may exist which are not particularly interesting because "everything is correlated". We also showed that two-stage statistical procedures are inefficient and that data peeking increases the prevalence of false positives. Fourth, and finally, we highlighted the issue of violation of independence, which we believe is a common problem in the field.

The current article is far from complete, and there are a variety of other pitfalls that can be acknowledged. Other common mistakes are (1) the usage of the Kaiser criterion (e.g., counting the number of eigenvalues greater than 1 for deciding on the number of factors to retain in factor analysis, (2) confusing standardized and nonstandardized coefficients in a regression analysis, and (3) multicollinearity in the regression model (e.g., inputting strongly correlated variables, such as driver age and driver experience in years).

We hope that the present article can be of use to the applied researcher. Recognition of the presented pitfalls may improve the robustness of published findings in the area of traffic psychology.

Supplementary Materials

A MATLAB script that reproduces the analyses, figures, and tables is available at https://doi.org/10.4121/13141289

References

Af Wåhlberg, A., 2017. Driver Behaviour and Accident Research Methodology: Unresolved Problems. CRC Press, Taylor and Francis Group, Boca Raton, FL.

Af Wåhlberg, A.E., 2010. Social desirability effects in driver behavior inventories. J. Safety Res. 41, 99–106, doi:10.1016/j.jsr. 2010.02.005.

Anscombe, F.J., 1960. Rejection of outliers. Technometrics 2, 123–146, doi:10.2307/1266540.

Antonakis, J., Dietz, J., 2011. Looking for validity or testing it? The perils of stepwise regression, extreme-scores analysis, heteroscedasticity, and measurement error. Pers. Individ. Differ. 50, 409–415, doi:10.1016/j.paid.2010.09.014.

Bakker, M., Wicherts, J.M., 2014. Outlier removal and the relation with reporting errors and quality of psychological research. PLOS ONE 9, e103360, doi:10.1371/journal.pone.0103360.

Bartram, D., 2007. Increasing validity with forced-choice criterion measurement formats. Int. J. Sel. Assess. 15, 263–272, doi:10.1111/j.1468-2389.2007.00386.x.

Bem, D.J., 2011. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. J. Pers. Soc. Psychol. 100, 407–425, doi:10.1037/a0021524.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J., Berk, R., Cesarini, D., 2018. Redefine statistical significance. Nat. Hum. Behav. 2, 6–10, doi:10.1038/s41562-017-0189-z.

Brown, A., Maydeu-Olivares, A., 2011. Item response modeling of forced-choice questionnaires. Educ. Psychol. Meas. 71, 460–502, doi:10.1177/0013164410375112.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–376, doi:10.1038/nrn3475.

Chavalarias, D., Wallach, J.D., Li, A.H.T., Ioannidis, J.P., 2016. Evolution of reporting P values in the biomedical literature 1990-2015. JAMA 315, 1141–1148, doi:10.1001/jama.2016.1952.

De Winter, J.C.F., Dodou, D., 2015. A surge of p-values between 0 041 and 0. 049 in recent decades (but negative results are increasing rapidly too). PeerJ 3, e733, 10.7717/peerj.733.

De Winter, J.C.F., Dodou, D., 2017.In: Human subject research for engineers. Cham: SpringerBriefs in Applied Sciences and Technology. Springer.

De Winter, J.C.F., Dodou, D., Stanton, N.A., 2015. A quarter of a century of the DBQ: some supplementary notes on its validity with regard to accidents. Ergonomics 58, 1745–1769, doi:10.1080/00140139.2015.1030460.

De Winter, J.C.F., Kovácsová, N., Hagenzieker, M., 2019. Cycling Skill Inventory: assessment of motor-tactical skills and safety motives. Traffic Inj. Prev. 20, 3–9, doi:10.1080/15389588.2019.1639158.

Evans, L., 2009. Reason, replication, and other scientific basics need increased emphasis to better advance safety knowledge. In: SWOV and Friends Symposium, The Hague, the Netherlands.

Factor, R., 2014. The effect of traffic tickets on road traffic crashes. Accid Anal. Prev. 64, 86–91, doi:10.1016/j.aap.2013.11.010.

Field, A., 2013. Discovering Statistics using IBM SPSS Statistics, 4th ed. Sage Publications.

Hartgerink, C.H., Van Aert, R.C., Nuijten, M.B., Wicherts, J.M., Van Assen, M.A., 2016. Distributions of *p*-values smaller than .05 in psychology: what is going on? PeerJ 4, e1935, doi:10.7717/peerj.1935.

Horn, J.L., McArdle, J.J., 2007. Understanding human intelligence since Spearman. In: Cudeck, R., MacCallum, R.C. (Eds.), Factor Analysis at 100. Historical Developments and Future Directions. Lawrence Erlbaum Associates, pp. 205–247.

Ioannidis, J.P., 2005. Why most published research findings are false. PLOS Med. 2, e124, doi:10.1371/journal.pmed.0020124.

Körber, M., Radlmayr, J., Bengler, K., 2016. Bayesian highest density intervals of take-over times for highly automated driving in different traffic densities. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 60, 2009–2013, doi:10.1177/1541931213601457.

Lakens, D., 2015a. Always use Welch's t-test instead of Student's t-test [blog]. Retrieved from http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html.

Lakens, D., 2015b. Comment: What p-hacking really looks like: A comment on Masicampo and Lalande (2012). Q. J. Exp. Psycho. 68, 829–832, 10.1080%2F17470218.2014.982664.

Lakens, D., Adolfi, F.G., Albers, C.J., Anvari, F., Apps, M.A., Argamon, S.E., Buchanan, E.M., 2018. Justify your alpha. Nat. Hum. Behav. 2, 168–171, doi:10.1038/s41562-018-0311-x.

Lord, F.M., 1953. On the statistical treatment of football numbers. Am. Psychol. 8, 750–751, doi:10.1037/h0063675.

Mattsson, M., 2012. Investigating the factorial invariance of the 28-item DBQ across genders and age groups: an Exploratory Structural Equation Modeling Study. Accid. Anal.Prev. 48, 379–396, doi:10.1016/j.aap.2012.02.009.

Meehl, P.E., 1990. Appraising and amending theories: the strategy of Lakatosian defense and two principles that warrant it. Psychol. Inq. 1, 108–141, doi:10.1207/s15327965pli0102_1.

Nederhof, A.J., 1985. Methods of coping with social desirability bias: a review. Eur. J. Soc. Psychol. 15, 263–280, doi:10.1002/ejsp.2420150303.

Nelson, L.D., Simmons, J., Simonsohn, U., 2018. Psychology's renaissance. Annu. Rev. Psychol. 69, 511–534, doi:10.1146/annurev-psych-122216-011836.

Nordhoff, S., De Winter, J., Madigan, R., Merat, N., Van Arem, B., Happee, R., 2018. User acceptance of automated shuttles in Berlin-Schöneberg: A questionnaire study. Transp. Res. Part F: Traffic Psychol. Behav. 58, 843–854, doi:10.1016/j.trf.2018.06.024.

Nosek, B.A., Ebersole, C.R., DeHaven, A.C., Mellor, D.T., 2018. The preregistration revolution. Proc. Natl. Acad. Sci. 115, 2600–2606, doi:10.1073/pnas.1708274114.

Nunnally, J., 1960. The place of statistics in psychology. Educ. Psychol. Meas. 20, 64–650, doi:10.1177/001316446002000401.

Rasch, D., Kubinger, K.D., Moder, K., 2011. The two-sample *t* test: pre-testing its assumptions does not pay off. Stat. Pap. 52, 219–231, doi:10.1007/s00362-009-0224-x.

Smith, P.L., Little, D.R., 2018. Small is beautiful: In defense of the small-*N* design. Psychon. Bull. Rev. 25, 2083–2101, doi:10.3758/s13423-018-1451-8.

Standing, L., Sproule, R., Khouzam, N., 1991. Empirical statistics: IV Illustrating Meehl's sixth law of soft psychology: everything correlates with everything. Psycho. Rep. 69, 123–126, doi:10.2466/pr0.1991.69.1.123.

Thompson, B., 1989. Why won't stepwise methods die? Meas. Eval. Couns. Dev. 21, 146–148, doi:10.1080/07481756.1989.12022899.

Wagenmakers, E.J., Wetzels, R., Borsboom, D., Van Der Maas, H.L., 2011. Why psychologists must change the way they analyse their data: the case of psi: comment on Bem (2011). J. Pers. Soc. Psychol. 100, 426–432, doi:10.1037/a0022790.

Ware, J.J., Munafò, M.R., 2015. Significance chasing in research practice: causes, consequences and possible solutions. Addiction 110, 4–8, doi:10.1111/add.12673.

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

Zhang, B., De Winter, J., Varotto, S., Happee, R., Martens, M., 2019. Determinants of take-over time from automated driving: A meta-analysis of 129 studies. Transp. Res. Part F: Traffic Psychol. Behav. 64, 285–307, doi:10.1016/j.trf.2019.04.020.

Zimmerman, D.W., 2011. A simple and effective decision rule for choosing a significance test to protect against non-normality. Br. J. Math. Stat. Psychol. 64, 388–409, doi:10.1348/000711010X524739.