Article:
Jones, A., Petrovskaya, E. and Stafford, T. orcid.org/0000-0002-8089-9479 (2024) Exploring the multiverse of analysis options for the alcohol Stroop. Behavior Research Methods, 56 (4). pp. 3578-3588. ISSN 1554-351X

https://doi.org/10.3758/s13428-024-02377-5

Behavior Research Methods (2024) 56:3578–3588
https://doi.org/10.3758/s13428-024-02377-5

ORIGINAL MANUSCRIPT

Exploring the multiverse of analysis options for the alcohol Stroop

Andrew Jones[1] · Elena Petrovskaya[2] · Tom Stafford[3]

Accepted: 21 February 2024 / Published online: 14 March 2024

Abstract

The alcohol Stroop is a widely used task in addiction science to measure the theoretical concept of attentional bias (a selective attention to alcohol-related cues in the environment), which is thought to be associated with clinical outcomes (craving and consumption). However, recent research suggests findings from this task can be equivocal. This may be because the task has many different potential analysis pipelines, which increase researcher degrees of freedom when analysing data and reporting results. These analysis pipelines largely come from how outlying reaction times on the task are identified and handled (e.g. individual reaction times > 3 standard deviations from the mean are removed from the distribution; removal of all participant data if > 25% errors are made). We used specification curve analysis across two alcohol Stroop datasets using alcohol-related stimuli (one published and one novel) to examine the robustness of the alcohol Stroop effect to different analytical decisions. We used a prior review of this research area to identify 27 unique analysis pipelines. Across both data sets, the pattern of results was similar. The alcohol Stroop effect was present and largely robust to different analysis pipelines. Increased variability in the Stroop effect was observed when implementing outlier cut-offs for individual reaction times, rather than the removal of participants. Stricter outlier thresholds tended to reduce the size of the Stroop interference effect. These specification curve analyses are the first to examine the robustness of the alcohol Stroop to different analysis strategies, and we encourage researchers to adopt such analytical methods to increase confidence in their inferences across cognitive and addiction science.

Keywords Alcohol Stroop · Alcohol · Craving · Multiverse · Specification curve analysis

Introduction

Over the past decade, there has been increased awareness and discussion of a '*reproducibility crisis* or *renaissance*' within science, and in particular psychology (Nosek et al., 2022). Many seemingly robust research findings have since failed to replicate or have come under increased scrutiny from the field. For example, in 2015 the Open Science Collaboration attempted to replicate 100 psychology studies published in high-impact journals (Open Science Collaboration, 2015), with increased statistical power. Only 36% of the studies were replicated and overall, the effect sizes from the replications were much smaller than those reported in the original studies. Subsequent large-scale replication attempts have also demonstrated a similar pattern (see Hagger et al., 2016; Klein et al., 2014). Whilst it is extremely difficult to identify the cause of unreliable findings, researchers have suggested it may be a combination of low-powered studies (Button et al., 2013), questionable research practices (Xie et al., 2021), poor transparency in reporting (Simmons et al., 2011), and opportunistic use of researcher degrees of freedom (Wicherts et al., 2016).

Researcher degrees of freedom are the numerous, often defensible, but arbitrary choices that are made when collecting, analysing, and reporting data for an experiment. This includes determining when to stop collecting data, how to randomise participants, how to identify and handle outlying data points, and which variable to select as a primary outcome (see Wicherts et al., 2016). In each of these cases,

Andrew Jones and Elena Petrovskaya contributed equally (joint lead authors).

✉ Andrew Jones
a.j.jones@ljmu.ac.uk

[1] School of Psychology, Tom Reilly Building, Liverpool John Moore's University, Byrom Street, L3 3AF Liverpool, UK

[2] Department of Computer Science, University of York, York, UK

[3] Department of Psychology, University of Sheffield, Sheffield, UK

a decision will need to be made from numerous possible options. Highlighting the potential problem with researcher degrees of freedom, Simmons et al. (2011) showed that by selectively choosing specific variables in their models and also failing to report certain analysis steps, listening to the Beatles song 'When I'm 64' literally made participants younger – which is of course physically impossible. This demonstrated the 'invisible multiplicity' researcher degrees of freedom can provide (Gelman & Loken, 2013), as only the final analytical decisions tend to be reported but may not be the only analyses that were attempted. Given that most data sets can be generated and analysed in a number of different ways, and under the assumption, a type 1 error will be made ~ 5% of the time (if there is no true effect), a determined enough researcher would be able to demonstrate a seemingly real effect by applying a relatively small number of design/analysis variations to any given data set. For example, Carp et al. demonstrated at least 6912 different analysis pipelines, from ten different pre-processing steps for neuro-imaging data (Carp, 2012). This could potentially lead to 345 analyses that could obtain a false-positive result ($p < .05$).

One mechanism of combatting unfettered researcher degrees of freedom is pre-registration (Yamada, 2018), in which analyses decisions are publicly stated ahead of time (usually before data has been collected). However, this generally limits researchers to one a-priori-defined analysis, which may or may not be viewed by others as the most appropriate method. It is reasonable to suggest that for most data sets, researchers will disagree on how best to analyse them. Indeed, several 'many analyst' projects have demonstrated this case. Silberzahn et al. (2018) provided 61 analysts across 29 teams with the same data to address the research question of whether dark-skin-toned soccer players are more likely to be punished than light-skin-toned soccer players. They demonstrated considerable variability in the analyses. Sixty-nine percent of analyses found a significant effect (31% did not), with effect sizes ranging from odds ratio = 0.89 (slightly negative) to odds ratio = 2.93 (moderately positive). Importantly, the analysts were not motivated to demonstrate a 'significant effect' and accounting for their statistical expertise was unable to account for this variability. Similarly, Botcinik-Nezer et al. (2020) demonstrated that across 70 teams who analysed the data, no two teams chose identical analysis pathways for fMRI data, and there was considerable disagreement across teams for the tested hypotheses.

To overcome these issues, there has been a shift to multiverse analyses/specification curve analysis (SCA) to examine whether findings are robust to different analysis strategies (see Simonsohn et al., 2020; Steegen et al., 2016). In this case, the raw data which is collected for an experiment is used to construct a multiverse of data sets by combining different data processing decisions. Rather than report one single statistical analysis, all reasonable analyses are reported, as long as they are consistent with the underlying theory, statistically valid, and are not redundant with other specifications. These techniques have been used across psychology. For example, examining birth order effects on personality, with thousands of separate analyses (Rohrer et al., 2017). Similarly, using 20,004 (out of a possible 2.5 trillion) specifications, Orben and Przybylski (2019) examined the link between well-being and technology usage in adolescents. As well as modelling variability based on analysis decisions, SCA help to identify trends (e.g. does the inclusion of specific covariates in models increase/decrease the effect) but also allow for an average effect size, based on all possible specifications (Flournoy et al., 2020).

Many published SCA analyses have examined the inclusion of different covariates into models, across multiple data sets. However, there has been little focus on whether pre-processing decisions, such as outlier removal taken can impact the magnitude of a given measure (outliers were addressed in the original paper describing specification curves: Simonsohn et al., 2020). These kinds of decisions are particularly important in cognitive/behavioural tasks, which measure reaction times and/or accuracy of responses. Outlier removal has been recognised as a potential data processing step which can impact findings (see Gress et al., 2018), and the reporting of outlier removal has increased during the replication crisis (Valentine et al., 2021). Often, there is no gold standard or widely accepted method of analysing data from these tasks. As such, there is considerable variability in their pre-processing across studies/lab groups which can impact their outcomes and reliability. For example, studies have shown that a priori decisions on the removal of outliers in reaction time distributions (e.g. using the mean vs. median, removing reaction times greater than 2 or 3 standard deviations around the individual mean) can impact the reliability of a widely used task to measure attentional bias (the Visual Probe task: Jones et al., 2018; Price et al., 2015), but also the Stroop and flanker tasks (Parsons, 2020), and contextual cueing tasks (Vadillo et al., 2023). However, other methods of removal also exist, such as transformation (e.g. recoding of extreme data points; Leys et al., 2019, provides a comprehensive overview of univariate outlier removal techniques).

The number of different decisions researchers might make when handling reaction time data is considerable and may have considerable impacts on theoretical and clinical findings. One research area where a large number of design and analysis decisions have been identified is the alcohol Stroop (see Jones et al., 2021). The alcohol Stroop is a widely used measure of 'attentional bias' in the addiction literature (Bollen et al., 2022). Attentional bias is the observation that individuals with an alcohol use history will show selective attention to substance-related cues in their environment, and

this attention is thought to be indicative of current motivation to drink alcohol (craving), but also predict consumption (Field et al., 2016; Bollen et al., 2022). Furthermore, a line of research has attempted to target attentional biases as a candidate for psychological treatment for alcohol (mis)use (Fadardi & Cox, 2009).

Despite the widespread use of the task, findings are equivocal (for an excellent review see Bollen et al., 2022). Wider observations of the literature have also suggested that poor methodological practices should reduce any enthusiasm for positive results (Christiansen et al., 2015). Novel techniques such as specification curve analyses may help to resolve debates on the veracity of alcohol Stroop findings, but also identify any specific analyses pipelines which might increase/decrease any observed effects. Given the similarities in data-processing across different cognitive tasks, a focus on a well-established task with clearly identified specifications (Jones et al., 2021) may provide a useful template for how to analyse and report data from these tasks moving forward.

Therefore, the aim of this project was to conduct specification curve analysis on (1) previously published alcohol Stroop data, and also (2) novel data collected for this purpose. We also collected data on craving and alcohol use to examine whether different analytic decisions impacted correlations between Stroop interference and these outcomes. We decided to focus on the alcohol Stroop, given the widespread use of the task and its importance in testing theoretical predictions of attentional bias (Cox et al., 2006). These findings will inform us whether researcher degrees of freedom can impact the Stroop effect (as a measure of attentional bias) which may, in turn, impact our confidence in the task as a robust measurement tool in the addiction field.

Methods

The (alcohol) Stroop task

The alcohol Stroop task is a variant of the Stroop task (Cox et al., 2006). The standard Stroop presents colour names in different colours (e.g. the word 'Red' is presented in the colour 'Blue'). Participants are asked to name the colour of the word whilst ignoring the semantic content. The inability to do this effectively is indicative of poor inhibitory control (Diamond, 2013).

The alcohol Stroop task uses two semantic-categories of words (alcohol-related and emotionally neutral) to generate a measure of attentional bias (the Stroop interference effect: reaction time to alcohol-related words – reaction time to neutral/comparison words). In line with the Stroop task, participants name the colour the words are presented in. Attentional bias is inferred through the difference in reaction times when alcohol-related words are presented compared to emotionally neutral words. For example, if it takes an individual longer to colour name alcohol-related words compared to the emotionally neutral words then they have an attentional bias to alcohol (i.e. their attention to the semantic content of the alcohol-related word means it takes them longer to identify the colour the word is presented in (Cox et al., 2006; Field & Cox, 2008). This paradigm has been used across various word categories to infer attentional bias towards or away from alcohol (see Fadardi & Cox, 2009).

Stroop task data sets

Previously published data

We re-analysed data from Spanakis et al. (2019). Specifically, we used data from an alcohol-related Stroop conducted via a mobile device using word stimuli. In this task, there were 66 critical trials, of which 33 were alcohol-related words and 33 were neutral/comparison words. There were 11 unique words in each category: alcohol (e.g. 'pub', 'beer', 'wine') and emotionally neutral (e.g. 'bog', 'ravine', 'valley'). Each word was presented in three possible colours (blue, green, red). Each trial began with a central fixation dot presented for 500 ms. Following this, the alcohol or neutral word was presented with three response boxes underneath containing the colour names. The order of these response boxes was randomised on each trial. Participants were required to make a touch screen response by pressing the box with the correct colour name. Participants had 3000 ms to respond before the trial timed out and it was coded as an incorrect response. The task used a blocked design – one block contained only the alcohol-related words and a second contained the neutral words. Block order was randomised.

Novel data set

We designed an alcohol Stroop to make it methodologically very different to the task utilized by Spanakis et al. (2019), to allow us to examine the robustness of analysis decisions across different task types. This is important, given that there are also considerable researcher degrees of freedom in how the alcohol Stroop is defined (Jones et al., 2021). The Stroop task was conducted online via a PC/laptop using a keyboard to record responses (rather than a touch screen).

The task was presented on a white background with a 300-ms fixation cross ('+') appearing in black before each word was presented. Words were presented in one of four possible colours; red, green, blue, or yellow, and participants were informed to ignore the content of the word and press a key indicating the colour. The key/colour information (D key for 'red', F key for 'green', J key for 'blue', K key for 'yellow') remained on screen at all times during the

task. Participants first completed 24 practice trials with the words ('one', 'two', 'three',… 'ten') in different colours. To progress, participants had to register at least 80% correct responses on the practice trials. If participants failed to do so, the experiment ended, and participants were unable to re-attempt their participation. If participants made an incorrect response a red 'x' appeared for 400 ms.

Following the practice trials, there were 168 critical trials; 84 alcohol trials and 84 neutral trials, presented in completely random order. There were 14 alcohol-related words (e.g. 'beer', 'alcopops', 'vodka') and 14 emotionally neutral words (e.g. 'box', 'queue', 'carpet'). If participants failed to respond after 3000 ms, a response was coded as incorrect. A completely randomised trial design was used (e.g. no blocks). In both tasks, the alcohol and emotionally neutral words were taken from previously established and widely implemented alcohol Stroop tasks (e.g. Fadardi & Cox, 2006), in which the emotionally neutral words have been matched for syllables, length and frequency within the English language, to control for any differences not related to the semantics of the word. This allows us to further generalise our findings to existing literature.

Procedure for novel data set

Participants were recruited via Prolific (Peer et al., 2017), to participate in a study titled '*Cognitive Processes and Alcohol consumption*' on December 10–11, 2020. The following Prolific screeners were used: aged 18+, resident in the UK, consumption of 14+ units of alcohol per week. Participants first read an information sheet and provided informed consent before providing demographic information (age in years and gender identity [male, female, non-binary, other]). Participants were then asked '*On a scale of 0 (no craving at all) to 100 (intense craving), how would you rate your craving for an alcoholic beverage at this moment in time*', followed by the Alcohol Use Disorders Identification Task-C (AUDIT-C: (Bush et al., 1998) data not reported here). Following this, participants completed the Stroop task. After completing the Stroop task, participants were asked if they had been distracted at all during the task (yes, no) and debriefed. The experiment lasted approximately 10 min total, and participants were reimbursed £1.10. The Stroop task was programmed and presented using Inquisit Web version 5 (Millisecond Software, Seattle, WA. USA). We aimed to recruit > 119 participants as a power calculation determined this would be enough to detect a small difference in RTs between alcohol and neutral words (d = .23, based on Jones et al. (2021), with 80% power and alpha = .05). Available funds allowed us to oversample, and 166 participants were recruited. Ten participants did not make it past the practice trials.

Specification curve and data analysis

For each specification curve, we begin with raw trial-level data from each participant on the alcohol Stroop task(s). First, all practice trials were removed, as is typical when calculating performance indices on these tasks. From the raw trial-level data, we recorded information on whether the trial was an alcohol-word trial or an emotionally neutral-word trial, and for each trial type, we recorded whether participants were correct (e.g. they identified the correct colour of the word) and the reaction time of their response. For each trial, reaction times on incorrect responses were not included when computing mean reaction times or standard deviations.

Following this initial raw data cleaning, we then passed this trial-level data through analytic code which determined each specification (e.g. removal of any reaction times > 4000 ms or removal of any reaction times > 3 standard deviations from the mean), before using the remaining data to compute a Stroop inference interference effect [RT to alcohol-related words – RT to neutral-related words]. This Stroop interference effect (in milliseconds) is our dependent variable for the main specification curves. Importantly, the specifications were not additive (e.g. the removal of any reaction times > 4000 ms, followed by the removal of reaction times > 3 SDs from the mean). Each specification was done in isolation on the individual trial data which had practice trials and reaction times from incorrect trials removed.

Our chosen specifications were taken from previous research which identified the different analytical approaches taken in the alcohol Stroop task (Jones et al., 2021). However, as these specifications themselves may only be present due to publication bias (for example, only specifications which led to a positive Stroop interference effect will be present in the published literature) we also identified two other measures of removing outliers discussed in Leys et al. (2019). These were the median absolute deviation, and Yuen's trimmed means approach. Median absolute deviation it relies on the median as the estimate of the centre of the distribution, and on the absolute difference (rather than standard deviations: see Leys et al., 2013). Yuen's trimmed means approach removes a percentage of extreme responses above and below the mean (here we used 10% and 20%) as cut-offs. The inclusion of these novel techniques would allow us to examine whether atypical techniques influence the overall Stroop effect.

Specification curve analyses are presented as a figure with an upper and lower panel. The upper panel presents the estimates and 95% confidence intervals of the Stroop interference effect for each specification ranked from

Table 1 Analysis decisions for reaction times/errors based on the alcohol-Stroop task used in the specification curve

No exclusion	Individual RT removed / replaced based on SDs	Individual RT removal based on raw RTs	Participant removal	Median absolute deviation	Trimmed mean
No exclusions (median RTs used to calculate interference)	Removal of RTs > 3 SDs or < 2 SDs from individual mean	Removal of all RTS > 2000 ms	Participant removed if their mean RT > 4 SDs from the sample mean	Removal of RTS > 2.5 MAD from the individual median	Removing 20% of individual participants based on extreme values (from the mean)
No exclusions (mean RTs used to calculate interference)	Removal of RTs > 3SDs or < 3 SDs from individual mean	Removal of all RTS > 1500 ms	Participant removed if RT > 2 SDs from the sample mean	Removal of RTS > 2.5 MAD or < 2.5 MAD from the individual median	Removing 10% of individual participants based on extreme values (from the mean)
	Removal of RTs > 2 SDs or < 2 SDs from individual mean	Removal of all RTS > 1000 ms	Participant removed if number of errors > 3 SDs from the sample mean		
	Removal of RTs > 2 SDs from individual mean	Removal of all RTS < 400 ms	Participant removed if > 33% of errors on the task		
	Removal of RTs > 3 SDs from individual mean	Removal of all RTS < 300 ms	Participant removed if > 25% of errors on the task		
	Replacing RTs > 3 SDs from the mean with the mean	Removal of all RTS < 200 ms	Participant removed if Interference score > or < 4 SDS from sample interference score		
	Replacing RTs > 3 SDs from the mean with the mean + 3 SDs	Removal of all RTS < 150 ms	Participant removed if > 25% of RTS < 200 ms		
		Removal of all RTS < 100 ms			

smallest to largest. The lower panel presents a description of each specification grouped by type (see Table 1) to aid interpretation. Individual ticks in the lower panel correspond to the relevant estimate in the upper panel. We also include the intraclass correlation coefficients for effects of the data removal category (see Table 1) on the variability in the outcome (see Scharkow, 2019). This informs us how much variance in the Stroop Interference effect is explained by these categorical decisions.

To supplement the specification curves, we also examined the distribution of correlation coefficients of the alcohol Stroop effect when no outlier removal technique was used (in this case simply using the median reaction times) vs all other techniques (see Hussey, 2023). Considerable variability in correlation coefficients would suggest that choosing a different analytic pipeline would lead to different outcomes (and greater flexibility for selective reporting), whereas a narrow distribution of strong positive correlation coefficients suggests consistency in outcome irrespective of the analytic pipeline.

Raw data and analysis scripts can be found on OSF [https://osf.io/utnx2/]. Data were analysed in R using the 'tidyverse' (Wickham et al., 2019), and 'specr' (Masur & Scharkow, 2020) packages. Note, we are unable to share raw data from Spanakis (2019) as it is not permitted under the ethical approval of the original project. However, we share a synthetic version of the data using the r package '*synthpop*' (Nowok et al., 2016) for any users who may be interested.

Results

Published data

Across the 27 specifications, the Stroop interference effect was robust, with a statistically significant positive score under all analysis decisions but one (20% trimmed mean). The median interference score was ~ 36.3 ms. The difference between the smallest and largest inference score was 25.4 ms (see Fig. 1 for the specification curve). The intraclass correlation coefficient of the data removal categories was 0.09, indicating about ~ 9% of variance in the Stroop interference effect estimates was explained by the categories.

Distributions of correlation coefficients between no exclusions (median) and all other exclusions

The distribution of correlation coefficients was narrow and clustered around strong positive correlations (mean $r = .76$, sd = .13; see Fig. 2). One correlation was clearly smaller (median ~ RTs > 1000) than the rest ($r = 0.17$). However,

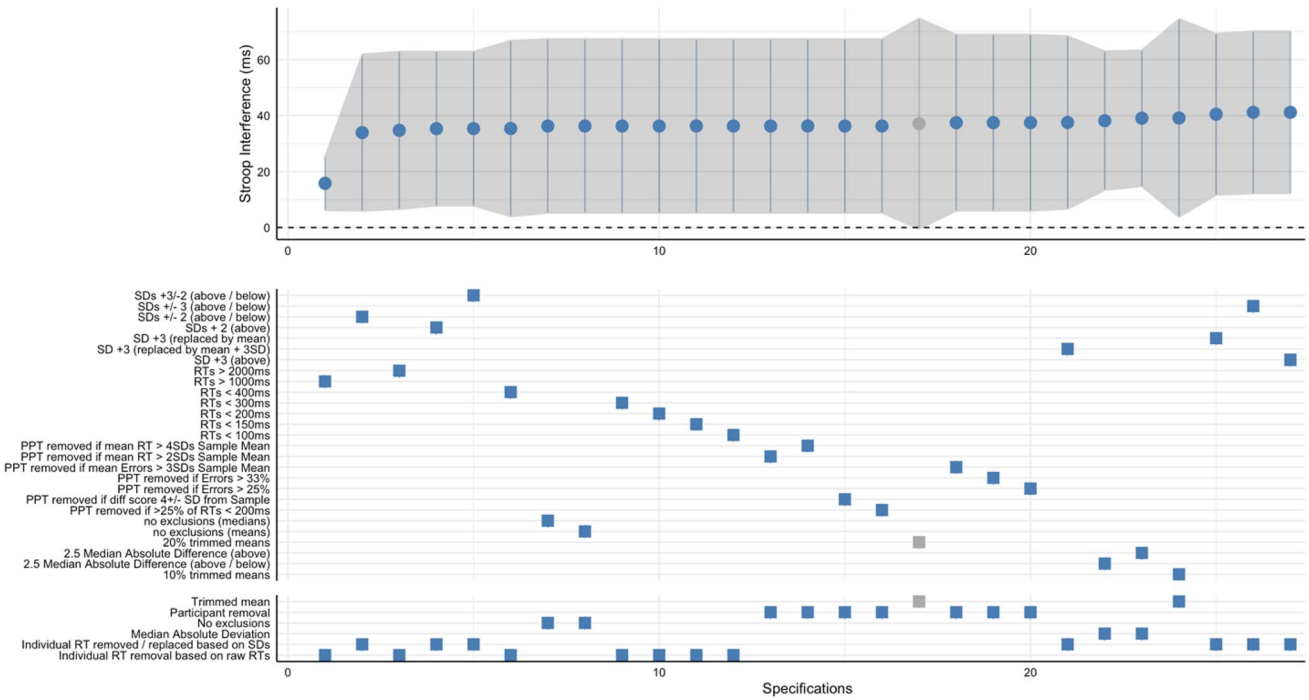

Fig. 1 Specification curve for the analysis decisions of the Stroop data published in Spanakis et al. (2019). The outcome is the Stroop interference effect (difference in RTs for alcohol – emotionally neutral words)

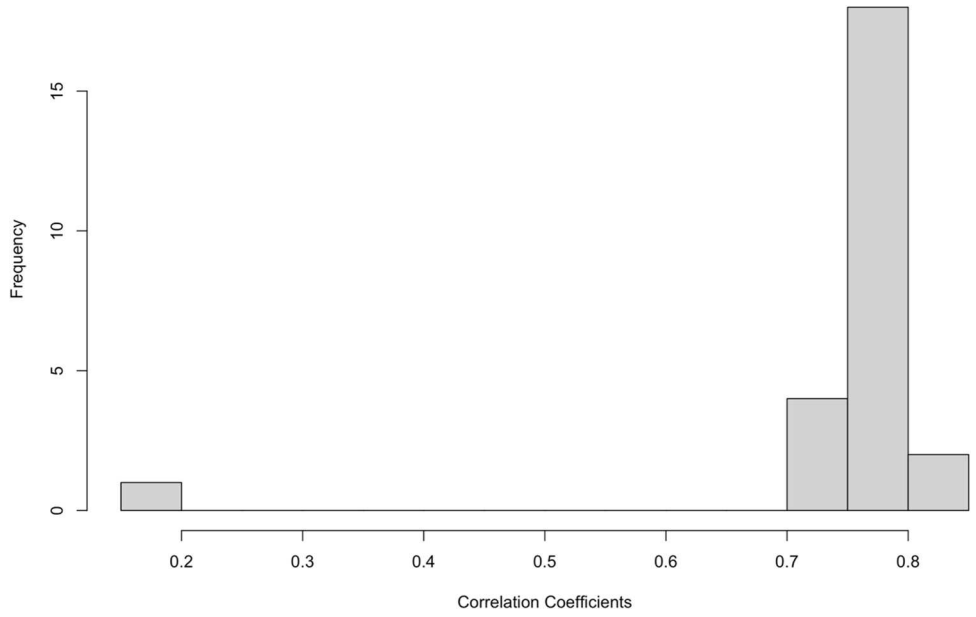

Fig. 2 Distributions of the correlations of the alcohol Stroop interference effect when comparing no exclusions (median) to all other possible strategies, in the published data set

this suggests that there is a consistency in the alcohol Stroop effect *within* individuals irrespective of most exclusions.

Novel data

Stroop interference

Across the 27 specifications, the Stroop interference effect was robust, with a statistically significant positive score under all analysis decisions ($ps < 0.001$). The median interference score was 25.2 ms. The difference between the largest and smallest Stroop interference score was 11.1 ms (see Fig. 3 for the specification curve). The intraclass correlation coefficient of the data removal categories was 0.74, indicating about ~ 74% of variance in the Stroop interference effect estimates was explained by the categories.

Distributions of correlation coefficients between no exclusions (median) and all other exclusions

The distribution of correlation coefficients was narrow and clustered around strong positive correlations (mean $r = .69$, sd = .04; see Fig. 4). In line with results from the published data, this suggests that there is a consistency in the alcohol Stroop effect *within* individuals irrespective of most exclusions.

Discussion

In this study, we used multiple SCA to examine the effects of analytical flexibility (researcher degrees of freedom; Wicherts et al., 2016) on the outcome of the alcohol Stroop task. Across two data sets (one published, one novel), we observed some variability in the size of the Stroop interference score, however, the overall Stroop interference effect was statistically significant across all specifications in both datasets, except one. We also demonstrated some consistency of the effect intra-individually, when comparing different analytic pipelines to no exclusions. These findings (and methodology) will be of importance to scientists who use the alcohol Stroop in their research, but also to any researchers who might use similar reaction-time-based tasks and be faced with a multitude of potential analysis pipelines. This study also adds to the increasing evidence base of the usefulness of SCA. These methods are another tool researchers have against the replication crisis and in improving the rigour of their findings.

This is the first study to our knowledge to directly examine whether the flexibility in analysis decisions can influence a commonly used metric of attentional bias in addiction psychology. Our specifications were not arbitrary and based on previously identified analyses from a systematic review, and thus could be justified by researchers when analysing their own data. Furthermore, we also included specifications from other sources (Leys et al., 2019) to reduce the likelihood of only 'successful' specifications being present in previously published research. We demonstrated the robustness of the

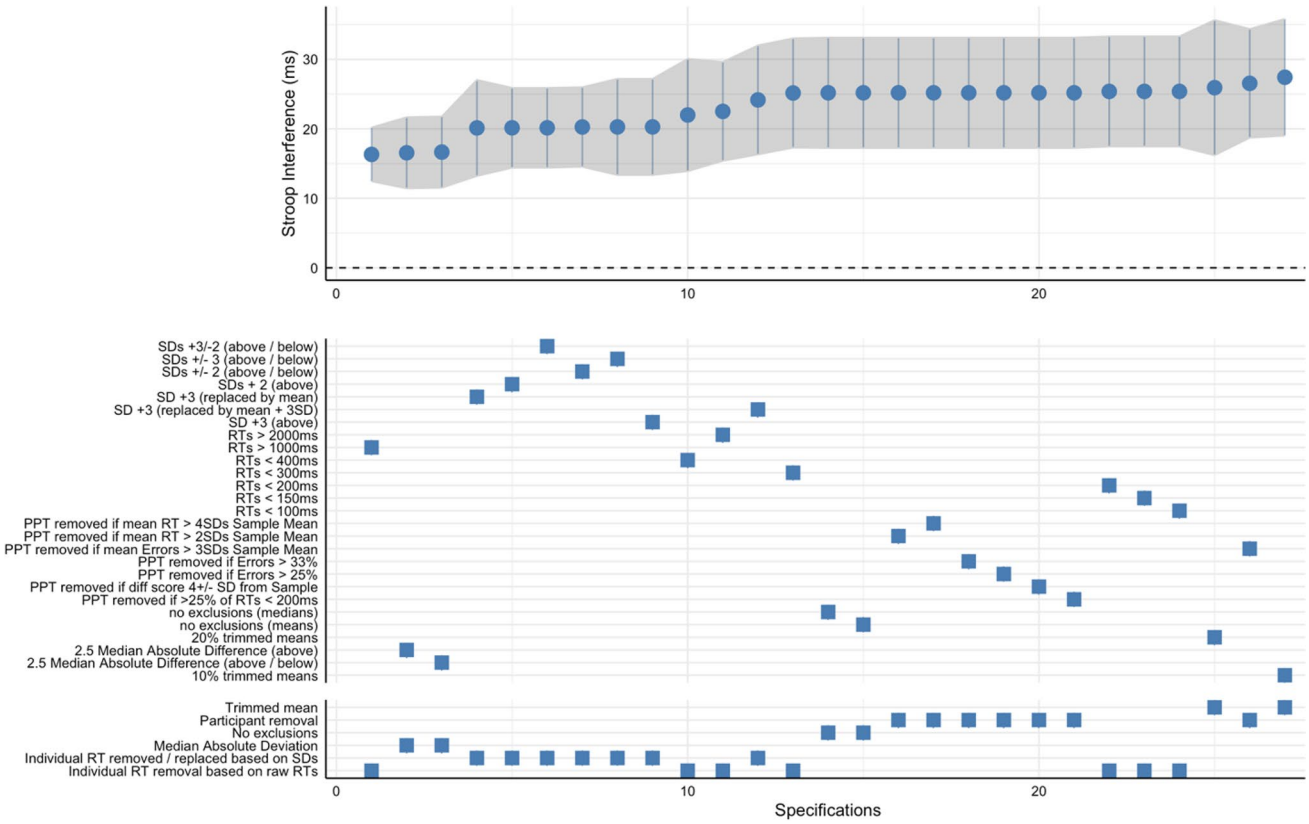

Fig. 3 Specification curve for different analysis decisions of the novel alcohol Stroop data. The outcome is the Stroop interference effect (difference in RTs for alcohol – emotionally neutral words)

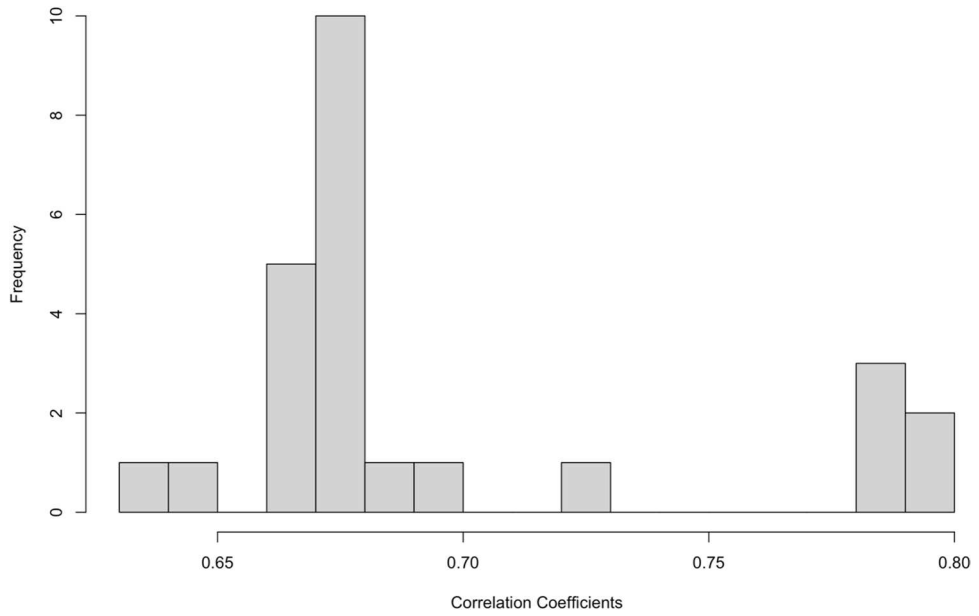

Fig. 4 Distributions of the correlations of the alcohol Stroop interference effect when comparing no exclusions (median) to all other possible strategies, in the novel data set

alcohol Stroop effect across a small but varied number of specifications. Our results also demonstrate that there was some variability in the effect sizes dependent upon analysis decisions, specifically when these decisions were based around outlier cut-offs using standard deviations from the individual mean and removal of upper bound reaction times. Removal of individual participants led to very little variability. This is likely because the criteria for removal (large number of errors) were strict. For example, Waters et al. (2009) demonstrated that < 1% of data were removed using these criteria. In our novel data, we likely constrained this further by removal of participants who made > 20% errors following the practice phase. Removal of individual RTs based on + 2/3SDs from the distribution generally led to smaller Stroop interference scores. It is likely that this is due to these cut-offs disproportionately remove longer reaction times to alcohol (rather than neutral) words. To our knowledge, when removing these 'outlying' reaction times, researchers tend to treat all reaction times as coming from the same distribution, rather than separate distributions for alcohol vs. neutral/control reaction times. However, if we assume a true Stroop effect, the overall distribution for alcohol reaction times should be different to that of neutral reaction times. The findings here support the alcohol Stroop interference effect as robust, and by extension theories that suggest drinkers have an attentional bias towards alcohol-related stimuli.

Only one specification led to a non-significant Stroop interference effect, in the published data. This was the 20% trimmed means approach. This specification requires the removal of the most data (from participants, rather than individual trial data). It is possible that this leads to a reduction of statistical power, as the overall estimate was similar to others; however the confidence intervals were considerably wider. Indeed, in the larger sample, the same pattern of results was not observed.

There are some limitations to our analyses. First, we focus on only the more prominent analysis decisions (identified in previous research; Jones et al., 2021) and as a result, our specification curves were relatively small. However, the overall pattern of results was consistent across all specifications for the Stroop effect. Furthermore, researchers have warned against overly inflating the analysis space with unnecessary specifications, which might serve to obscure reasonable effect estimates (Del Giudice & Gangestad, 2021). Secondly, our Stroop data were generated from atypical designs in which the Stroop was administered via a mobile device or online. Whilst psychology (and addiction science) transitions towards greater mobile and online testing (Gosling & Mason, 2015; Jones et al., 2022), these studies are still in the minority, making any generalisation to laboratory-based Stroop tasks more difficult (it is worth noting that our average Stroop interference scores were similar to interference scores in the lab: (Cox et al., 2006)). However, one might reasonably expect greater variability in reaction times when administered via mobile devices/online (Backx et al., 2020; Holden et al., 2019), making these data sets excellent candidates for our specification curves identifying more extreme responses. Nevertheless, future research should examine the full spectrum of possible specifications for the alcohol Stroop from data collected under laboratory-conditions to ensure the reliability of these effects.

In conclusion, we present the first SCA of the alcohol Stroop task. We chose reasonable specifications based on pre-existing analysis decisions as reported in a systematic review, as well as novel techniques not typically used in the field. The Stroop interference effect was robust but somewhat variable, supporting 'attentional bias' as a theoretical construct in alcohol use. We encourage researchers to consider the implementation of specification curve/multiverse analyses when analysing data from Stroop or similar cognitive tasks, to allow for the presentation of all possible analysis strategies. This will increase confidence in their findings, but also in the field moving forward.

Funding None.

Data Availability Raw data and analysis scripts can be found on OSF [https://osf.io/utnx2/].

Declarations

Ethics approval The research was approved by the Local Research Ethics Committee ('ID:5743, Psychological factors associated with substance use').

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication Not applicable.

Conflict of interest All authors report no conflicts of interest.

References

Backx, R., Skirrow, C., Dente, P., Barnett, J. H., & Cormack, F. K. (2020). Comparing web-based and lab-based cognitive assessment using the Cambridge Neuropsychological Test Automated

Battery: A within-subjects counterbalanced study. *J Med Internet Res, 22*(8), e16792. https://doi.org/10.2196/16792

Bollen, Z., Field, M., Billaux, P., & Maurange, P. (2022). Attentional bias in alcohol drinkers: A systematic review of its link with consumption variables. *Neuroscience and Biobehavioural Reviews, 139*, 104703.

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature, 582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Bush, K., Kivlahan, D. R., McDonell, M. B., Fihn, S. D., & Bradley, K. A. (1998). The AUDIT alcohol consumption questions (AUDIT-C): An effective brief screening test for problem drinking. *Archives of Internal Medicine, 158*(16), 1789–1795. https://doi.org/10.1001/archinte.158.16.1789

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365–376. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-84876665206&partnerID=40&md5=1ad34d5f809fb3bc56e78be53e40b0f0

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience, 6*, 149. https://doi.org/10.3389/fnins.2012.00149

Christiansen, P., Schoenmakers, T., & Field, M. (2015). Less than meets the eye: Reappraising the clinical relevance of attentional bias in addiction. *Addictive Behaviors, 33*, 43–50.

Collaboration, O.S. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Cox, M., Fadardi, J. S., & Pothos, E. (2006). The addiction-stroop test: Theoretical considerations and procedural recommendations. *Psychological Bulletin, 132*(3), 443–476. https://doi.org/10.1037/0033-2909.132.3.443

Del Giudice, M., & Gangestad, S. W. (2021). A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science, 4*(1), 2515245920954925. https://doi.org/10.1177/2515245920954925

Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168.

Fadardi, J., & Cox, M. (2006). Alcohol attentional bias: Drinking salience or cognitive impairment? *Psychopharmacology, 185*, 169–178.

Fadardi, J., & Cox, M. (2009). Reversing the sequence: Reducing alcohol consumption by overcoming alcohol attentional bias. *Drug, Alcohol Depend, 101*, 137–45.

Field, M., & Cox, W. M. (2008). Attentional bias in addictive behaviors: A review of its development, causes, and consequences. *Drug and Alcohol Dependence, 97*(1–2), 1–20. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-46649106485&partnerID=40&md5=183ab5bc55b9c7cddd20f7271fb380df

Field, M., Werthmann, J., Franken, I., Hofmann, W., Hogarth, L., & Roefs, A. (2016). The role of attentional bias in obesity and addiction. *Health Psychology, 35*(8), 767–780. https://doi.org/10.1037/hea0000405

Flournoy, J. C., Vijayakumar, N., Cheng, T. W., Cosme, D., Flannery, J. E., & Pfeifer, J. H. (2020). Improving practices and inferences in developmental cognitive neuroscience. *Development Cognitive Neuroscience, 45*, 100807. https://doi.org/10.1016/j.dcn.2020.100807

Gelman, A. & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf. Accessed 8-10-2023.

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66*, 877–902. https://doi.org/10.1146/annurev-psych-010814-015321

Gress, T., Denvir, J., & Shapiro, J. (2018). Effect of removing outliers on statistical inference: Implications to interpretation of experimental data in medical research. *Marshall Journal of Medicine, 4*, 9.

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Holden, J., Francisco, E., Lensch, R., Tommerdahl, A., Kirsch, B., Zai, L., … Tommerdahl, M. (2019). Accuracy of different modalities of reaction time testing: Implications for online cognitive assessment tools. *bioRxiv*, 726364. https://doi.org/10.1101/726364

Hussey, I. (2023). *Meta-methods analysis*. YouTube: https://www.youtube.com/watch?v=H66HegclUTQ&t=631s. Accessed 05-10-2023.

Jones, A., Christiansen, P., & Field, M. (2018). Failed attempts to improve the reliability of the alcohol visual probe task following empirical recommendations. *Psychology of Addictive Behaviors, 32*(8), 922–932. https://doi.org/10.1037/adb0000414

Jones, A., Worrall, S., Rudin, L., Duckworth, J. J., & Christiansen, P. (2021). May I have your attention, please? Methodological and analytical flexibility in the addiction Stroop. *Addiction Research & Theory, 29*(5), 413–426. https://doi.org/10.1080/16066359.2021.1876847

Jones, A., Earnest, J., Adam, M., Clarke, R., Yates, J., & Pennington, C. R. (2022). Careless responding in crowdsourced alcohol research: A systematic review and meta-analysis of practices and prevalence. *Experimental and Clinical Psychopharmacology, 30*(4), 381–399. https://doi.org/10.1037/pha0000546

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766.

Leys, C., Delacre, M., Mora, Y., Lakens, D., & Ley, C. (2019). How to classify, detect and manage univariate and multivariate outliers, with emphasis on pre-registration. *International Review of Social Psychology, 32*, 5.

Masur, P. K., & Scharkow, M. (2020). *specr: Conducting and visualizing specification curve analyses*. Available from https://CRAN.R-project.org/package=specr

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., … Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Nowok, B., Raab, G., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software, 74*(11), 1–26. https://doi.org/10.18637/jss.v074.i11

Orben, A., & Przybylski, A. K. (2019). The association between adolescent well-being and digital technology use. *Nature Human Behaviour, 3*(2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

Parsons, S. (2020). Exploring reliability heterogeneity with multiverse analyses: Data processing decisions unpredictably influence measurement reliability. https://doi.org/10.31234/osf.io/y6tcz

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Price, R. B., Kuckertz, J. M., Siegle, G. J., Ladouceur, C. D., Silk, J. S., Ryan, N. D., … Amir, N. (2015). Empirical recommendations for improving the stability of the dot-probe task in clinical research. *Psychological Assessment, 27*(2), 365–376. https://doi.org/10.1037/pas0000036

Rohrer, J. M., Egloff, B., & Schmukle, S. C. (2017). Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science, 28*(12), 1821–1832. https://doi.org/10.1177/0956797617723726

Scharkow, M. (2019). *Getting more information out of the specification curve*. https://underused.org/2019-01-spec-curve. Accessed 05-10-2023.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour, 4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

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. Retrieved from http://www.scopus.com/inward/record.url?eid=2-s2.0-80555145867&partnerID=40&md5=aed01f73904f880f8f18617f685cfd0a

Spanakis, P., Jones, A., Field, M., & Christiansen, P. (2019). A Stroop in the hand is worth two on the laptop: Superior reliability of a smartphone-based alcohol Stroop in the real world. *Substance Use & Misuse, 54*(4), 692–698. https://doi.org/10.1080/10826084.2018.1536716

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Vadillo, A. , Malejka, S. & Shanks, D. R. (2023). *Mapping the reliability multiverse of contextual cuing*. Preprint. https://osf.io/vncfx

Valentine, K., et al. (2021). Have psychologists increased reporting of outliers in response to the reproducibility crisis. *Social and Personality Psychology Compass, 15*, e12591.

Waters, A. J., Carter, B. L., Robinson, J. D., Wetter, D. W., Lam, C. Y., Kerst, W., & Cinciripini, P. M. (2009). Attentional bias is associated with incentive-related physiological and subjective measures. *Experimental and clinical psychopharmacology, 17*(4), 247–257. https://doi.org/10.1037/a0016658

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics, 27*(4), 41. https://doi.org/10.1007/s11948-021-00314-9

Yamada, Y. (2018). How to crack pre-registration: Toward transparent and open science. *Frontiers in Psychology, 9*, 1831. https://doi.org/10.3389/fpsyg.2018.01831