**May I have your attention, please? Methodological and Analytical Flexibility in the Addiction Stroop**

Andrew Jones1,2

Semra Worrall1

Lara Rudin1

Jay J. Duckworth1

Paul Christiansen1

1University of Liverpool, UK.

2Liverpool Centre for Alcohol Research

Author for correspondence: Department of Psychology, Eleanor Rathbone Building, University of Liverpool, L697ZA

Email: [ajj@liv.ac.uk](mailto:ajj@liv.ac.uk)

Twitter: @ajj\_1988

**Abstract**

Theoretical models of addiction predict that an attentional bias towards substance-related cues plays a role in development and maintenance of addictive behaviours although empirical data testing these predictions are somewhat equivocal. This may in part be a consequence of substantial variability in methods used to operationalise attentional bias. Our aim was to examine the variability in key design and analysis decisions of the addiction Stroop. Using a pre-registered design, we identified 95 studies utilising an addiction Stroop (46 alcohol, 25 smoking, 24 drug-related). We extracted key information about the design of the Stroop tasks, including; administration (paper-and-pencil vs. computerised), response (key-press vs. voice), design (block vs. mixed). For analysis decisions we extracted information on upper- and lower-bound reaction time cut-offs, removal of data based on standard error cut-offs, removal of participants based on overall performance, type of outcome used, and removal of errors. Based on variability from previous research there are at least 1,451,520 different possible designs of the computerised Alcohol Stroop, 77,760 designs of the computerised Smoking Stroop and 112,640 for the Drug Stroop. Many key design decisions were unreported. Similarly, variability in analyses decisions would allow for 9,000 different methods for analysing the Alcohol Stroop, 5,376 for the Smoking Stroop and 768 for the Drug Stroop. P-curves suggest data provided evidential value and exploratory meta-regressions suggest that the addiction Stroop effect was not associated with design and analysis decisions. Nevertheless, adoption of consistent reporting guidelines of the addiction Stroop is necessary to aid reliability and reproducibility.

**Key words: addiction; attentional bias; alcohol; smoking; Stroop; methods; degrees of freedom**

**Introduction**

Attentional bias refers to the tendency of specific types of stimuli, such as alcohol- and smoking-related cues, to capture, hold and make it difficult to disengage attention from (Field et al., 2016). Theoretical models of addiction predict that this selective and sustained attention towards drug-related cues is associated with craving and can play a potentially causal role in the development and maintenance of substance (mis)use (Field et al., 2014; Robbins & Ehrman, 2016). Empirical evidence has demonstrated some support for these claims. In a seminal meta-analytic investigation Field et al (2009) demonstrated a robust, albeit small (r = .19 [95% CI .15 - .23]), correlation between various indices of attentional bias and craving. Experimental manipulations of attentional bias have had a causal influence on alcohol consumed in the lab (Fadardi & Cox, 2009; Field & Eastwood, 2005), although see (Field, Duka, et al., 2009)), which has led to the development of psychological interventions to reduce alcohol- and tobacco-seeking in the ‘real-world’(Kerst & Waters, 2014). Whilst many researchers see promise in measuring and targeting these biases (Wiers et al., 2013), others have suggest poor methodology and lack of convincing evidence should reduce enthusiasm for most published findings (Christiansen et al., 2015; Cristea et al., 2016).

In order to accurately test the hypotheses made by incentive-motivational models of addiction, specifically that attentional bias to substance-related cues is present in substance users and associated with substance-related outcomes, attentional biases need to be operationalised. Various computerised tasks have been designed to isolate and measure attention (both indirectly, through reaction times, and directly using eye-tracking), such as the Visual Probe (MacLeod et al., 1986) and attentional blink paradigms (Shapiro et al., 1997) - interested readers can read a comprehensive review on measurement of attention biases here (Field & Cox, 2008). However, one of the most widely used measures is the ‘addiction’ Stroop, which is a variation of the original Stroop task. The original Stroop task was designed to measure the inhibition of cognitive interference (Stroop, 1935). Participants are shown a list of colour-words (e.g. ‘RED’, ‘BLUE’, ‘GREEN’) printed in different colours, and participants are asked to name the colour in which the word is printed in, whilst ignoring the content of the word. There are two types of trials; congruent trials in which the word is presented in its corresponding colour (e.g. the word ‘RED’ printed in red ink), and incongruent trials in which the word is presented in a different colour to its content (e.g. the word ‘RED’ printed in blue ink). Participants should be faster and make fewer errors on congruent trials compared to incongruent trials, and numerous studies have demonstrated the robustness of this effect across a variety of populations including; healthy controls, children, Attention Deficit Hyperactivity Disorder, depression and so on (Mourik et al., 2005; Verhaeghen & De Meersman, 1998).

The ‘addiction’ Stroop is an emotional Stroop variation, in which the emotional content of the words causes interference (through grabbing attention). Similar designs using relevant words have been used to a examine a range of attentional biases, such as safety-related words in the work place (Xu et al., 2014), and trauma-related words in war veterans with Post Traumatic Stress Disorder (Ashley et al., 2013). The ‘addiction’ Stroop replaces the standard colour words with alcohol/drug-related words and a comparison category (e.g. neutral words within a semantic category such as household or musical). Similarly, images rather than words can be used, where an image will have a coloured hue or background (pictorial Stroop). In these versions if participants are slower to name the colour in which alcohol/drug-related words/images are printed compared to neutral/comparison-words, this is interpreted as the content or meaning of the alcohol/drug-related word/image ‘grabbing’ the attention making it more difficult to identify the colour, than the control/comparison words. Using this task (or a variation), various studies have demonstrated an ‘attentional bias’ in heavy drinkers (Field et al., 2007), alcohol dependent patients (Lusher et al., 2004), as well as smokers (Cane et al., 2009) and illicit drug users (Marissen et al., 2006). However, there are ‘null’ results for the Stroop effect present throughout the literature (Jones et al., 2018; Asmaro, 2014 Franken et al, 2004), which may suggest any effect is not robust.

It is likely that the perceived simplicity of the design has led to widespread adoption of the ‘addiction’ Stroop as a measure of attentional bias, as it can be delivered in various settings with minimal/no programming experience (using only some paper and a stopwatch). However, subtle methodological differences in the design of the task may lead to widely different performance outcomes. A meta-analysis demonstrated the effect-sizes in alcohol-related Stroop tasks were moderated by a number of methodological variables (Cox et al., 2006). The number of word controls (i.e., semantic relatedness, word frequency, and word length), the use of paper-and-pencil (vs. computerised), vocal responses vs. button presses, and combined responses, were significantly, positively associated with increased effect sizes. Whereas the number of trials was significantly negatively associated with the effect size. There was no difference between blocked (substance-related cues and neutral cues presented in separate blocks) vs. non-blocked designs (substance-related and neutral cues in mixed blocks). Within the smoking Stroop the use of paper-and-pencil (vs. computerised) and unblocked (vs. blocked) designs were associated with significantly larger effect sizes. At first glance, these methodological decisions may seem trivial. However, increasing variation in tasks might impede our ability to generalise findings across studies. Furthermore, each modification to the design may also impact the psychometric properties of the task (Cooper et al., 2017). In support of this, Ataya et al (2012) re-analysed data collected from 7 addiction Stroop tasks conducted in their laboratory and demonstrated ‘acceptable’ reliability overall, but with considerable variability across task versions (ɑ = .74; 95% CI . 46 - .89). Other studies have demonstrated different estimates of internal reliability (Robinson et al., 2015; Spanakis et al., 2019), however most do not report reliability despite its integral role in effect estimates and reproducibility (Baugh, 2002).

The estimates of attentional bias may also be moderated by analysis decisions made by the researcher once data collection is completed. Various studies have demonstrated that analysing the same raw data in different ways can lead to dramatically different outcomes. For example, Carp et al (Carp, 2012a) demonstrated that across 10 pre-processing steps, each with between 2-4 different options led to 6,912 unique analysis pipelines in fMRI research, which when corrected for multiple comparisons using different methods increased to 34,560 different outcomes. Similarly, Silberzahn (2018) crowdsourced data analysis by giving 29 research teams (involving 61 analysts) the same data set and asked them to address the research question as to whether soccer referees are more likely to punish (give a red-card) dark-skin-toned players, compared to light skin-toned players. The analytic approaches, statistical significance and resulting effect sizes varied considerably.

This analytical flexibility has been termed ‘the garden of forking paths’ or ‘researcher degrees of freedom’. In support of this, Wicherts et al (2016) demonstrated at least 15 broad analysis decisions which might be taken, including; specifying pre-processing methods, deciding on how to deal with outliers and selecting the primary outcome. Of further concern are analysis decisions that might go unreported or be selectively reported. In a tongue-in-cheek example Simmons and colleagues (2011) demonstrated that unreported analysis decisions may lead to participants getting younger after listening to The Beatles.

The flexibility in designing and analysing research has been linked to the reproducibility ‘crisis’ in psychology and science as it may allow for questionable research practices such as p-hacking and HARKing (Munafò et al., 2017). P-hacking (or selective reporting) is using the flexibility in analytical pathways to calculate a p-value which might lead to the rejection of a null hypothesis (and as such supporting an alternative hypothesis of interest). HARKing (or Hypothesising After Results are Known) is generating a hypothesis based on the data, rather than in advance (a-priori). Selective reporting has been linked to increases in false-positive findings (Ioannidis, 2005). Tasks used to measure attentional bias and analyse the subsequent data allow extensive decision flexibility. Zhang et al (2019) demonstrated considerable variability and (lack of) reporting in the design of Visual Probe tasks for attentional bias modification. Similarly, we identified similar variability in analysis options when dealing with outliers on individual trials within the Visual Probe task, and how these decisions can influence psychometric properties (Jones et al., 2018). However, to our knowledge there has been limited investigation into the addiction Stroop.

Therefore, the aim of this study was to review the variability of potential design and analysis decisions a researcher might undertake for the addiction Stroop based on previous literature. Specifically, we wanted to provide an estimate of the number of i) design and ii) analysis decisions which could be made for the alcohol, smoking and drug-related Stroops respectively, based on what was reported in the literature. We also aimed to quantify the level of unreported methodological and analysis decisions. Finally, to examine potential p-hacking/selective reporting we conducted p-curve analyses on the overall Stroop effect (a difference in responding to one type of word, compared to another). It is possible that p-hacking in the Stroop literature may be present if there are a large number of analytical options, combined with null results. Similarly, there may be implausible effects reported in the literature, given poor reliability of measures constrains the maximum observable correlation between measures but also obscures effects through increased error variance (Parsons et al, 2019).

We pre-registered our study design and analyses strategy on Open Science Framework [https://osf.io/yp2wc]. We conducted formal searches similar to that of a systematic review, including search strategies, eligibility criteria and methodological data extraction. There are some deviations from our pre-registered protocol. Rather than include total number of trials used in the Stroop task we extracted information on the number of alcohol-, smoking-, drug-related cues used and the number of times each cue was presented. This provides greater information about the design of the Stroop task and taken together should contribute to the total number of trials, regardless.

**Methods**

*Search Strategy*

We searched PsychInfo (N = 613), Pubmed (N = 529) and Scopus (N = 228) in October 2019, using the following search terms: addiction OR substance AND use OR substance AND misuse OR drugs OR alcohol OR smoking OR cigarette OR nicotine OR amphetamine OR cocaine OR cannabis OR marijuana OR heroin OR polydrug OR opiate. As discussed, the majority of studies used either an alcohol or smoking Stroop, therefore our analyses focused on these.

*Inclusion and Exclusion Criteria*

In order to be eligible for inclusion studies had to include an addiction Stroop e.g. any study which included alcohol or smoking words or images within a traditional Stroop task, designed to measure attentional bias.

*Data extraction and analyses*

We extracted data in relation to the design / methodological aspects of the Stroop, as well as data related to analytic decisions. For design / methodological aspects we extracted information on administration (computerised vs paper and pencil), participant response (key press vs. voice), inclusion of practice trials (yes vs. not reported), number of experimental trials (not including practice trials), number of drug-related stimuli, type of stimuli (words vs. pictures), number of times each stimuli was presented, design (blocked vs. mixed), control stimuli (category), number of comparison categories, number of colours used; response timeout. Our extraction code book can be found on Open Science Framework. Extraction was shared across four authors (AJ, PC, LR, SW), with cross-checking. A fifth author (JD) independently extracted data from 21 of the articles with 85.7% agreement. All disagreements were resolved within the team.

For analysis decisions we extracted information on the lower-bound *reaction time (RT)* cut-off (in ms), upper-bound RT cut-off (in ms), removal of individual RTs based on standard deviation cut offs, exclusion of trials in which errors were made, the primary outcome of the task (e.g. Raw RTs, or ‘difference’ scores), and whether participants were removed from analyses based on Stroop performance.

In order to conservatively estimate the potential number of different *(hypothetical)* unique Stroop tasks an investigator could design based on previous research we multiplied the number of different variations for key design decisions,

*Number of variations = participant response type x inclusion of practice trials x number of drug-related stimuli x type of stimuli x number of times stimuli presented x design x control category x number of colours x timeout*.

We did not include administration type, as this generally does not include variation in response type (only voice), timeouts, or design (block designs only). We also did not include the number of critical trials, as this is often determined by other decisions we did include (e.g. number of stimuli, number of colours used) and this would considerably inflate our estimates. Similarly, we did not include the number of comparison conditions as this is largely related to specific hypotheses within papers.

In order to estimate the potential number of different analytical pipelines an investigator could attempt based on previous research we multiplied the number of different variations in key analysis variables,

*Analysis pipelines = lower-bound RT cut-off x upper-bound RT cut-off x SD removal x removal of errors x removal based on performance x primary outcome used.*

If a design decision was not reported, we took that to assume it was not carried out (e.g. no lower-bound RT cut-off was implemented), as it is impossible to infer selective reporting.

The conducted P-Curve tests for evidential value of the reported p-values using the P-Curve app (http://p-curve.com). The rationale for p-curve is as follows. First, unlike previous techniques for combining p-values (e.g. Fisher’s combined test) the p-curve only uses statistically significant p-values (p < .05). In the absence of a ‘real’ Stroop effect (i.e. no attentional bias to alcohol / smoking / drug related words) the distribution of p-values from p < .0001 to p ~ 1.0 should be uniform (all outcomes equally as likely: a flat p-curve). If there is a true ‘Stroop’ effect the distribution of p-values should be right skewed (more p values closer to p < .01 than p = .05) and the right skew should be larger with greater power of the studies. However, according to (Simonsohn, Nelson & Simmons, 2014) if selective reporting or p-hacking exists in the literature (in this case due to the large number of analysis decisions) then the p-curve should be left skewed (a greater number of p-values ~.05 than p < .01), due to experimenters’ limited ambition to achieve a small p-value rather than one that simply crosses the threshold of conventional statistical significance.

We report the continuous test which computes pp-values for each test, before standardizing the values (Z-scored). The sum of the Z scores is divided by the number of tests and the resulting Z score and corresponding p-value is the test for evidential value (Simonsohn et al., 2015). We also report the half-curve. The half-curve focuses on the distribution of p-values < .025, which may be a solution to deal with more ambitious p-hacking (Simonsohn, Simmons & Nelson, 2015). In order to conduct supplementary P-Curve analyses, we also extracted information on hypotheses for the Stroop effect (e.g. a test of reaction time / error differences to different stimulus types) and the statistical test for this hypothesis. Type of tests identified included paired samples t-tests (e.g. comparing mean reaction times to alcohol vs. control words), one sampled t-tests comparing a Stroop difference score (mean reaction times to control words – mean reaction times to neutral words) to 0, or omnibus tests (e.g. ANOVAS) if multiple images categories (e.g. alcohol vs. neutral vs. positive vs. negative) were used or the analyses included group comparisons (e.g. smokers vs. non-smokers). If this was the case, we extracted the F-value and degrees of freedom for the main effect of cue. Similarly, we also extracted statistical tests for hypotheses which examined that the Stroop effect was moderated by substance use status, (e.g. ‘compared to controls, abstinent alcoholics would show delayed colour-naming of alcohol-related words’ (Field et al., 2013)) or was associated with consumption/craving (e.g. ‘smoking Stroop effect should be associated with craving for the positive effects of smoking’ (Waters et al., 2009)). We also conducted robustness checks (e.g. if studies reported both a group difference between heavy and light drinkers in the Stroop, but also a correlation with alcohol consumption). These robustness checks did not substantially influence our findings. P-curve reporting table is available on Open Science Framework.

*Exploratory Meta-analysis*

In an attempt to explore whether the effect sizes from the Stroop task were associated with analysis decisions we conducted exploratory meta-analyses on the effect sizes for the Stroop effect across alcohol, smoking and drug Stroops. This analysis was not pre-registered, but was the consequence of a helpful peer-review process.

We extracted information from studies which allowed us to compute an effect size for the Stroop effect. This was either the mean or median reaction times / number of errors to substance alcohol words and the standard deviation, and the mean or median reaction times / number of errors to neutral comparison words and the standard deviation. If standard deviations were not available, but means were, we imputed these by estimating the average standard deviation in relation to the mean for all other studies of the same stimulus type. If standard errors were provided we converted them to standard deviations (standard deviation = standard error \* √N). If data was available in figures, we used webplot digitizer to obtain it (Rohatgi, 2015). If data was not available in the paper we contacted authors and requested this data. We computed a Standardised Mean Difference using the formula (meanSubstance – meanNeutral / pooledSD), and the associated standard error. If data was provided as an interference score, we used the formula for SMD as (meanInterference – 0 / SDInterference). Given the lack of reporting of a correlation between substance and neutral RTs/Errors, we imputed a correlation of 0.59, in line with previous research (Balk et al, 2012).

If studies reported a Stroop effect across different drinking groups (e.g. Heavy and Light Drinkers: Adams et al 2012) we took the overall Stroop effect to reduce any variability in effect sizes from sampling. For smoking and drug-related Stroops we did not include the Stroop effect from non-using controls groups (e.g. Copersino et al, 2004). One study (Lusher et al, 2004) had an implausibly large effect size SMD > 50[[1]](#footnote-1) and was excluded.

Meta-analysis and -regressions were conducted in R, using the ‘metafor’ package. We used a random effects, restricted maximum likelihood model. Heterogeneity was measured using the I2 statistic (Higgins et al, 2003), with I2 > 50% representing moderate heterogeneity and > 75% indicative of substantial heterogeneity. We conducted moderator analyses on a number of design and analysis decisions including i) computerised vs paper-and-pencil tasks, ii) words vs pictures, iii) blocked vs mixed design, iv) inclusions of practice trials vs non-stated, v) inclusion of RT cut offs (upper or lower bound) vs no exclusion. Data and analyses files are on OSF [https://osf.io/fc8qn/]

**Results**

*Study selection*

After duplicates were removed, we retained 1073 articles for title and abstract screening. Following title and abstract screening 183 papers remained for full text screen. Of these, we identified n = 46 studies containing an Alcohol Stroop, n = 25 containing a Smoking Stroop and n = 24 containing a Drug-related Stroop. Full details of papers included is provided on Open Science Framework. The remaining studies did not meet our inclusion criteria (e.g. presenting an emotion Stroop with no alcohol / smoking cues, using a modified Stroop to manipulate attentional bias, duplicates that were not picked up by automatic filtering in EndNote / study protocols). PRISMA flowchart is shown in figure 1.

*Figure 1: PRISMA diagram outlining systematic searches of the literature*

A screenshot of a cell phone

Description automatically generated

**Alcohol Stroop**

We identified 46 studies (published from 1994 – 2019) which implemented a version of the addiction Stroop with alcohol-related stimuli.

*Variation in design / methodological variables (see Table 1)*

The majority of studies (N = 35) used a computer to administer the Alcohol Stroop, compared to a paper-and-pencil version (N = 10). One study did not report how the Stroop task was administered. All paper-and-pencil Stroop tasks required a voice response from the participant, whereas 26 of the computerised tasks required a key-press, and 7 required a verbal response. Two studies required both a voice response and corresponding key-press, however one study recorded only the voice response and the other recorded only the key press response.

The majority of the Stroop tasks used words (N = 43), with 3 using pictures. Twenty-five used a blocked design, whereas 16 used a mixed design. Two studies used a combination mixed and block design, and three studies did not report the design used. The majority (N = 31) included four-word colours (range 2 – 5). Furthermore, most studies reported the inclusion of practice trials (N = 33). The number of alcohol-related stimuli ranged from 4 – 28 (mean = 14.1), and the number of times the stimuli were presented ranged from 1 – 48 (mean = 6.4). There was considerable variability in control / comparison cues, which we attempted to categorise into 10 distinct themes (building-related, clothing-related, emotional, environment-related, household-related, music-related, office-related, soft-drinks, transport, other). The most common comparator was household-related words (N = 17)

Finally, there was considerable variability in the reporting of trial time outs in the computer administrated Stroops. Ten studies did not report this information. Of the studies that did, 12 required a response from the participant to move on to the next trial. Where a timeout was reported it ranged from 1500 ms – 6000 ms (mean = 2714 ms).

We estimate that based on previous research there are 1,451,520 potential iterations of the alcohol addiction Stroop task that could be designed for administration on a computer, and 11,520 paper-and-pencil versions[[2]](#footnote-2), based on previous research.

[table 1 here]

*Variation in analysis decisions (see Table 2).*

Paper-and-pencil Stroop tasks do not record reaction times or error rates on a trial-by-trial basis and are not included in the frequencies for lower / upper bound cut-offs or individual trial removals based on standard deviations. Twenty-five studies did not report a lower bound RT cut-off. Of those that did, cut-offs ranged from 100 ms – 400 ms, with one study not reporting what the cut off was (RT trimming was stated). Twenty-five studies did not report an upper-bound cut off. Of those that did, cut-offs ranged from 1500 ms – 10000 ms, with one study not reporting the cut off used. Thirty-two studies did not remove RTs if they were a number of SDs around the mean. Two studies removed RTs that were 3 SDs above the mean and one study removed RTs that were 3 SDs above or below the mean. Fourteen studies did not explicitly state trials on which errors were made were not removed, whereas 19 studies did. One study (Cox et al., 2003) stated that due to error infrequency they were not considered in analyses. One study recoded errors as the mean RT + 2 SDs.

Of all the studies, 42 did not remove any participants based on overall Stroop scores. Studies removed participants if they were >2, 3, or 4 SDs from the sample mean. One study removed participants based on ‘totally incorrect responses’ but did not provide further information.

Finally, the majority of all Stroop studies (N = 25) conducted primary analyses on Mean Reaction Times to alcohol-stimuli vs. comparison stimuli, with 1 study analysing medians rather than means. Twelve studies analysed difference / interference scores[[3]](#footnote-3) based on means, 1 analysed difference / interference scores based on medians and 1 analysed difference / interference scores based on D scores. Three studies used errors as a primary outcome (2 using raw error scores and 1 using difference / interference error scores). Two studies analysed individual reaction times in a general linear model.

We estimated based on the available research there are 9,000 potential analysis pipelines for the computerised alcohol Stroop.

[table 2 here]

*P-Curve for Stroop effect (online supplementary Fig1)*

We were able to extract 22 statistical tests for a Stroop effect (e.g. the test for difference in reaction times / errors / difference scores based on stimuli). Of these 22 significant statistical tests, 3 had a p value > .025. Continuous tests for half curve (Z=-12.97, p<.0001) and full curve (Z=-13.03, p<.0001) were statistically significant. The p-curve demonstrated support for evidential value.

*P-Curve for Moderation with or associations with drinking status / craving (online supplementary Fig2)*

We were able to extract 17 significant tests. Of these significant statistical tests, 6 had a p value > .025. Continuous tests for half curve (Z=-4.72, p<.001) and full curve (Z=-3.99, p<.001) were statistically significant. Half and Full curve tests remained significant (ps < .001) for robustness checks. The p-curve demonstrated support for evidential value.

**Smoking Stroop**

We identified 25 studies (published from 2006 – 2019) which implemented a version of the addiction Stroop with smoking-related stimuli.

*Variation in design / methodological variables (see Table 3)*

The majority of studies (N = 21) used a computer to administer the smoking Stroop, compared to a paper-and-pencil version (N = 4). All paper-and-pencil Stroop tasks required a voice response from the participant, whereas 16 of the computerised tasks required a key-press, and 3 required a verbal response. One study required both a voice response and corresponding key-press, and one study did not report the response required.

All but one (N = 20) of the computerised smoking Stroops used words. Eighteen used a blocked design, whereas 5 used a mixed design. Two studies did not report the design used. The majority of all Stroops (N=12) included three word-colours (range 3 – 5). Furthermore, most studies reported the inclusion of practice trials (N =14). The number of smoking-related stimuli ranged from 5 – 26 (mean = 12.8), and the number of times the stimuli were presented ranged from 1 – 10 (mean = 5.4). There was some variability in control / comparison cues, which we categorised into 3 distinct themes (household-related, mixed/other, unknown /not reported). The most common comparators were unknown / not reported (N = 9).

Finally, there was considerable variability in the reporting of trial timeouts. In the computerised versions 11 studies did not report this information. Of the studies that did, two required a response from the participant to move on to the next trial. Where a timeout was reported it ranged from 1200 ms – 4000 ms (mean = 2462 ms).

We estimate that based on previous research there are 77,760 potential iterations of the smoking addiction Stroop task that could be designed to be administered on a computer, and 1,080 paper-and-pencil versions, based on previous research.

[table 3 about here]

*Variation in analysis decisions (see Table 4).*

Ten studies did not report a lower-bound RT cut-off. Of the 11 that did that did, cut-offs ranged from 100 ms – 200 ms. Fourteen studies did not report an upper-bound RT cut-off, of those that did it ranged from 1000 – 2000 ms. Fourteen studies did not remove RTs if they were a number of SDs around the mean. One study removed RTs >2 SDs above and below the mean, one study removed RTs >2 SDs above the mean only, two studies removed RTs > 3 SDs above and below the mean, one study removed RTs > 2 SDs below and >3 SDs above the mean, one study replaced RTs > 3 SDs above the mean with mean RT + 3 SDs, and one study replaced RTs > 3 SDs above the mean with the mean RT. Thirteen studies did not explicitly state trials on which errors were made were not removed, whereas 7 did.

Twenty studies did not report removal of participant responses based on Stroop performance. Two studies removed participants who made >33% errors, 1 study removed participants who made >25% errors or a difference score was 4 standard deviations outside the sample mean, and 1 study removed participants with mean RTs outside three standard deviations from the sample mean.

Finally, the nine studies conducted primary analyses on Mean Reaction Times to alcohol-stimuli vs. comparison stimuli, with 1 study analysing medians rather than means. Eleven studies analysed difference / interference scores[[4]](#footnote-4) based on means, 2 analysed difference / interference scores based on medians. In 2 studies the main outcome measure was unclear.

We estimated based on the available research there are 5,376 potential analysis pipelines for the alcohol Stroop.

[table 4 about here]

*P-Curve for Stroop effect (see online supplementary Fig3)*

We were able to extract 13 statistical tests for a Stroop effect.

Of these 13 significant statistical tests, 0 had a p value > .025. Continuous tests for half curve (Z=-5.06, p<.0001) and full curve (Z=-6.90, p<.0001) were statistically significant. The p-curve demonstrated support for evidential value.

*P-Curve for Moderation with or associations with smoking status / craving (online supplementary Fig4).*

We were able to extract 5 significant tests. Of these significant statistical tests, 1 had a p value > .025. Continuous tests for half curve (Z=-1.71, p=.0437) and full curve (Z=-2.30, p=.0106) were statistically significant. The p-curve demonstrated some support for evidential value.

**Drug-related Stroop**

We identified 24 studies (published from 2000 – 2018) which implemented a version of the addiction Stroop with drug-related stimuli (either: mixed, cocaine-, methamphetamine-, heroin-, cannabis-related cues)

*Variation in design / methodological variables (see Table 5)*

The majority of studies (N = 21) used a computer to administer the drug Stroop, compared to a paper-and-pencil version (N = 3). All paper-and-pencil Stroop tasks required a voice response from the participant, whereas 19 of the computerised tasks required a key-press, and 2 required a verbal response.

Twenty-two of the computerised drug-related Stroops used words (1 used images and 1 used both). Twelve used a blocked design, whereas 10 used a mixed design. Two studies did not report the design used. The majority of all Stroops (N=17) included four word-colours (range 3 – 4). Furthermore, most studies reported the inclusion of practice trials (N = 15). The number of smoking-related stimuli ranged from 2 – 319 (mean = 11.2[[5]](#footnote-5)), and the number of times the stimuli were presented ranged from 1 – 32 (mean = 6.6). There was some variability in control / comparison cues, which we categorised into 8 distinct themes. The most common comparators were mixed (N = 9).

Finally, there was considerable variability in the reporting of trial timeouts. In the computerised versions 5 studies did not report this information. Of the studies that did, one required a response from the participant to move on to the next trial. Where a timeout was reported it ranged from 1750 ms – 6000 ms (mean = 3170 ms).

We estimate that based on previous research there are 112,640 potential iterations of the drug addiction Stroop task that could be designed to be administered on a computer, and 2,816 paper-and-pencil versions, based on previous research.

*[table 5 about here]*

*Variation in analysis decisions (see Table 6)*

Thirteen studies did not report a lower-bound RT cut-off. Of the 8 that did that did, cut-offs ranged from 100 ms – 200 ms. Nineteen studies did not report an upper-bound RT cut-off, the two that did were 1500 ms and 3000 ms. One study removed RTs >3 SDs above and below the mean. Eleven studies did not explicitly state trials on which errors were made were not removed.

Twenty studies did not report removal of participant responses based on Stroop performance. One study removed participants with RTs > 3 SDs above the mean, one study removed participants with >25% reaction times < 200 ms, and one study removed participants with ‘excessive errors’. A further study removed all participants with a negative inference score (e.g. no attentional bias), although this was to test a specific theoretical prediction.

Finally, fourteen studies conducted primary analyses on Mean Reaction Times to drug-related-stimuli vs. comparison stimuli, with 1 study analysing mean reaction times in the first blocks of the task only. Seven studies analysed difference / interference scores based on means, and 2 analysed difference / interference scores based on medians.

We estimated based on the available research there are 768 potential analysis pipelines for the alcohol Stroop.

*[table 6 about here]*

*P-Curve for Drug Stroop effect (see online supplementary Fig5)*

We were able to extract 10 statistical tests for a Stroop effect. Of these 10 significant statistical tests, 0 had a p value > .025. Continuous tests for half curve (Z=-9.07, p<.0001) and full curve (Z=-10.65, p<.0001) were statistically significant. The p-curve demonstrated support for evidential value

*P-Curve for Moderation with or associations with smoking status / craving (online supplementary Fig6).*

We were able to extract 8 significant tests. Of these significant statistical tests, 7 had a p value < .025. Continuous tests for half curve (Z=-2.25, p<.01) and full curve (Z=-3.75, p<.001) were statistically significant. Robustness checks demonstrated the half curve p value changed to (p = .002), whilst full curve remained < .001. The p-curve demonstrated support for evidential value.

***Exploratory Meta-analyses***

We were able to extract k = 77 effect sizes relating to the alcohol, smoking and drug Stroops. The pooled effect size was SMD = 0.23 [95% CI = 0.17; 0.29], Z = 7.29, p < .001], see supplementary Figure 7 for forest plot. There was considerable heterogeneity across effect sizes (I2 = 80%). Leave-one-out analyses demonstrated the pooled effect was not largely influenced by one effect size (SMDs ranged between .22 to .23, all p values < .001). There was no evidence of moderation of the effect size for the majority of our extracted variables (see supplementary Table 1). There was weak evidence increased number of alcohol, smoking, drug-related stimuli was associated (X2(1) = 4.19, p = .040) with larger effect sizes (coefficient = .002 (95% CI: .001 to .004)).

**DISCUSSION**

The aim of this study was to quantify the design and analysis decisions a researcher might make when utilising the addiction Stroop task. Based on available information in the literature a considerable number of potential design and analysis decisions can be taken for each task. Given these decisions, it is possible that based on previous information a researcher could design 1,451,520 different computerised Stroop tasks to assess alcohol attentional bias, 77,760 to assess smoking attentional bias and 112,640 to assess attentional bias to drug-related cues. This makes the number of tasks that are precisely the same very limited across studies, which in turn limits their generalisability. Whilst researchers often have to make difficult decisions on task design, which might be constrained by pragmatic (e.g. time) or methodological issues, small changes to task design can considerably influence psychometric properties (Cooper et al., 2017).

Similarly, we found evidence of analytical flexibility (‘*researcher degrees of freedom*’) when examining the data from the addiction Stroops. These analysis pipelines may increase the risk of false-positive findings in the literature (Ioannidis, 2005; Simmons et al., 2011), which might also partially explain the inconsistent findings reported across attentional bias studies (Christiansen et al., 2015; Heitmann et al., 2018). Encouragingly our p-curve analyses suggest evidential value and limited evidence of p-values clustering around the commonly accepted threshold for statistical significance (p < .05; (Chavalarias et al., 2016)). This suggests that researchers are unlikely to be utilising analysis pipelines in search of a ‘statistically significant’ effect, known as p-hacking. However, we note that the p-curve is not immune to p-hacking if parallel analyses are conducted and the ‘strongest’ result is chosen (Ulrich & Miller, 2015), rather than conducting sequential analyses until the p < .05 is reached.

A wider issue identified in our review is the lack of reporting for key methodological and analytical variables. Whilst this issue is not limited to the Stroop task (it has been reported across psychology, ranging from fMRI research (Carp, 2012b) to human laboratory studies of eating behaviour (Robinson et al., 2018)) it does limit the ability of researchers to attempt direct replications (Brandt et al., 2014). In line with many of the positive reforms in psychology, researchers should endeavour to establish guidelines or a checklist on reporting of the addiction Stroop and cognitive tests more generally, alongside reliability estimates (Parsons et al., 2019). Here we have examined some key variables which should at the very least be reported however, we invite researchers to add to this list and strive toward transparent and comprehensive reporting. Similar efforts should also be undertaken for other tasks designed to measure attentional bias (e.g. Visual Probe).

There is also a lack of research into the psychometric properties of the emotional/addictive Stroop, which should be surprising given their widespread use. Very few studies reported on the psychometric properties of the task used. In a comprehensive analysis Hedge et al (2018) demonstrated that many cognitive tasks including the standard Stroop offer poor reliability, below what is acceptable for research into individual differences. These observations were replicated by Wilson et al (Wilson et al., 2019) when examining the emotional Stroop with suicide related words (*r*s = -.09-.13), suggesting unacceptable measurement error in these tasks.

Finally, we note that the vast majority of design and analysis decisions made by researchers were not justified (this has also been noted for cognitive tasks elsewhere, specifically for trial numbers ‘*different studies can choose very different numbers without any explicit discussion or justification*’ Hedges et al (2019, p. 1174). There was some consistency in the design and analysis decisions within research groups suggesting heritability passed from peers / mentors (discussed in Smaldino & McElreath, 2016). Whilst this may serve to increase productivity as less effort is spend designing and validating new research materials, it can also increase the proliferation of inadequate practices and increase false positives. When analysis decisions were justified with a reference it was often inaccurately attributed. For example, studies often cited Ratcliffe (Ratcliff, 1993) as justification for implementing RT cut-offs, however did not provide any information on distribution testing, amount of data removed or comparing their method with other methods to determine whether it was optimal. All of which are recommended in the original paper.

Our exploratory meta-analysis demonstrated a small but robust alcohol Stroop effect, in line with previous meta-analyses (Cox et al, 2006). Our examination of design and analytic variables as potential moderators of the effect suggested limited influence. There was weak evidence that a larger number of unique substance-related cues increased the Stroop effect, this may be due to the reduced likelihood of habitation (and attenuating of any biases) to a larger number of images (Hall & Rodriquez, 2017). However, examination of each potential variable in isolation may not lead to substantial differences in effect sizes across studies, due to myriad other factors which also contribute to the phenomenon. The combination of analytic / methodological decisions may substantially influence effect sizes (for example, a small number of trial repetitions combined with no practice trials and no reaction time trimming might lead to effect sizes computed on highly variable RTs). Future avenues of research might include multiverse analyses to examine the variability of effect sizes and statistical significance within studies, as a result of design and methodological decisions (see Steegan et al., 2016). Importantly, if the overall Stroop effect remains irrespective of different methodologies and analysis pathways (known as conceptual replication: Crandell & Sherman, 2016) then this can increase our confidence that the effect is robust. Such conceptual replications, as well as direct replications, are important in overcoming the replication crisis.

Our review is not without limitations. First, we did not analyse data from Stroop tasks which were designed to measure attentional bias to illicit drugs due to the limited number in the literature. It is entirely plausible we have not provided an exhaustive list of potential design and analysis decisions researchers might make, as we decided to focus on prominent and previously discussed variables. Indeed it is possible that variables such as inter-trial-intervals might influence Stroop performance, based on research in other cognitive domains (Auchter et al., 2017; Cooper et al., 2017). Similarly, the counterbalancing (or not) of blocked designs may also moderate Stroop performance (Waters & Feyerabend, 2000). We also noted (but did not analyse) variability in the number of different category comparisons, e.g. alcohol vs. neutral vs. positive vs. negative words (Fridrici et al., 2013). Increasing the number of comparisons serves to increase complexity and length of the task which may also influence psychometric properties. Furthermore, for a number of variables (e.g. lower / upper-bound RT cut-offs) it is impossible to infer whether lack of reporting is due to these procedures not being carried out or careless / undisclosed reporting. Therefore, it is possible our estimations of analysis pipelines are conservative, with the potential that they are even larger than reported here.

**Concluding remarks**

This is the first study to examine in detail the variability in the methodological and analytic decisions researchers might be faced with when conducting research into attentional bias using the addiction Stroop. The observed variability means a large number of different tasks can be designed and analyses carried out by justifying decisions based on published literature. This has a number of implications for the reproducibility and reliability of the attentional bias research in addiction. Despite this, researchers should have some confidence that attentional bias is a seemingly robust phenomenon. We observed no evidence of p-hacking (at least in the studies we identified), and effect sizes demonstrating the presence of bias towards-substance related cues were unaffected by the identified analysis/methodological choices. Nevertheless, in order to achieve further progress in the field researchers should develop and adopt stringent reporting guidelines, and investigate whether methodological and analytic decisions influence the psychometric properties of the addiction Stroop.

**Funding details:**

No funding.

**Disclosure statement:**

The authors report no conflicts of interest

**References:**

Ashley, V., Honzel, N., Larsen, J., Justus, T., & Swick, D. (2013). Attentional bias for trauma-related words: Exaggerated emotional Stroop effect in Afghanistan and Iraq war veterans with PTSD. *BMC Psychiatry*, *13*(1), 86. https://doi.org/10.1186/1471-244X-13-86

Ataya, A. F., Adams, S., Mullings, E., Cooper, R. M., Attwood, A. S., & Munafò, M. R. (2012). Internal reliability of measures of substance-related cognitive bias. *Drug and Alcohol Dependence*, *121*(1–2), 148–151. https://doi.org/10.1016/j.drugalcdep.2011.08.023

Auchter, A., Cormack, L. K., Niv, Y., Gonzalez-Lima, F., & Monfils, M. H. (2017). Reconsolidation-Extinction Interactions in Fear Memory Attenuation: The Role of Inter-Trial Interval Variability. *Frontiers in Behavioral Neuroscience*, *11*. https://doi.org/10.3389/fnbeh.2017.00002

Baugh, F. (2002). Correcting Effect Sizes for Score Reliability: A Reminder that Measurement and Substantive Issues are Linked Inextricably. *Educational and Psychological Measurement*, *62*(2), 254–263. https://doi.org/10.1177/0013164402062002004

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van ’t Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Cane, J., Sharma, D., & Albery, I. (2009). The addiction Stroop task: Examining the fast and slow effects of smoking and marijuana-related cues. *Journal of Psychopharmacology*, *23*(5), 510–519. https://doi.org/10.1177/0269881108091253

Carp, J. (2012a). On the Plurality of (Methodological) Worlds: Estimating the Analytic Flexibility of fMRI Experiments. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00149

Carp, J. (2012b). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage*, *63*(1), 289–300. https://doi.org/10.1016/j.neuroimage.2012.07.004

Chavalarias, D., Wallach, J. D., Li, A. H. T., & Ioannidis, J. P. A. (2016). Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. *JAMA*, *315*(11), 1141–1148. https://doi.org/10.1001/jama.2016.1952

Christiansen, P., Schoenmakers, T. M., & Field, M. (2015). Less than meets the eye: Reappraising the clinical relevance of attentional bias in addiction. *Addictive Behaviors*, *44*, 43–50. https://doi.org/10.1016/j.addbeh.2014.10.005

Cooper, S. R., Gonthier, C., Barch, D. M., & Braver, T. S. (2017). The Role of Psychometrics in Individual Differences Research in Cognition: A Case Study of the AX-CPT. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01482

Cox, W. M., Fadardi, J. S., & Pothos, E. M. (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

*Crandall, C., Sherman, J. (2016). On the scientific superiority of conceptual replications for scientific progress. Journal of Experimental Social Psychology, 66, 93-99.*

Cristea, I. A., Kok, R. N., & Cuijpers, P. (2016). The Effectiveness of Cognitive Bias Modification Interventions for Substance Addictions: A Meta-Analysis. *PLoS ONE*, *11*(9). https://doi.org/10.1371/journal.pone.0162226

Fadardi, J. S., & Cox, W. M. (2009). Reversing the sequence: Reducing alcohol consumption by overcoming alcohol attentional bias. *Drug and Alcohol Dependence*, *101*(3), 137–145. https://doi.org/10.1016/j.drugalcdep.2008.11.015

Field, M, & Cox, W. (2008). Attentional bias in addictive behaviors: A review of its development, causes, and consequences. *Drug and Alcohol Dependence*, *97*(1–2), 1–20. https://doi.org/10.1016/j.drugalcdep.2008.03.030

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

Field, Matt, Christiansen, P., Cole, J., & Goudie, A. (2007). Delay discounting and the alcohol Stroop in heavy drinking adolescents. *Addiction*, *102*(4), 579–586. https://doi.org/10.1111/j.1360-0443.2007.01743.x

Field, Matt, Duka, T., Tyler, E., & Schoenmakers, T. (2009). Attentional bias modification in tobacco smokers. *Nicotine & Tobacco Research*, *11*(7), 812–822. https://doi.org/10.1093/ntr/ntp067

Field, Matt, & Eastwood, B. (2005). Experimental manipulation of attentional bias increases the motivation to drink alcohol. *Psychopharmacology*, *183*(3), 350–357. https://doi.org/10.1007/s00213-005-0202-5

Field, Matt, Marhe, R., & Franken, I. H. A. (2014). The clinical relevance of attentional bias in substance use disorders. *CNS Spectrums*, *19*(3), 225–230. https://doi.org/10.1017/S1092852913000321

Field, Matt, Mogg, K., Mann, B., Bennett, G. A., & Bradley, B. P. (2013). Attentional biases in abstinent alcoholics and their association with craving. *Psychology of Addictive Behaviors*, *27*(1), 71–80. https://doi.org/10.1037/a0029626

Field, Matt, Munafò, M. R., & Franken, I. H. A. (2009). A Meta-Analytic Investigation of the Relationship Between Attentional Bias and Subjective Craving in Substance Abuse. *Psychological Bulletin*, *135*(4), 589–607. https://doi.org/10.1037/a0015843

Fridrici, C., Leichsenring-Driessen, C., Driessen, M., Wingenfeld, K., Kremer, G., & Beblo, T. (2013). The individualized alcohol Stroop task: No attentional bias toward personalized stimuli in alcohol-dependents. *Psychology of Addictive Behaviors: Journal of the Society of Psychologists in Addictive Behaviors*, *27*(1), 62–70. https://doi.org/10.1037/a0029139

Hall, G., Rodriguez (2017). Habituation and Conditioning: Salience Change in Associative Learning. Journal of Experimental Psychology: Animal Learning and Cognition, 43, 48-61.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Heitmann, J., Bennik, E. C., van Hemel-Ruiter, M. E., & de Jong, P. J. (2018). The effectiveness of attentional bias modification for substance use disorder symptoms in adults: A systematic review. *Systematic Reviews*, *7*. https://doi.org/10.1186/s13643-018-0822-6

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

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

Kanyongo, G., Brook, G., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and Statistical Power: How Measurement Fallibility Affects Power and Required Sample Sizes for Several Parametric and Nonparametric Statistics. *Journal of Modern Applied Statistical Methods*, *6*(1). https://doi.org/10.22237/jmasm/1177992480

Kerst, W. F., & Waters, A. J. (2014). Attentional retraining administered in the field reduces smokers’ attentional bias and craving. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *33*(10), 1232–1240. https://doi.org/10.1037/a0035708

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science (New York, N.Y.)*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Lusher, J., Chandler, C., & Ball, D. (2004). Alcohol dependence and the alcohol Stroop paradigm: Evidence and issues. *Drug and Alcohol Dependence*, *75*(3), 225–231. https://doi.org/10.1016/j.drugalcdep.2004.03.004

MacLeod, C., Mathews, A., & Tata, P. (1986). Attentional bias in emotional disorders. *Journal of Abnormal Psychology*, *95*(1), 15–20. https://doi.org/10.1037/0021-843X.95.1.15

Marissen, M. A. E., Franken, I. H. A., Waters, A. J., Blanken, P., van den Brink, W., & Hendriks, V. M. (2006). Attentional bias predicts heroin relapse following treatment. *Addiction (Abingdon, England)*, *101*(9), 1306–1312. https://doi.org/10.1111/j.1360-0443.2006.01498.x

Mourik, R. V., Oosterlaan, J., & Sergeant, J. A. (2005). The Stroop revisited: A meta-analysis of interference control in AD/HD. *Journal of Child Psychology and Psychiatry*, *46*(2), 150–165. https://doi.org/10.1111/j.1469-7610.2004.00345.x

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1–9. https://doi.org/10.1038/s41562-016-0021

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements: *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/2515245919879695

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, *114*(3), 510–532. https://doi.org/10.1037/0033-2909.114.3.510

Robbins, S. J., & Ehrman, R. N. (2016). The Role of Attentional Bias in Substance Abuse: *Behavioral and Cognitive Neuroscience Reviews*. https://doi.org/10.1177/1534582305275423

Robinson, E., Bevelander, K. E., Field, M., & Jones, A. (2018). Methodological and reporting quality in laboratory studies of human eating behavior. *Appetite*, *125*, 486–491. https://doi.org/10.1016/j.appet.2018.02.008

Shapiro, K. L., Raymond, J. E., & Arnell, K. M. (1997). The attentional blink. *Trends in Cognitive Sciences*, *1*(8), 291–296. https://doi.org/10.1016/S1364-6613(97)01094-2

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., … 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*. https://doi.org/10.1177/2515245917747646

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*. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L., Simmons, J. (2014). P-curve: a key to the file-drawer. Journal of Experimental Psychology: General, 143, 534-47.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology. General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

*Steegan, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. Perspectives in Psychology Science, 11, 702-712.*

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662. https://doi.org/10.1037/h0054651

Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General, 144*, 1137–1145

Verhaeghen, P., & De Meersman, L. (1998). Aging and the Stroop effect: A meta-analysis. *Psychology and Aging*, *13*(1), 120–126. https://doi.org/10.1037/0882-7974.13.1.120

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

Waters, A. j., & Feyerabend, C. (2000). Determinants and effects of attentional bias in smokers. *Psychology of Addictive Behaviors*, *14*(2), 111–120. https://doi.org/10.1037/0893-164X.14.2.111

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. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01832

Wiers, R. W., Gladwin, T. E., Hofmann, W., Salemink, E., & Ridderinkhof, K. R. (2013). Cognitive Bias Modification and Cognitive Control Training in Addiction and Related Psychopathology: Mechanisms, Clinical Perspectives, and Ways Forward. *Clinical Psychological Science*, *1*(2), 192–212. https://doi.org/10.1177/2167702612466547

Wilson, K. M., Millner, A. J., Auerbach, R. P., Glenn, C. R., Kearns, J. C., Kirtley, O. J., Najmi, S., O’Connor, R. C., Stewart, J. G., & Cha, C. B. (2019). Investigating the psychometric properties of the Suicide Stroop Task. *Psychological Assessment*, *31*(8), 1052–1061. https://doi.org/10.1037/pas0000723

Xu, Y., Li, Y., Wang, G., Yuan, X., Ding, W., & Shen, Z. (2014). Attentional bias toward safety predicts safety behaviors. *Accident Analysis & Prevention*, *71*, 144–153. https://doi.org/10.1016/j.aap.2014.05.013

Zhang, M., Fung, D. S. S., & Smith, H. (2019). Variations in the Visual Probe Paradigms for Attention Bias Modification for Substance Use Disorders. *International Journal of Environmental Research and Public Health*, *16*(18). https://doi.org/10.3390/ijerph16183389

**Table 1: Extracted data of the design decisions of the Alcohol Stroop**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Lead author** | **Year** | **Admin** | **Response** | **N stimuli** | **Times presented** | **Word/picture** | **Block / mixed** | **Control stimuli** | **Colours** | **Practice** | **TIMEOUT (MS)** |
| Adams | 2012 | Comp | Key | 20 | 4 | Words | Both | Music | 4 | Yes | 2500 |
| Albery | 2015 | Comp | Key | 25 | 4 | Words | Block | Environment | 4 | Yes | UR |
| Bailey | 2016 | Comp | Key | 4 | 48 | Words | Mixed | Household | 4 | NR | UR |
| Bauer | 1998 | Comp | Voice | 10 | 8 | Words | Mixed | Household | 4 | Yes | 1500 |
| Bruce | 2004 | Comp | Key | 10 | 12 | Pictures | Mixed | Household | 3 | Yes | UR |
| Carrigan | 2004 | Comp | Key | 10 | 2 | Words | Block | Household | 4 | Yes | 3000 |
| Choi | 2015 | Comp | Key | 8 | 3 | Words | NR | Household | 3 | Yes | NR |
| Christiansen | 2014 | Paper | Voice | 8 | 14 | Words | Block | Soft drinks | 4 | NR | NA |
| Cox | 1999 | Comp | Key | 20 | 1 | Words | Block | Music | 4 | Yes | UR |
| Cox | 2002 | Comp | Voice | 10 | 3 | Words | Block | Household | 4 | Yes | UR |
| Cox | 2003 | Paper | Voice | 20 | 4 | Words | Block | Household | 4 | Yes | NA |
| Cox | 2007 | Comp | Key | NR | NR | Words | Mixed | Household | 4 | Yes | NR |
| Duka | 2002 | Paper | Voice | 8 | NR | Words | Block | Emotional | NR | NR | NA |
| Duka | 2004 | Paper | Voice | 20 | NR | Words | Block | Office | NR | NR | NA |
| Fadardi | 2006 | Comp | Key | 28 | 4 | Words | Mixed | Building | 4 | Yes | 3000 |
| Fadardi | 2009 | Comp | Key | 7 | 8 | Words | Mixed | Household | 4 | Yes | 3000 |
| Field | 2013 | Paper | Voice | 25 | 4 | Words | Block | Environment | 4 | Yes | NA |
| Field | 2007 | Paper | Voice | 20 | 5 | Words | Block | Music | 4 | Yes | NA |
| Field | 2007 | Paper | Voice | 20 | 5 | Words | Block | Music | 4 | Yes | NA |
| Flaudias | 2013 | Comp | Voice | 4 | 6 | Words | Block | Mixed | 3 | Yes | NR |
| Fridrici | 2013 | Comp | Key | 8 | 8 | Words | Block | Mixed | 4 | Yes | 1500 |
| Garland | 2012 | Comp | Key | 20 | 3 | Pictures | Mixed | Household | 3 | NR | NR |
| Grant | 2007 | Comp | Voice | 20 | 2 | Words | Mixed | Clothing | 5 | Yes | 2000 |
| Johnsen | 1994 | Comp | Both | 20 | 1 | Words | Block | Household | 4 | NR | 6000 |
| Jones | 2000 | Comp | Both | 24 | 10 | Words | Block | Emotional | 2 | Yes | UR |
| Klein | 2007 | Comp | Key | 8 | 3 | Words | Mixed | Mixed | 2 | Yes | UR |
| Luehring-Jones | 2017 | Comp | Key | 25 | 4 | Words | Block | Environment | 4 | NR | NR |
| Lusher | 2004 | Comp | Key | 8 | 8 | Words | Mixed | Household | 4 | Yes | UR |
| Moss | 2013 | NR | NR | NR | NR | Words | NR | Environment | NR | NR | NR |
| Murphy | 2011 | Comp | Key | 20 | 1 | Words | Block | Music | 4 | Yes | UR |
| Oliver | 2015 | Comp | Key | 10 | 3 | Words | Mixed | Household | 3 | NR | 2000 |
| Ostafin | 2019 | Comp | Key | 10 | 9 | Words | Block | Music | 3 | Yes | UR |
| Read | 2017 | Comp | Voice | NR | NR | Words | Block | Household | 4 | Yes | 1500 |
| Rofey | 2007 | Comp | Key | 8 | NR | Words | NR | Mixed | 4 | NR | NR |
| Rose | 2008 | Paper | Voice | 8 | 4 | Words | Block | Office | 4 | Yes | NA |
| Ryan | 2002 | Paper | Voice | 5 | 10 | Words | Block | Household | 4 | NR | NA |
| Shamloo | 2014 | Comp | Key | 14 | NR | Words | Mixed | Household | 4 | Yes | NR |
| Sharma | 2001 | Comp | Key | 25 | 4 | Words | Mixed | Environment | 4 | Yes | NR |
| Snelleman | 2015 | Comp | Key | 11 | 3 | Words | Block | Office | 3 | Yes | NR |
| Spanakis | 2018 | Comp | Key | 11 | 3 | Pictures | Block | Soft drinks | 3 | NR | 3000 |
| Steinberg | 2011 | Comp | Voice | 20 | NR | Words | Mixed | Music | 4 | Yes | UR |
| Stetter | 1995 | Paper | Voice | 10 | 10 | Words | Block | Household | 4 | Yes | NA |
| Stewart | 2002 | Comp | Voice | 20 | NR | Words | Mixed | Clothing | 5 | Yes | 2000 |
| Stormark | 2000 | Comp | Key | 4 | 4 | Words | Block | Mixed | 4 | NR | 4000 |
| Van den wildenberg | 2006 | Comp | Key | 8 | 10 | Words | Both | Transport | 4 | Yes | 3000 |
| Zetteler | 2006 | Comp | Key | 12 | 1 | Words | Mixed | Mixed | 4 | Yes | UR |
| ***Variations*** |  |  | ***3*** | ***12*** | ***12*** | ***2*** | ***3*** | ***10*** | ***4*** | ***2*** | ***7*** |

*Legend: Comp = computerised; Paper = paper and pencil; NR = not reported; NA = not applicable; UR = until response*

**Table 2: Extracted data of the analysis decisions on the Alcohol Stroop task**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Lead Author** | **Year** | **Lower bound RT cut off (ms)** | **Upper bound RT cut off (ms)** | **SD removal** | **Exclude errors** | **Primary outcome** | **Participants removed** |
| Adams | 2012 | NR | NR | NR | Yes | Raw RT (median) | NR |
| Albery | 2015 | NR | NR | NR | Yes | Raw RT (mean) | NR |
| Bailey | 2016 | 100 | 2000 | NR | NR | Raw RT (mean) | NR |
| Bauer | 1998 | NR | NR | NR | NR | Difference (mean RT) | NR |
| Bruce | 2004 | 300 | 1500 | NR | Yes | Difference (mean RT) | NR |
| Carrigan | 2004 | NR | NR | NR | Yes | Raw RT (mean) | NR |
| Choi | 2015 | NR | NR | NR | NR | Difference (mean RT) | NR |
| Christiansen | 2014 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Cox | 1999 | NR | NR | NR | NR | Raw RT (mean) | NR |
| Cox | 2002 | NR | NR | NR | NR | Raw RT (mean) | NR |
| Cox | 2003 | NA | NA | NA | NO\* | Raw RT (mean) | YES (2SD of Sample) |
| Cox | 2007 | NR | NR | NR | Yes | Difference (mean RT) | NR |
| Duka | 2002 | NR | NR | NR | NR | Errors | NR |
| Duka | 2004 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Fadardi | 2006 | 400 | 2000 | NR | Yes | Difference (mean RT) | NR |
| Fadardi | 2009 | NR | NR | NR | NR | Difference (mean RT) | NR |
| Field | 2013 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Field | 2007 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Field | 2007 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Flaudias | 2013 | 300 | 1500 | NR | NR | Raw RT (mean) | NR |
| Fridrici | 2013 | NR | NR | NR | Yes | Raw RT (mean) | NR |
| Garland | 2012 | TRIM | TRIM | 3 above | Yes | Difference (mean RT) | NR |
| Grant | 2007 | NR | NR | NR | NR | Difference (mean RT) | NR |
| Johnsen | 1994 | NR | 6000 | NR | NR | Raw RT (mean) | YES (incorrect responses) |
| Jones | 2000 | NR | NR | NR | Yes | Difference (median RT) | NR |
| Klein | 2007 | NR | NR | NR | Yes | Raw RT (mean) | YES (4 SD of sample) |
| Luehring-Jones | 2017 | NR | NR | NR | Yes | Difference (mean RT) | NR |
| Lusher | 2004 | NR | NR | NR | Yes | Raw RT (mean) | NR |
| Moss | 2013 | NR | NR | NR | NR | Raw RT (mean) | NR |
| Murphy | 2011 | NR | NR | NR | Yes | Difference (mean RT) | NR |
| Oliver | 2015 | 200 | NR | 3 above/below | Yes | Single RT (MLM) | NR |
| Ostafin | 2019 | 200 | 2000 | NR | Yes | Difference (errors) | YES (error 3SD of Sample) |
| Read | 2017 | 200 | NR | NR | Yes | Single RT (MLM) | NR |
| Rofey | 2007 | NR | NR | NR | NR | NR | NR |
| Rose | 2008 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Ryan | 2002 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Shamloo | 2014 | NR | NR | NR | Yes | Raw RT (mean) | NR |
| Sharma | 2001 | NR | NR | NR | NR | Errors | NR |
| Snelleman | 2015 | 300 | 10000 | NR | Correct+2SD | Difference (d-scores) | NR |
| Spanakis | 2018 | 200 | 2000 | 3 above | Yes | Difference (mean RT) | NR |
| Steinberg | 2011 | NR | NR | NR | NR | Raw RT (mean) | NR |
| Stetter | 1995 | NA | NA | NA | NR | Difference (mean RT) | NR |
| Stewart | 2002 | NR | 2000 | NR | Yes | Raw RT (mean) | NR |
| Stormark | 2000 | NR | NR | NR | NR | Raw RT (mean) | NR |
| van den Wildenberg | 2006 | NR | NR | NR | Yes | Raw RT (mean) | NR |
| Zetteler | 2006 | NR | NR | NR | NR | Raw RT (mean) | NR |
| ***Variations*** |  | 5 | 5 | 3 | 3 | 8 | 5 |

*Legend: MLM= multilevel model; NR = not reported; coded as NR in variation.*

**Table 3: Extracted data of the design decisions of the Smoking Stroop**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Lead Author** | **Year** | **Admin** | **Response** | **N Stimuli** | **Times presented** | **Words / Pictures** | **Block / Mixed** | **Control stimuli** | **Colours** | **Practice** | **Timeout (ms)** | |
| Begh | 2016 | Comp | Key | 12 | 8 | Pictures | Block | Neutral (unknown) | 4 | Yes | 2000 |
| Canamar | 2012 | Comp | Key | 26 | 1 | Words | Block | Mixed | 3 | Yes | 2000 |
| Cane | 2009 | Comp | Key | 5 | NR | Words | Block | Mixed | 4 | Yes | UR |
| Field | 2007 | Paper | Voice | 12 | 8 | Words | Block | Household | 4 | Yes | NA |
| Freeman | 2014 | Paper | Voice | NR | NR | Words | Block | Neutral (unknown) | NR | NR | NA |
| Fucito | 2010 | Comp | Key | NR | NR | Words | Block | Household | 4 | NR | NR |
| Greenaway | 2012 | Paper | Voice | 12 | 8 | Words | Block | Household | 4 | Yes | NA |
| Hendricks | 2006 | Comp | Key | 22 | NR | Words | Block | Neutral (unknown) | 3 | Yes | NR |
| Hitsman | 2007 | Paper | Voice | 9 | 10 | Words | Block | Household | 5 | NR | NA |
| Janes | 2010 | Comp | Key | 11 | 6 | Words | Block | Neutral (unknown) | 3 | Yes | 3000 |
| Janes | 2010 | Comp | Key | 11 | 6 | Words | Block | Neutral (unknown) | 3 | Yes | 3000 |
| Klein | 2009 | Comp | Voice | 10 | 3 | Words | Block | Mixed | 4 | NR | 1200 |
| Larsen | 2014 | Comp | Key | 8 | 4 | Words | Mixed | Mixed | 4 | Yes | 4000 |
| Masiero | 2019 | Comp | Voice | 18 | NR | Words | NR | Mixed | NR | NR | NR |
| McCarthy | 2009 | Comp | Voice | 20 | 4 | Words | Block | Mixed | 3 | Yes | 1500 |
| Munafo | 2007 | Comp | NR | 12 | 6 | Words | Block | Household | NR | Yes | NR |
| Munafo | 2008 | Comp | Key | 12 | 6 | Words | Block | Household | NR | Yes | NR |
| Poltavski | 2012 | Comp | Key | 11 | 3 | Words | NR | Mixed | 3 | Yes | NR |
| Robinson | 2015 | Comp | Key | NR | NR | Words | Mixed | Neutral (unknown) | 3 | NR | NR |
| Robinson S2 | 2015 | Comp | Key | NR | NR | Words | Mixed | Not reported | NR | NR | NR |
| Rzetelny | 2008 | Comp | Both | 12 | NR | Words | Block | Neutral (unknown) | NR | NR | NR |
| Sofuoglu | 2008 | Comp | Key | NR | NR | Words | Mixed | Neutral (unknown) | 3 | NR | NR |
| Waters | 2014 | Comp | Key | 11 | 3 | Words | Block | Household | 3 | Yes | 3000 |
| Waters | 2009 | Comp | Key | NR | NR | Words | Block | Neutral (unknown) | 3 | NR | UR |
| Waters | 2009 | Comp | Key | 10 | NR | Words | Mixed | Household | 3 | NR | NR |
| ***Variations*** |  |  | **3** | **10** | **6** | **2** | **2** | **3** | **3** | **2** | **6** |

*Legend: Comp = computerised; Paper = paper and pencil; NR = not reported; NA = not applicable; UR = until response*

**Table 4: Extraction table of Analysis Decisions of the Smoking Stroop.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Lead Author** | **Year** | **Lower bound RT cut off (ms)** | **Upper bound RT cut off (ms)** | **SD removal** | **Exclude errors** | **Primary outcome** | **Participants removed** |
| Begh | 2016 | NR | NR | NR | NR | Difference (median RT) | >3 SD Population |
| Canamar | 2012 | 200 | 1500 | 2 above / below | NR | Raw RT (mean) | NR |
| Cane | 2009 | NR | NR | NR | NR | Raw RT (median) | NR |
| Field | 2007 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Freeman | 2014 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Fucito | 2010 | 100 | 1500 | NR | Yes | Difference (mean RT) | NR |
| Greenaway | 2012 | NA | NA | NA | NR | Difference (mean RT) | NR |
| Hendricks | 2006 | NR | NR | NR | NR | NR | NR |
| Hitsman | 2007 | NA | NA | NA | NR | Difference (mean RT) | NR |
| Janes | 2010 | 150 | 1500 | 3 above / below | NR | Difference (mean RT) | NR |
| Janes | 2010 | 150 | 1500 | 3 above / below | NR | Difference (mean RT) | NR |
| Klein | 2009 | NR | NR | NR | NR | Difference (mean RT) | Latencies > 1500 |
| Larsen | 2014 | 200 | 2000 | NR | NR | Difference (mean RT) | NR |
| Masiero | 2019 | NR | NR | NR | NR | Raw RT (mean) | NR |
| McCarthy | 2009 | 200 | NR | 3 (replaced by mean + 3SD) | Yes | Raw RT (mean) | NR |
| Munafo | 2007 | 200 | 2000 | 2 above | Yes | Raw RT (mean) | NR |
| Munafo | 2008 | 200 | 1000 | 2 above, 3 below | Yes | Difference (mean RT) | NR |
| Poltavski | 2012 | NR | NR | NR | NR | Raw RT (mean) | NR |
| Robinson | 2015 | NR | NR | NR | NR | Difference (mean RT) | NR |
| Robinson S2 | 2015 | NR | NR | NR | NR | NR | NR |
| Rzetelny | 2008 | NR | NR | 3 (replaced by mean) | Yes | Raw RT (mean) | NR |
| Sofuoglu | 2008 | NR | NR | NR | NR | Raw RT (mean) | NR |
| Waters | 2014 | 100 | NR | NR | Yes | Difference (median RT) | >25% errors / 4 SD interference score |
| Waters | 2009 | 100 | NR | NR | Yes | Difference (mean RT) | >33% errors |
| Waters | 2009 | 100 | NR | NR | Yes | Difference (mean RT) | >33% errors |
| **Variations** |  | **4** | **4** | **7** | **2** | **4** | **6** |

*Legend: NR = not reported*

**Table 5: Extraction of data for design decisions of the Drug Stroop**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Lead Author** | **Year** | **Admin** | **Response** | **N Stimuli** | **Times presented** | **Word/Picture** | **Block / Mixed** | **Control stimuli** | **Colours** | **Practice** | **Timeout (ms)** |
| Anastasio | 2014 | Comp | Key | 10 | 6 | Words | Block | NR | 3 | Yes | NR |
| Asmaro | 2014 | Comp | Key | NR | NR | Pictures | Mixed | Mixed | 4 | Yes | NR |
| Carpenter | 2006 | Comp | Key | 20 | 1 | Words | Mixed | NR | 4 | Yes | 6000 |
| Carpenter | 2012 | Comp | Key | 20 | NR | Words | Mixed | NR | NR | Yes | 6000 |
| Copersino | 2004 | Paper | Voice | 5 | 10 | Words | Block | Mixed | NR | Yes | NA |
| Cousijn | 2013 | Paper | Voice | 14 | 4 | Words | Block | Office | 4 | NR | NA |
| Cousijn | 2015 | Paper | Voice | 14 | 4 | Words | Block | Office | 4 | NR | NA |
| DeVito | 2018 | Comp | Key | NR | NR | Words | NR | Furniture | NR | Yes | NR |
| Ersche | 2010 | Comp | Key | 16 | NR | Words | Block | NR | 4 | Yes | 1900 |
| Fadardi | 2010 | Comp | Key | 7 | 8 | Words | Mixed | Building | 4 | Yes | 3000 |
| Franken | 2000 | Comp | Voice | 10 | 10 | Words | Mixed | Transport | 4 | NR | 3000 |
| Franken | 2004 | Comp | Key | 10 | 4 | Words | NR | Transport | 4 | Yes | 3000 |
| Gardini | 2009 | Comp | Key | 319 | NR | Words | Block | Mixed | 4 | Yes | 3000 |
| Haifeng | 2015 | Comp | Key | 2 | 32 | Words | Mixed | Mixed | 4 | Yes | 3000 |
| Hester | 2010 | Comp | Key | 6 | 3 | Words | Mixed | Mixed | 4 | Yes | NR |
| Hester | 2006 | Comp | Key | 20 | 1 | Both | Block | Mixed | 4 | NR | UR |
| Liu | 2012 | Comp | Key | 10 | 6 | Words | Block | Furniture | 3 | Yes | 3000 |
| Liu | 2013 | Comp | Key | 10 | 6 | Words | Block | Mixed | 3 | Yes | 3000 |
| Marhe | 2013 | Comp | Key | 10 | 6 | Words | Block | Furniture | 4 | Yes | 1750 |
| Marissen | 2006 | Comp | Voice | 10 | 5 | Words | Mixed | Transport | 4 | NR | 3000 |
| Nuijten | 2016 | Comp | Key | NR | NR | Words | Block | Transport | 4 | NR | NR |
| Smith | 2014 | Comp | Key | NR | NR | Words | Block | Music | 4 | Yes | 1900 |
| Waters | 2012 | Comp | Key | 11 | 3 | Words | Mixed | Transport | 3 | Yes | 3000 |
| Ziaee | 2016 | Comp | Key | 7 | 4 | Words | Mixed | Building | 4 | Yes | 3000 |
| **Variations** |  | 2 | 2 | 11 | 8 | 2 | 2 | 8 | 2 | 2 | 5 |

*Legend: Comp = computerised; Paper = paper and pencil; NR = not reported; NA = not applicable; UR = until response; Cousijn et al (2015) also contained alcohol Stroop.*

**Table 6: Extracted data of the analysis decisions on the Drug Stroop task**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Lead Author** | **Year** | **Lower bound RT cut off (ms)** | **Upper bound RT cut off (ms)** | **SD removal** | **Exclude errors** | **Primary outcome** | **Participants removed** |
| Anastasio | 2014 | 200 | NR | NR | Yes | Raw RT (mean) | NR |
| Asmaro | 2014 | 150 | 1500 | NR | Yes | Raw RT (mean) | Excessive errors |
| Carpenter | 2006 | NR | NR | NR | NR | Raw RT (mean) | NR |
| Carpenter | 2012 | NR | NR | NR | NR | Difference (mean RT) | NR |
| Copersino | 2004 | NA | NA | NA | Yes\* | Raw RT (mean) | NR |
| Cousijn | 2013 | NA | NA | NA | NR | Raw RT (mean) | NR |
| Cousijn | 2015 | NA | NA | NA | NR | Difference (mean RT) | NR |
| DeVito | 2018 | 100 | NR | NR | Yes | Raw RT (mean) | NR |
| Ersche | 2010 | NR | NR | NR | Yes | Difference (median RT) | NR |
| Fadardi | 2010 | NR | NR | NR | NR | Difference (mean) | NR |
| Franken | 2000 | 200 | 3000 | NR | Yes | Raw RT (mean) | NR |
| Franken | 2004 | NR | NR | 3 above, below | NR | Raw RT (mean) | NR |
| Gardini | 2009 | NR | NR | NR | NR | Difference (mean RT) | NR |
| Haifeng | 2015 | 200 | NR | NR | Yes | Raw RT (mean) | NR |
| Hester | 2010 | NR | NR | NR | Yes | Raw RT (mean) | NR |
| Hester | 2006 | NR | NR | NR | Yes | Raw RT (mean) | >3SDs from mean |
| Liu | 2012 | 200 | NR | NR | Yes | Raw RT (mean) - first block | NR |
| Liu | 2013 | NR | NR | NR | NR | Raw RT (mean) | NR |
| Marhe | 2013 | NR | NR | NR | Yes | Raw RT (mean) | NR |
| Marissen | 2006 | 200 | NR | NR | NR | Raw RT (mean) | NR |
| Nuijten | 2016 | NR | NR | NR | NR | Difference (mean RT) | >25% reaction times (<200 ms) |
| Smith | 2014 | NR | NR | NR | Yes | Difference (median RT) | NR |
| Waters | 2012 | 100 | NR | NR | Yes | Difference (mean RT) | NR |
| Ziaee | 2016 | NR | NR | NR | NR | Difference (mean RT) | negative inference scores were removed |
| **Variations** |  | 4 | 3 | 2 | 2 | 4 | 4 |

*Legend: NR = not reported*

*\*subjects were told to self correct errors*

1. This may have been due to a mislabelling of standard deviations in the original paper, however we were unable to contact the author for clarification. [↑](#footnote-ref-1)
2. Here we looked at variability only in number of alcohol stimuli, number of times presented, control category, number of colours and practice trials. [↑](#footnote-ref-2)
3. If it wasn’t specifically stated otherwise we assumed difference scores were based on mean RTs [↑](#footnote-ref-3)
4. If it wasn’t specifically stated otherwise we assumed difference scores were based on mean RTs [↑](#footnote-ref-4)
5. Not including 319 in estimation which would significantly skew the mean (26.6, if included) [↑](#footnote-ref-5)