Careless responding in crowdsourced alcohol research: a systematic review and meta-analysis of practices and prevalence

Andrew Jones[1,2*]

Jessica Earnest[3]

Martyna Adam[1]

Ross Clarke[3]

Jack Yates[1]

Charlotte R. Pennington[3,4]

[1] Psychology, The University of Liverpool, UK.

[2] Liverpool Centre for Alcohol Research.

[3] School of Psychology, Aston University, Birmingham, UK.

[4] Institute of Health & Neurodevelopment, Aston University, Birmingham, UK.

***Corresponding Author**: Andrew Jones,email: ajj@liv.ac.uk

Funding: None.

Conflicts of Interest: None.

Author note: This manuscript was posted as a pre-print to PsyArXiv (https://psyarxiv.com/rs9xh).

Public Significance Statement

Careless responding poses a threat to the otherwise advantageous method of crowdsourcing research participants. This meta-analysis assessed practices and prevalence of careless responding in crowdsourced alcohol research and found that, out of 96 studies, 51 (53%) included at least one measure of careless responding, identifying ~11.7% [95% CI: 7.6% to 16.5%] participants as careless responders. We provide practical recommendations for handling careless responding and highlight the importance of this to ensure researchers report robust, reliable results with minimal bias.

Abstract

Crowdsourcing — the process of using the internet to outsource research participation to 'workers' — has considerable benefits, enabling research to be conducted quickly, efficiently, and responsively, diversifying participant recruitment, and allowing access to hard-to-reach samples. One of the biggest threats to this method of online data collection however is the prevalence of careless responders who can significantly affect data quality. The aims of this preregistered systematic review and meta-analysis were to: i), examine the prevalence of screening for careless responding in crowdsourced alcohol-related studies; ii), examine the pooled prevalence of careless responding; and iii) identify any potential moderators of careless responding across studies. Our review identified 96 eligible studies (~126,130 participants), of which 51 utilised at least one measure of careless responding (53.2%: 95% CI 42.7% to 63.3%; ~75,334 participants). Of these, 48 reported the number of participants identified by careless responding method(s) and the pooled prevalence rate was ~11.7% [95% CI: 7.6% to 16.5%]. Studies using the MTurk platform identified more careless responders compared to other platforms, and the number of careless response items was positively associated with prevalence rates. The most common measure of careless responding was an attention check question, followed by implausible response times. We suggest that researchers plan for such attrition when crowdsourcing participants and provide practical recommendations for handling and reporting careless responding in alcohol research.

Key words: alcohol use research; crowdsourcing; careless responding; insufficient effort responding; meta-analysis

Introduction

The prevalence of research being conducted online has rapidly increased over the previous decade, and online studies are quickly becoming the standard in many areas of psychology (Gosling & Mason, 2015; Stewart, Chandler, & Paolacci, 2017). This is particularly evident in the field of alcohol-use research as the complete anonymity afforded by online surveys may increase the likelihood of individuals responding without fear of stigma whilst also reducing socially desirable responses, leading to more reliable self-reports of drinking behaviours (Aust, Diedenhofen, Ullrich, & Musch, 2013; Strickland & Stoops, 2019). Online research also allows for considerable flexibility in research design, which has led to large cross-sectional surveys (Reynolds et al., 2019), sophisticated prospective or extensive longitudinal designs (Strickland & Stoops, 2018), randomised controlled trials for alcohol use interventions (Cunningham, Godinho, & Kushnir, 2017), and even qualitative research (Strickland & Victor, 2020). Finally, the COVID-19 pandemic has accelerated online data collection due to concerns around face-to-face recruitment and testing (De Man, Campbell, Tabana, & Wouters, 2021).

In line with the transition to data collection online, the use of novel methods for recruiting participants into studies has proliferated. One such method is ‘crowdsourcing’, which generally refers to the process of using the internet to outsource work to the crowd (Strickland & Stoops, 2019; Wazny, 2017) and usually involves paying ‘workers’ (in this case participants) to take part in research studies via a number of different websites or worker pools. Perhaps the most popular of these is Amazon's Mechanical Turk ('MTurk'). MTurk allows ‘requestors’ to post tasks that can be accomplished via a computer, such as problem-solving or content analyses, for a small financial gain. ‘Workers’ then browse the site and

choose whether to undertake the tasks. The use of MTurk for academic research across a range of disciplines has become increasingly favourable and is now the primary use of the platform (Silvana & Kim, 2019). Following the success of MTurk, numerous other crowdsourcing websites have been developed or identified for use by researchers, such as Prolific (www.prolific.co), CrowdFlower (www.crowdflower.com), and Qualtrics Panels (www.qualtrics.com) to name a few (Palan & Schitter, 2018; Paolacci & Chandler, 2014).

This increase in popularity is likely due to the considerable benefits of crowdsourcing participants, particularly in overcoming the limitations of traditional laboratory-based research. First, crowdsourcing facilitates recruitment of large samples in a relatively short period of time; meaning research can be conducted quickly, efficiently, and responsively in 'real-time' (Wazny, 2017). Indeed, crowdsourcing is able to greatly reduce the 'cost per observation' of studies (Gupta, Rigotti, & Wilson, 2021), as well as expanding opportunities to recruit diverse and representative samples (Henrich, Heine, & Norenzayan, 2010, Ghai, 2021), and traditionally hard to reach populations (Mullen, Fox, Goshorn, & Warraich, 2021). For example, MTurk has more than 500,000 workers from over 190 countries and can provide nationally representative samples (at least for the USA: (Burnham, Le, & Piedmont, 2018)). Furthermore, the demographics of this participant pool has remained consistent over time (Moss, Rosenzweig, Robinson, & Litman, 2020). Similarly, as of November 2021, Prolific currently has >250,000 unique users and allows for generalisable UK sampling, based on over 100 demographic screeners (Palan & Schitter, 2018). Given these benefits, many researchers agree that crowdsourcing has the potential to greatly improve global health research (Morris et al., 2017; Ranard et al., 2014; Wazny, 2017).

There are also considerable limitations which come with outsourcing participants to the internet. Leaving the laboratory surrenders the high experimental control that researchers have over their environment. For example, in a study by Clifford and Jerit (2014) participants who completed surveys in their natural environment rather than the laboratory reported greater use of their mobile phone (21% vs. 9%, respectively), talking to another person (21% vs. 2%), and browsing the internet (11% vs. 1%). Therefore, one of the biggest limitations is the potential for these distractions to affect data quality. This may be particularly prominent when participants are not intrinsically motivated to complete the research, but gain incentives for doing so (Brühlmann, Petralito, Aeschbach, & Opwis, 2020; Chmielewski & Kucker, 2020). Careless responding (sometimes known as insufficient effort responding), which has been defined as '*any behaviour, regardless of intention that results in the reporting of data that does not accurately reflect the true nature of the participant'* (Nichols & Edlund, 2020, p. 626), is thought to be prevalent, and has been measured via different methods throughout the psychology literature. For example, implausible completion times might occur when participants aim to gain payment or compensation for the minimal amount of time or effort involved, and therefore provide fast but unreliable/unthoughtful responses (Dominik Johannes, 2019). Non-mutually exclusive might be a lack of attentive responses or failure to read and follow instructions, which can be identified by 'attention checks' (Göritz, Borchert, & Hirth, 2019). Importantly for the researcher, poor quality data can increase noise (non-random error variance), making it much more difficult to ascertain any signal within the noise (Schroeders, Schmidt, & Gnambs, 2021). Furthermore, if left unaccounted for, careless responders '*pose a great threat to replicability*' (Curran, 2016, p. 5).

Given the rapid shifts to online research it is important to draw attention to these issues, but also to overcome them, to ensure the evidence base is not biased by poor quality

studies with unreliable data. Some researchers have suggested that the quality of data from crowdsourcing platforms is worsening. For example, a longitudinal design from 2015 to 2019 saw a dramatic increase in statistically improbable responses (e.g. reporting > 4 children of a single age), response inconsistencies from previous waves, and failed pre-screen questions via MTurk (Chmielewski & Kucker, 2019). It is unlikely these issues are limited to one platform (Boas, Christenson, & Glick, 2020), however the magnitude of noise may be greater in MTurk (Gupta et al., 2021).

Some studies suggest careless responding should be expected in 10-15 % of samples, but the overall prevalence may be study specific (Goldammer, Annen, Stöckli, & Jonas, 2020). Such estimates are important to quantify given that even lower prevalence of careless responding (>5%) have been demonstrated to bias study results in simulations (Credé, 2010), or if detected and removed can substantially reduce statistical power (Conijn, Franz, Emons, de Beurs, & Carlier, 2019), or even completely change data interpretations (Arias, Garrido, Jenaro, Martínez-Molina, & Arias, 2020; Maniaci & Rogge, 2014). Importantly, in alcohol and addiction research careless responses can lead to the misattribution of individuals on clinical or diagnostic measures (Meyer, Faust, Faust, Baker, & Cook, 2013), or inflate correlations between alcohol use and other variables of interest (King, Kim, & McCabe, 2018).

Given the growing importance of crowdsourced research into alcohol use the aims of this systematic review and meta-analysis were as follows: i) examine the prevalence of screening for careless responding in crowdsourced alcohol-related studies, ii) examine the pooled prevalence of participants identified as careless responders in these studies, iii) identify any potential moderators of careless responding across studies, and iv) provide

practical recommendations for screening and handling for carelessness for researchers moving forward.

Methods

Transparency & Openness

The design and analysis plan for this review was pre-registered via the standard Open Science Framework protocol on 18th May 2021 (https://osf.io/p837n/registrations) and the full data was extracted between July 2021 and August 2021. Data, materials, and analysis code are publicly available (https://osf.io/p837n/). A lack of data availability limited our ability to conduct individual participant meta-analysis, which was pre-registered as a possibility.

Search Strategy

We searched three comprehensive and widely used academic databases (PsycInfo, Scopus, Pubmed) as well as the pre-print server PsyArXiv from 2011 to 2021. We limited searches from 2011 as Chandler and Shapiro (2016) demonstrated this was the start of the proliferation of journal articles using MTurk, and a 10-year period would provide us with a representative sample of studies. Search terms included [(crowdsour* OR online) AND alcohol*]. Once we had identified reoccurring crowdsourcing platforms that fit our definition (see below), we also conducted simpler searches for the crowdsourcing platform name + alcohol [e.g. (CrowdFlower + alcohol*)] in the first 100 hits in Google Scholar to identify any further studies. Identification of relevant articles and data extraction was conducted in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Statement (PRISMA; Page et al., 2021; see supplementary online Table 1 for PRISMA checklist).

Eligibility criteria

To be eligible for inclusion in the review and meta-analysis, studies were required to (i) collect data about an individual's alcohol use, and (ii) be published in English. To determine crowdsourcing platforms, we used the criteria that a wide variety of participants could opt into the platform rather than being selected purposively (Stewart et al., 2017), and were paid or rewarded for their participation. As such, we excluded studies which recruited student samples online through course-credit recruitment systems (such as SONA systems) and targeted market research (e.g. Ipsos Mori). We reasoned that these are not crowdsourcing platforms because they sample a narrow heterogeneous sample and don't provide financial gain or are invite-only. This decision was made post-hoc. Studies were also excluded if they were longitudinal follow-up studies (e.g. if brought about by COVID-19 changes to recruitment), were a systematic review or meta-analysis, or if data was duplicated across multiple papers. In the latter case, we included the paper which provided the most information about careless responding.

Extraction and coding

Our main outcome variables were: i) the reporting of measurement of careless responding [yes, no] and ii) if careless responding was reported, the percentage of participants that were identified as careless responders (number of careless responders / number of sample recruited). For our moderator variables we extracted the number of unique careless responding measures used (e.g Attention checks + implausible completion times = 2), number of total carelessness items included (e.g. if a study had 4 attention checks and

implausible completion time = 5), crowdsourcing platform, sample location, type of careless responding, length of time of the study (in minutes), the year that data was collected and the year the study was published. To be eligible for moderation analysis a subgroup needed at least 5 effect sizes, as defined in our pre-registration.

Statistical analyses

For our first aim (identifying the measurement of careless responding) we conducted a frequency count. To examine the percentage of participants identified as careless responders we conducted a prevalence meta-analysis. We used a Restricted Maximum Likelihood random effects model in anticipation of considerable heterogeneity. In a deviation from our pre-registered analysis we used Freeman-Tukey arcsine transformation (Miller, 1978), rather than square root transformation, using the 'escalc' function in 'metafor'. This was appropriate as some studies had a proportion of 0, and the Freeman-Tukey transformation does not require corrections and improves variance stability in these cases (Lin & Xu, 2020). We report back transformed values in text and Figures.

To examine the robustness of any pooled prevalence estimate we conducted a number of additional analyses, which included (i) Trim and Fill; (ii) removal of outlying effects; (iii) computation of 'influence' statistics; and (iv) Graphical Displays of Study Heterogeneity (GOSH). Trim and Fill analysis trims any studies which are thought to contribute to funnel plot asymmetry (usually small studies with large effects) before imputing effects to improve the symmetry of the funnel plot (Duval & Tweedie, 2000). To remove outlying effect sizes, we conducted a box plot on the included effect sizes (see supplementary Figure 1); using the 'influence' command in R we examined if any studies were having an outlying influence on the effect size. Finally, GOSH conducts meta-analyses on all possible subsets of effect sizes

included in the meta-analysis which allows for robust estimates of heterogeneity (Olkin, Dahabreh, & Trikalinos, 2012). We examined the effects of these additional analyses on both the pooled effect, but also between-study heterogeneity. Heterogeneity was measured using the I^2 (Inconsistency) statistic. We used $I^2 > 50\%$ as indicative of moderate heterogeneity and $I^2 > 75\%$ as indicative of substantial heterogeneity (Higgins, Thompson, Deeks, & Altman, 2003).

We conducted moderation analyses on i) the number of unique measures of careless responding used (1 vs >1); the absolute number of careless response items; the crowdsourcing platform (MTurk vs Other); and the length of time of the study (in minutes, as reported). We also conducted exploratory analyses on the year in which the data was collected and the year in which the study was published (of which information was available in 28 articles). We examined the prevalence of individuals failing eligibility criteria in studies, and whether this predicted carelessness responding rates. Within studies using the MTurk platform we examined if studies which explicitly reported pre-screening participants using approval rates and previous task completion (e.g. >95% approval rates) influenced crowdsourcing rates (compared to no reporting). We could not perform moderation on type of careless measure due to heterogeneity and small subgroup sizes.

Results

Articles identified

The initial searches included 5,942 articles, which decreased to 3,293 after removal of duplicates. Titles and abstracts were screened and cross checked by all authors, with high levels of agreement (>95%). At this stage, 2,827 studies were removed. The full text of all

remaining articles was screened against inclusion and exclusion criteria and cross checked by a second coder. Following this, 96 articles remained. See Figure 1 for PRISMA flowchart.

[insert Figure 1 here]

Study characteristics

The majority of studies recruited samples from the USA (N = 52: 54.1%) and primarily using the MTurk crowdsourcing platform (N = 52: 54.1%). The average age of the samples was ~34.2 and the gender distribution was ~51.1% female. Most studies sampled majority White / Caucasian ethnicity. Across all studies identified (including those that did not measure careless responding but did use crowdsourcing) the sample size was ~126,130 (min = 78, max = 7058).

Frequency of studies which measured careless responding

Out of 96 identified studies, 51 studies (53.1% [95% CI: 42.7% to 63.3%]: ~ 75,334 participants) utilised at least one measure of careless responding (see supplementary online Table 1 for full list of included studies). The most common measure of careless responding was an attention check question (see Table 1). The majority of studies (27 / 52.9%) included one measure of careless responding only.

Pooled prevalence of careless responders

From the 51 studies that utilised measures of careless responding, 48 reported the number of participants which were identified by the method(s). Of these studies, the pooled

prevalence rate of careless responding was 11.7% ([95% CI: 7.6% to 16.5%], I^2 = 99.7%, see Figure 2).

[insert Figure 2 here]

Trim and Fill analyses demonstrated that 14 effect sizes were missing on the right side of the funnel plot (see online supplementary Figure 2), and inclusion of these 'filled' effect sizes increased the pooled prevalence of careless responding to 18.6% [95% CI: 13.5% to 24.3%]. Removal of three outliers identified by a box plot reduced this prevalence rate to 8.9% ([95% CI: 6.3% to 11.9%], I^2 = 99.4%: see online supplementary Figure 3). Outliers overlapped with influential cases (N = 2: Rodriguez et al, 2020;Huhn et al, 2021). Finally, GOSH analyses on 50,000 subsets identified the average heterogeneity to be high under every iteration (>96%: see online supplementary Figure 4). Given the influence of outlying prevalence rates for three studies we removed these studies from all subsequent moderation analyses but note any changes to the patterns of results with these included.

Moderation Analyses

Number of unique careless measures used

We coded the prevalence of careless responding within studies, where the study used only one measure of careless responding (N = 23) vs. more than one unique measure (N = 22)[1]. There was no statistical evidence of moderation ($X^2(1) = 0.07$, $p = .798$). Studies which used only one measure of careless responding had a pooled prevalence rate of 8.5% ([95%

[1] *Note, these numbers are different to those reported above due to the exclusion of outliers and missing data on prevalence.*

CI: 4.9% to 13.1%], I^2 = 99.4%), and studies which used multiple measures had a pooled prevalence rate of 9.3% ([95% CI: 5.9% to 13.4%], I^2 = 99.3%). Treating the number of different careless measures used as continuous variable in meta-regression was not significant (*b* = .02 [95% CI: -.07 to .12]).

Total number of carelessness items

We examined the total number of carelessness items used in the study using a meta-regression (M = 2.9, Max = 8, based on 32 studies with available information). The association was significant (*b* = .04 [95% CI: .01 to .07], *p* = .013), indicating that the number of carelessness items was positively associated with identification of careless responders (see Figure 3).

[insert Figure 3 here]

Crowdsource platform

The majority of studies were conducted using MTurk (N = 31), followed by Prolific (N = 3). Given the smaller number of other crowdsourcing platforms used we compared MTurk to others (N = 14, total). Moderation analysis was significant (X^2(1) = 4.11, *p* = .042). Studies using the MTurk platform identified more careless responders (10.9% [95% CI: 7.4% to 14.9%], I^2 = 99.2%) compared to other platforms (5.2% [95% CI: 2.5% to 8.8%], I^2 = 99.3%). If outliers were included in this analysis, the effect of moderation was not significant (X^2(1) = 1.17, *p* = .278).

Length of online study

We were able to extract the average length of the online study in minutes from 16 studies. Across the studies the average length was 18 minutes [min = 5 minutes, max = 32 minutes]. A meta-regression demonstrated that the length of time did not significantly influence careless responding (b = 0.00 [95% CI: -0.01 to 0.01], z = 0.03, p = .976).

Year

A meta-regression examining the association between year of data collection and prevalence of careless responding was not significant (b < -.01 [95% CI: -.02 to .01], p = .621). When examining the year of data publication, this remained non-significant (b < -.001 [95% CI: -.02 to .02], p = .977).

Pre-screening and eligibility as a method of reducing carelessness

Within studies that screened for carelessness, we identified 17 studies that also reported removing participants who did not meet eligibility criteria and the pooled prevalence rate was approximately 38.3% [95% CI: 25.8% to 50.9%] of participants removed. Interestingly, increased percentage of participants removed for failing eligibility checks was positively associated with increased prevalence of careless responding within studies (b = .50 [95% CI: .07 to .93], p = .024).

Within studies conducted on MTurk we identified 12 studies which did not explicitly state whether participants were pre-screened, and 19 where information was provided. Moderation analyses on pre-screening was not significant ($X^2(1) = 0.26$, p = .608).

Discussion

In this systematic review and meta-analysis we observed that approximately 53% crowdsourced studies in the field of alcohol-research accounted for careless responding. Within those studies which utilised some measures of careless responding, approximately 12% of participants were identified as providing poor quality data, however there was considerable heterogeneity in these prevalence rates. The methods by which researchers identified careless responding were also highly variable.

Careless responding has been identified as a widespread issue in survey studies, particularly when participants have been recruited through crowdsourcing platforms (Brühlmann et al., 2020). Our data suggests a small majority of studies reported on attempts to identify careless responders, but this reporting was often suboptimal (which has also been observed elsewhere; Arndt, Ford, Babin, & Luong, 2021). This is particularly surprising given research has identified the impact of careless responding on inferences made (Credé, 2010; King et al., 2018), as well as statistical power across a wide range of study types (Conijn et al., 2019). Furthermore, careless responding in alcohol and addiction research has been found to result in misattribution of clinical or diagnostic indices (Meyer, Faust, Faust, Baker, & Cook, 2013), and unreliable, inflated correlations between alcohol use and other variables (King, Kim, & McCabe, 2018). To our knowledge this is the first attempt at characterising the extent (and practices for identifying) careless responding within crowdsourced data in alcohol research.

Our pooled prevalence estimates of ~12% of individuals who could be categorised as careless responders is similar to estimates across different samples and fields of study (between 10 – 15%: Goldammer et al., 2020; Schroeders, Schmidt, & Gnambs, 2021). These comparisons suggest, at least for the studies we were able to identify, that careless responding

is no more prevalent in alcohol-related research than other fields. In all cases, these participants were removed from subsequent study analyses, which can impact any a-priori power calculations. Whilst some online studies report oversampling to account for potential data removal (e.g. Robinson, Smith, & Jones, 2021), this isn't common practice. Our data therefore suggests that, at least as a rule of thumb, studies in alcohol and addiction research should be prepared to oversample by ~12% to ensure that statistical power is retained if excluding careless responders. Moreover, we found no studies that attempted to examine the impact of removal or inclusion of careless responders on study inferences.

Perhaps contrary to expectations we observed that increased proportion of participants that are removed from a study for failing eligibility criteria was associated with increased prevalence of careless responding, as we might expect removal of participants who failed eligibility criteria to capture individuals who respond carelessly. Alternatively, these studies may be better at capturing carelessness in general. However, this exploratory analysis should be interpreted with caution because it was conducted on a small sub-set of data and needs clarifying in future studies. Similarly, we demonstrated that studies which explicitly stated pre-screening using MTurk criteria (e.g. >95% approval ratings and 100 previous task completed) did not significantly differ in carelessness compared to studies which did not explicitly state the use of pre-screening. In both cases above, it is impossible to distinguish whether pre-screening and eligibility checks took place but were simply not reported (and therefore studies were coded as not including these steps). Future research may benefit from clear reporting guidelines for crowdsourced studies (see Ramírez et al., 2021).

We were able to demonstrate some evidence that careless responding was influenced by our chosen moderator variables. First, MTurk samples tended to have more careless

responders than other samples, which supports other research to suggest increasing rates of poor-quality data from MTurk (Arndt et al., 2021; Chmielewski & Kucker, 2019; Gupta et al., 2021). However, this finding was not robust to outlier inclusion. We also demonstrated that the total number of carelessness items led to increased identification of carelessness (but not the number of unique careless response measures). This is in line with research suggesting that one careless response measure is inadequate (Berinsky, Margolis, & Sances, 2014). The most common type of method for identifying careless responding was an attention check (e.g. asking participants to select a specific response such as 'strongly agree' or providing a statement or question with a clear meaningful answer '*which planet do you live on?*'), followed by implausible completion times, but we were unable to reliably test for differences across types of measure. Finally, we found no evidence that typical survey length was associated with careless responding. However, most studies did not report the time taken to complete the survey, and further did not specify at which point in the questionnaire measures of careless responding were included. This may have important consequences given research which has demonstrated that longer surveys are more prone to careless responding (Bowling, Gibson, Houpt, & Brower, 2020), and participants respond more carelessly as surveys progress (Bowling et al., 2020).

It is possible that researchers intentionally choose not to include carelessness checks into their research designs, as some argue that such inclusions are a threat to external validity (Berinsky, Margolis, & Sances, 2016; Kung, Kwok, & Brown, 2018). For example, respondents who fail attention checks are unlikely to represent a random subset of the sampled population (Berinsky et al., 2014), and may have a common underlying personality trait which contributes to their failure. Indeed, Ward et al., (2017) demonstrated that individuals lower in conscientiousness and agreeableness provide more careless responses

which, in turn, biases the average conscientiousness and agreeableness of the sample if removed (see also, Bowling et al., 2020). Both factors are demonstrably associated with quantity and frequency of alcohol consumption (Hakulinen et al., 2015). Similarly, the inclusion of measures to identify careless responding may also increase subsequent social desirability. Clifford and Jerit (2015) demonstrated that explicit warnings about careless responding and feedback led to increases in socially desirable responses, particularly amongst educated participants. Again, social desirability has been shown to bias estimates of alcohol consumption (Davis, Thake, & Vilhena, 2010). Finally, there is some concern that poorly thought-out careless response checks might incorrectly categorise individuals. For example, the careless response item '*I can eat as much as a horse*' has a high false positive rate of capturing careless responding (up to 42%), as participants are able to justify why they might agree with this item (Curran & Hauser, 2019). These authors recommend careless response items which focus on clear impossibility/truth (e.g. 'Oranges are Fruit' or 'I work fourteen months in a year').

It may be possible to overcome issues with explicit carelessness items with the development and implementation of *covert* measures. Various covert measures exist, such as identifying weak correlations between positively and negatively worded questions, long string index (e.g. consecutive identical responses on a Likert scale), Mahalanobis distance (e.g. distance between data points and a distribution), and individual consistency (e.g. the consistency of response strings within an individual: (Huang, Curran, Keeney, Poposki, & DeShon, 2012)). It is possible that these are less common due to the increased effort and expertise required to analyse them, however we note that various step-by-step guides, syntax, and software packages exist to help researchers implement these (Curran, 2016; Huang et al., 2012; Landers, 2016; Meade & Craig, 2012). Furthermore, perhaps one of the most intuitive

detection methods (long string index) is simple to calculate and requires researchers to identify the longest unbroken sequence of the response. For example, in the case of '3, 4, 4, 4, 4, 4, 1, 3, 3, 2, 3', this is clearly '4' which is repeated 5 times. In this case the max long string score would be 5. Rules of thumb suggest that individuals with a max long string score of greater than half the number of items on the scale should be considered as careless responders (Curran, 2016).

There are some limitations to this meta-analysis. First, we were unable to resolve between-study heterogeneity in our moderator analyses, meaning that these findings should be interpreted with caution (Imrey, 2020; Sun & Feng, 2019), but also future research should attempt to elucidate both individual and study-level variables that might predict careless responding or poor-quality data, specifically in alcohol and addiction research. Second, as we limited our data analysis to alcohol-related research we effectively analysed a narrow sample of studies and individuals who likely drank alcohol. Therefore, we are unable to make robust generalisations to, or comparisons with, other research fields. Finally, we were unable to directly address the issue of inauthentic users ('bots') which have been thought to proliferate crowdsourcing platforms and were first identified as an issue in research studies circa 2018 (Rea, Kleeman, Zhu, Gilbert, & Yue, 2020). Many platforms block repeat internet protocol (IP) addresses in attempts to circumvent this issue; however, this can be overcome with Virtual Private Networks. Other safeguards include public Turing tests (e.g. CAPTCHA), individual study links, or open-ended questions (e.g. Chmeilewski & Kucker, 2020). It is possible that the prevalence of careless responders we identified may have included some inauthentic users/bots, and researchers should certainly be mindful of this issue when designing studies. Indeed, Godinho et al. (2020) identified no addiction researchers whom have considered the issues of bots in their designs.

Moving forward we emphasise the importance of examining careless responding in crowdsourced alcohol-related data and ensuring detailed reporting. One such approach would be to include reporting requirements which could be integrated into pre-existing checklists for internet research, such as The Checklist for Reporting Results of Internet E-Surveys (CHERRIES: (Eysenbach, 2004)), as well as specific guidelines for crowdsourced samples (Stewart et al., 2017). Indeed, Chmielewski and Kucker (2019) suggest all studies using MTurk should adopt specific reporting criteria such as validity indicators, screening decisions and the number of participants dropped, which could be broadened to other platforms. Where there is good reason not to include these (e.g., threats to validity; Kung et al., 2018), such decisions should be justified explicitly. Individual studies should also i) compare the sample of individuals who might respond carelessly to the non-careless sample on key demographic information, and variables critical to hypothesis testing; and ii) report any inferential statistics in both the full sample and non-careless sample to examine the extent to which their removal might influence findings (Waites & Ponder, 2016). Researchers should consider which careless responding measures are most appropriate given their study (Huang et al., 2012), but also the individual scales used (McCredie, Harris, Regan, Morey, & Fields, 2021). Researchers should also transparently detail the measures used for careless responding, avoiding terms such as ‘data quality checks’ or ‘catch questions’ to aid reproducibility. Finally, it is important to consider at what point in a study to measure careless responding; if this is measured early it may not capture later lapses in responding (especially if only one measure is used), as well as influencing social desirability (Clifford & Jerit, 2015).

To conclude, this systematic review and meta-analysis examined reporting of careless responding methods in alcohol-related crowdsourced research, but also the pooled prevalence

of individuals who might be categorised as careless responders in these studies. Approximately 53% of studies explicitly reported on careless responding, whilst the remaining provided no justifications of not doing so. Within these studies the prevalence of careless responding was approximately 12%, which is consistent with reports in other fields. We present some wider recommendations for researchers handling careless responding and hope this review highlights the importance of assessing this to ensure reporting of robust, reliable results with minimal bias.

References:

Albertella, L., Le Pelley, M. E., Chamberlain, S. R., Westbrook, F., Lee, R., Fontenelle, L. F., Grant, J. E., Segrave, R. A., McTavish, E., & Yücel, M. (2020). Reward-related attentional capture and cognitive inflexibility interact to determine greater severity of compulsivity-related problems. *Journal of Behavior Therapy and Experimental Psychiatry, 69*, 101580. https://doi.org/10.1016/j.jbtep.2020.101580

Altendorf, M. B., van Weert, J., Hoving, C., & Smit, E. S. (2019). Should or could? Testing the use of autonomy-supportive language and the provision of choice in online computer-tailored alcohol reduction communication. *Digital Health, 5, 2055207619832767*. https://doi.org/10.1177/2055207619832767

Arch J. J. (2013). Pregnancy-specific anxiety: which women are highest and what are the alcohol-related risks?. *Comprehensive Psychiatry, 54(3),* 217–228. https://doi.org/10.1016/j.comppsych.2012.07.010

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods, 52*(6), 2489-2505. https://doi.org/10.3758/s13428-020-01401-8

Arndt, A. D., Ford, J. B., Babin, B. J., & Luong, V. (2021). Collecting samples from online services: How to use screeners to improve data quality. *International Journal of Research in Marketing, advanced online publication*. https://doi.org/10.1016/j.ijresmar.2021.05.001

Angus, D. J., Pickering, D., Keen, B., & Blaszczynski, A. (2021). Study framing influences crowdsourced rates of problem gambling and alcohol use disorder. *Psychology of*

Addictive Behaviors, 10.1037/adb0000687. Advance online publication. https://doi.org/10.1037/adb0000687

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods, 45*(2), 527-535. https://doi.org/10.3758/s13428-012-0265-2

Bergman, B. G., Wu, W., Marsch, L. A., Crosier, B. S., DeLise, T. C., & Hassanpour, S. (2020). Associations Between Substance Use and Instagram Participation to Inform Social Network-Based Screening Models: Multimodal Cross-Sectional Study. *Journal of Medical Internet research, 22(9),* e21916. https://doi.org/10.2196/21916

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science, 58*(3), 739-753. https://doi.org/10.1111/ajps.12081

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers? *Journal of Experimental Social Psychology, 66,* 20-28. https://doi.org/10.1016/j.jesp.2015.09.010

Blackwell, A., De-Loyde, K., Hollands, G. J., Morris, R. W., Brocklebank, L. A., Maynard, O. M., Fletcher, P. C., Marteau, T. M., & Munafò, M. R. (2020). The impact on selection of non-alcoholic vs alcoholic drink availability: an online experiment. *BMC Public Health, 20(1),* 526. https://doi.org/10.1186/s12889-020-08633-5

Boas, T. C., Christenson, D. P., & Glick, D. M. (2020). Recruiting large online samples in the United States and India: Facebook, Mechanical Turk, and Qualtrics. *Political Science Research & Methods, 8*(2), 232-250. https://doi.org/10.1017/psrm.2018.28

Booth, L., Jongenelis, M. I., Drane, C., Miller, P. G., Chikritzhs, T., Hasking, P., Hastings, G., Thorn, M., & Pettigrew, S. (2021). Attitudinal factors associated with drink

counting. *Drug and Alcohol Review, 40(6),* 1056–1060. https://doi.org/10.1111/dar.13277

Booth, L., Norman, R., & Pettigrew, S. (2020). The potential effects of autonomous vehicles on alcohol consumption and drink-driving behaviours. *Drug and Alcohol Review, 39(5),* 604–607. https://doi.org/10.1111/dar.13055

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods,* 1094428120947794. https://doi.org/10.1177/1094428120947794

Britton, M., Derrick, J. L., Shepherd, J. M., Haddad, S., Garey, L., Viana, A. G., & Zvolensky, M. J. (2021). Associations between alcohol consumption and smoking variables among Latinx daily smokers. *Addictive Behaviors, 113,* 106672. https://doi.org/10.1016/j.addbeh.2020.106672

Brown, K. G., Stautz, K., Hollands, G. J., Winpenny, E. M., & Marteau, T. M. (2016). The Cognitive and Behavioural Impact of Alcohol Promoting and Alcohol Warning Advertisements: An Experimental Study. *Alcohol and Alcoholism (Oxford, Oxfordshire), 51(3),* 354–362. https://doi.org/10.1093/alcalc/agv104

Brühlmann, F., Petralito, S., Aeschbach, L. F., & Opwis, K. (2020). The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology*, *2*, 100022. https://doi.org/10.1016/j.metip.2020.100022

Burnham, M. J., Le, Y. K., & Piedmont, R. L. (2018). Who is Mturk? Personal characteristics and sample consistency of these online workers. *Mental Health, Religion & Culture, 21*(9-10), 934-944. https://doi.org/10.1080/13674676.2018.1486394

Buykx, P., Gilligan, C., Ward, B., Kippen, R., & Chapman, K. (2015). Public support for alcohol policies associated with knowledge of cancer risk. *The International Journal on Drug Policy, 26(4),* 371–379. https://doi.org/10.1016/j.drugpo.2014.08.006

Buykx, P., Li, J., Gavens, L., Hooper, L., Gomes de Matos, E., & Holmes, J. (2018). Self-Reported Knowledge, Correct Knowledge and use of UK Drinking Guidelines Among a Representative Sample of the English Population. *Alcohol and alcoholism (Oxford, Oxfordshire), 53(4),* 453–460. https://doi.org/10.1093/alcalc/agx127

Buykx, P., Li, J., Gavens, L., Hooper, L., Lovatt, M., Gomes de Matos, E., Meier, P., & Holmes, J. (2016). Public awareness of the link between alcohol and cancer in England in 2015: a population-based survey. *BMC Public Health, 16(1),* 1194. https://doi.org/10.1186/s12889-016-3855-6

Campbell, E. M., & Strickland, J. C. (2019). Reliability and validity of the Brief DSM-5 Alcohol Use Disorder Diagnostic Assessment: A systematic replication in a crowdsourced sample. Addictive Behaviors, 92, 194–198. https://doi.org/10.1016/j.addbeh.2019.01.007

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*(1), 53-81. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Chmielewski, M., & Kucker, S. C. (2019). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological & Personality Science, 11*(4), 464-473. https://doi.org/10.1177/1948550619875149

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science, 1*(2), 120-131. https://doi.org/10.1017/xps.2014.5

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly, 79*(3), 790-802. https://doi.org/10.1093/poq/nfv027

Conijn, J. M., Franz, G., Emons, W. H. M., de Beurs, E., & Carlier, I. V. E. (2019). The assessment and impact of careless responding in routine outcome monitoring within mental health care. *Multivariate Behavioral Research, 54*(4), 593-611. https://doi.org/10.1080/00273171.2018.1563520

Coomber, K., Jones, S. C., Martino, F., & Miller, P. G. (2017). Predictors of awareness of standard drink labelling and drinking guidelines to reduce negative health effects among Australian drinkers. Drug and alcohol review, 36(2), 200–209. https://doi.org/10.1111/dar.12383

Coomber, K., Martino, F., Barbour, I. R., Mayshak, R., & Miller, P. G. (2015). Do consumers 'Get the facts'? A survey of alcohol warning label recognition in Australia. *BMC Public Health, 15*, 816. https://doi.org/10.1186/s12889-015-2160-0

Corrigan, P. W., Lara, J. L., Shah, B. B., Mitchell, K. T., Simmes, D., & Jones, K. L. (2017). The Public Stigma of Birth Mothers of Children with Fetal Alcohol Spectrum Disorders. *Alcoholism, Clinical and Experimental Research, 41(6)*, 1166–1173. https://doi.org/10.1111/acer.13381

Craig, D. G., Dakkak, M., Gilmore, I. T., Hawkey, C. J., Rhodes, J. M., Sheron, N., & British Society of Gastroenterology (2012). A drunk and disorderly country: a nationwide cross-sectional survey of alcohol use and misuse in Great Britain. *Frontline Gastroenterology, 3(1)*, 57–63. https://doi.org/10.1136/flgastro-2011-100047

Crane, C. A., Schlauch, R. C., & Miller, K. E. (2019). The association between caffeinated alcoholic beverages and the perpetration of intimate partner violence. *The American*

Journal of Drug and Alcohol Abuse, 45(5), 538–545. https://doi.org/10.1080/00952990.2019.1605522

Crane, C. A., Umehira, N., Berbary, C., & Easton, C. J. (2018). Problematic alcohol use as a risk factor for cyber aggression within romantic relationships. *The American Journal on Addictions, 10.1111/ajad.12736. Advance online publication. https://doi.org/10.1111/ajad.12736*

Credé, M. (2010). Random responding as a threat to the validity of effect size eEstimates in correlational research. *Educational & Psychological Measurement, 70*(4), 596-612. https://doi.org/10.1177/0013164410366686

Critchlow, N., MacKintosh, A. M., Thomas, C., Hooper, L., & Vohra, J. (2019). Awareness of alcohol marketing, ownership of alcohol branded merchandise, and the association with alcohol consumption, higher-risk drinking, and drinking susceptibility in adolescents and young adults: a cross-sectional survey in the UK. *BMJ open, 9(3),* e025297. https://doi.org/10.1136/bmjopen-2018-025297

Cunningham, J. A., Godinho, A., & Kushnir, V. (2017). Using Mechanical Turk to recruit participants for internet intervention research: Experience from recruitment for four trials targeting hazardous alcohol consumption. *BMC Medical Research Methodology, 17*(1), 156. https://doi.org/10.1186/s12874-017-0440-3

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66,* 4-19. https://doi.org/10.1016/j.jesp.2015.07.006

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality, 82,* 103849. :https://doi.org/10.1016/j.jrp.2019.103849

Davis, C. G., Thake, J., & Vilhena, N. (2010). Social desirability biases in self-reported alcohol consumption and harms. *Addictive Behaviors, 35*(4), 302-311. https://doi.org/10.1016/j.addbeh.2009.11.001

de Beukelaar, M. F., Janse, M. L., Sierksma, A., Feskens, E. J., & de Vries, J. H. (2019). How full is your glass? Portion sizes of wine, fortified wine and straight spirits at home in the Netherlands. *Public Health Nutrition, 22(10),* 1727–1734. https://doi.org/10.1017/S1368980019000442

Dekker, M. R., Jones, A., Maulik, P. K., & Pettigrew, S. (2020). Public support for alcohol control initiatives across seven countries. *The International Journal on Drug Policy, 82*, 102807. https://doi.org/10.1016/j.drugpo.2020.102807

Dekker, M. R., Jongenelis, M. I., Hasking, P., Kypri, K., Chikritzhs, T., & Pettigrew, S. (2020). Factors Associated with Engagement in Protective Behavioral Strategies among Adult Drinkers. *Substance Use & Misuse, 55(6),* 878–885. https://doi.org/10.1080/10826084.2019.1708944

De Man, J., Campbell, L., Tabana, H., & Wouters, E. (2021). The pandemic of online research in times of COVID-19. *BMJ Open, 11*(2), e043866. https://doi.org/10.1136/bmjopen-2020-043866

Dominik Johannes, L. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods, 13*(3). https://doi.org/10.18148/srm/2019.v13i3.7403

Dumas, T. M., Davis, J. P., Maxwell-Smith, M. A., & Bell, A. (2018). From Drinking Group Norms to Individual Drinking Consequences: A Moderated Mediation Model Examining the Role of Members' Status, Identification with the Group and with Emerging Adulthood. *Substance Use & Misuse, 53(8),* 1311–1323. https://doi.org/10.1080/10826084.2017.1408651

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Eysenbach, G. (2004). Improving the quality of web surveys: The Checklist for Reporting Results of Internet E-Surveys (CHERRIES). *Journal of Medical Internet Research, 6*(3), e34. https://doi.org/10.2196/jmir.6.3.e34

Fendrich, M., Becker, J., Park, C., Russell, B., Finkelstein-Fox, L., & Hutchison, M. (2021). Associations of alcohol, marijuana, and polysubstance use with non-adherence to COVID-19 public health guidelines in a US sample. *Substance Abuse, 42(2),* 220–226. https://doi.org/10.1080/08897077.2021.1891603

Forkus, S. R., Breines, J. G., & Weiss, N. H. (2020). PTSD and alcohol misuse: Examining the mediating role of fear of self-compassion among military veterans. Psychological trauma : theory, research, practice and policy, 12(4), 364–372. https://doi.org/10.1037/tra0000481

French, M.T., Mortensen, K. & Timming, A.R (2020). Changes in self-reported health, alcohol consumption and sleep quality during the COVID-19 pandemic in the United States. *Applied Economics Letters,* DOI: 10.1080/13504851.2020.1861197

Fugitt, J. L., Ham, L. S., & Bridges, A. J. (2017). Undifferentiated Gender Role Orientation, Drinking Motives, and Increased Alcohol Use in Men and Women. *Substance Use & Misuse, 52(6),* 760–772. https://doi.org/10.1080/10826084.2016.1264963

Ghai, S. (2021). It's time to reimagine sample diversity and retire the WEIRD dichotomy. *Nature Human Behaviour, 5,* 971-972

Gratz, K. L., Scamaldo, K. M., Vidaña, A. G., Richmond, J. R., & Tull, M. T. (2021). Prospective interactive influence of financial strain and emotional nonacceptance on problematic alcohol use during the COVID-19 pandemic. *The American Journal of*

Drug and Alcohol Abuse, 47(1), 107–116. https://doi.org/10.1080/00952990.2020.1849248

Greene, K. M., Hedstrom, A. M., & Murphy, S. T. (2019). Driving/riding after alcohol and marijuana use among young adults: Is residing with family protective?. *Traffic Injury Prevention, 20(7),* 679–684. https://doi.org/10.1080/15389588.2019.1641597

Gupta, N., Rigotti, L., & Wilson, A. (2021). The experimenters' dilemma: Inferential preferences over populations. https://arxiv.org/pdf/2107.05064.pdf

Godinho, A., Schell, C., & Cunningham, J. A. (2020). Out damn bot, out: Recruiting real people into substance use studies on the internet. *Substance Abuse,* 41(1), 3-5. doi:10.1080/08897077.2019.1691131

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly, 31*(4), 101384. https://doi.org/10.1016/j.leaqua.2020.101384

Göritz, A. S., Borchert, K., & Hirth, M. (2019). Using attention testing to select crowdsourced workers and research participants. *Social Science Computer Review, 39*(1), 84-104. https://doi.org/10.1177/0894439319848726

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

Hakulinen, C., Elovainio, M., Batty, G. D., Virtanen, M., Kivimäki, M., & Jokela, M. (2015). Personality and alcohol consumption: Pooled analysis of 72,949 adults from eight cohort studies. *Drug & Alcohol Dependence, 151,* 110-114. https://doi.org/10.1016/j.drugalcdep.2015.03.008

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral & Brain Sciences, 33*(2-3), 61-83. https://doi.org/10.1017/S0140525X0999152X

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ (Clinical research ed.), 327*(7414), 557-560. https://doi.org/10.1136/bmj.327.7414.557

Hobday, M., Lensvelt, E., Gordon, E., Liang, W., Meuleners, L., & Chikritzhs, T. (2017). Distance travelled to purchase alcohol and the mediating effect of price. *Public health, 144*, 48–56. https://doi.org/10.1016/j.puhe.2016.11.019

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business &Psychology, 27*(1), 99-114. https://doi.org/10.1007/s10869-011-9231-8

Huhn, A. S., Strain, E. C., Jardot, J., Turner, G., Bergeria, C. L., Nayak, S., & Dunn, K. E. (2021). Treatment Disruption and Childcare Responsibility as Risk Factors for Drug and Alcohol Use in Persons in Treatment for Substance Use Disorders During the COVID-19 Crisis. *Journal of addiction medicine, 10.1097/ADM.0000000000000813.* Advance online publication. https://doi.org/10.1097/ADM.0000000000000813

Imrey, P. B. (2020). Limitations of meta-analyses of studies with high heterogeneity. *JAMA Network Open, 3*(1), e1919325-e1919325. https://doi.org/10.1001/jamanetworkopen.2019.19325

Jain, J. P., Offer, C., Rowe, C., Turner, C., Dawson-Rose, C., Hoffmann, T., & Santos, G. M. (2021). The Psychosocial Predictors and Day-Level Correlates of Substance Use Among Participants Recruited via an Online Crowdsourcing Platform in the United States: Daily Diary Study. *JMIR public health and surveillance, 7(4),* e23872. https://doi.org/10.2196/23872

Jongenelis, M. I., Pratt, I. S., Slevin, T., Chikritzhs, T., Liang, W., & Pettigrew, S. (2018). The effect of chronic disease warning statements on alcohol-related health beliefs and

consumption intentions among at-risk drinkers. *Health Education Research, 33(5),* 351–360. https://doi.org/10.1093/her/cyy025

Jones, A., & Field, M. (2015). Alcohol-related and negatively valenced cues increase motor and oculomotor disinhibition in social drinkers. *Experimental and Clinical Psychopharmacology, 23(2),* 122–129. https://doi.org/10.1037/pha0000011

Kaplan, B. A., & Reed, D. D. (2018). Happy hour drink specials in the Alcohol Purchase Task. *Experimental and Clinical Psychopharmacology, 26(2),* 156–167. https://doi.org/10.1037/pha0000174

Kim, H. S., & Hodgins, D. C. (2017). Reliability and validity of data obtained from alcohol, cannabis, and gambling populations on Amazon's Mechanical Turk. *Psychology of Addictive Behaviors 31(1),* 85–94. https://doi.org/10.1037/adb0000219

King, K. M., Kim, D. S., & McCabe, C. J. (2018). Random responses inflate statistical estimates in heavily skewed addictions data. *Drug & Alcohol Dependence, 183,* 102-110. https://doi.org/10.1016/j.drugalcdep.2017.10.033

Kristan, J., & Suffoletto, B. (2015). Using online crowdsourcing to understand young adult attitudes toward expert-authored messages aimed at reducing hazardous alcohol consumption and to collect peer-authored messages. *Translational Behavioral Medicine, 5(1),* 45–52. https://doi.org/10.1007/s13142-014-0298-4

Kuerbis, A., Muench, F. J., Lee, R., Pena, J., & Hail, L. (2016). An exploratory pilot study of mechanisms of action within normative feedback for adult drinkers. *PeerJ, 4,* e2114. https://doi.org/10.7717/peerj.2114

Kung, F. Y. H., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity? *Applied Psychology, 67*(2), 264-283. https://doi.org/10.1111/apps.12108

Landers, R. (2016). Calculating LongString in Excel to detect careless responders. *Excel Macro.* https://doi.org/10.22541/au.160590023.35753324/v1

Lang, B., & Rosenberg, H. (2017). Public perceptions of behavioral and substance addictions. *Psychology of Addictive Behaviors 31(1),* 79–84. https://doi.org/10.1037/adb0000228

Levitt, E. E., Amlung, M. T., Gonzalez, A., Oshri, A., & MacKillop, J. (2021). Consistent evidence of indirect effects of impulsive delay discounting and negative urgency between childhood adversity and adult substance use in two samples. *Psychopharmacology, 238(7),* 2011–2020. https://doi.org/10.1007/s00213-021-05827-6

Li, E., Hing, N., Russell, A., & Vitartas, P. (2020). Impulsive Sports Betting: The Effects of Food or Substance Consumption. *Journal of gambling studies, 36(2),* 539–554. https://doi.org/10.1007/s10899-020-09938-1

Linden-Carmichael, A. N., Masters, L. D., & Lanza, S. T. (2020). "Buzzwords": Crowd-sourcing and quantifying U.S. young adult terminology for subjective effects of alcohol and marijuana use. *Experimental and Clinical Psychopharmacology, 28(6),* 632–637. https://doi.org/10.1037/pha0000344

Lin, L., & Xu, C. (2020). Arcsine-based transformations for meta-analysis of proportions: Pros, cons, and alternatives. *Health Science Reports, 3*(3), e178. doi:https://doi.org/10.1002/hsr2.178

Lovett, D. E., Ham, L. S., & Veilleux, J. C. (2015). Psychometric evaluation of a standardized set of alcohol cue photographs to assess craving. *Addictive behaviors, 48*, 58–61. https://doi.org/10.1016/j.addbeh.2015.05.002

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

McCredie, M. N., Harris, B., Regan, T., Morey, L. C., & Fields, S. A. (2021). Development and validation of a validity scale for use with the UPPS-P and Short UPPS-P Impulsive Behavior Scales. *Journal of Personality Assessment*, 1-18. https://doi.org/10.1080/00223891.2020.1866588

McKetin, R., Chalmers, J., Sunderland, M., & Bright, D. A. (2014). Recreational drug use and binge drinking: stimulant but not cannabis intoxication is associated with excessive alcohol consumption. *Drug and Alcohol Review*, *33*(4), 436–445. https://doi.org/10.1111/dar.12147

McPhee, M. D., Keough, M. T., Rundle, S., Heath, L. M., Wardell, J. D., & Hendershot, C. S. (2020). Depression, Environmental Reward, Coping Motives and Alcohol Consumption During the COVID-19 Pandemic. *Frontiers in Psychiatry, 11*, 574676. https://doi.org/10.3389/fpsyt.2020.574676

Meade, A., & Craig, S. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437-455. https://doi.org/10.1037/a0028085

Meadows, A. L., Strickland, J. C., Kerr, M. S., Rayapati, A. O., & Rush, C. R. (2019). Adverse Childhood Experiences, Tobacco Use, and Obesity: A Crowdsourcing Study. *Substance Use & Misuse, 54(10)*, 1743–1749. https://doi.org/10.1080/10826084.2019.1608254

Meisel, S. N., Colder, C. R., & Read, J. P. (2016). Addressing Inconsistencies in the Social Norms Drinking Literature: Development of the Injunctive Norms Drinking and Abstaining Behaviors Questionnaire. *Alcoholism, Clinical and Experimental Research, 40(10),* 2218–2228. https://doi.org/10.1111/acer.13202

Meredith, S. E., Sweeney, M. M., Johnson, P. S., Johnson, M. W., & Griffiths, R. R. (2016). Weekly Energy Drink Use Is Positively Associated with Delay Discounting and Risk

Behavior in a Nationwide Sample of Young Adults. *Journal of Caffeine Research, 6(1),* 10–19. https://doi.org/10.1089/jcr.2015.0024

Metcalf, D. A., Saliba, A., McKenzie, K., & Gao, A. (2021). Relationships between consumption patterns, health beliefs, and subjective wellbeing in Chinese Baijiu consumers. *Substance Abuse Treatment, Prevention, and Policy, 16(1),* 31. https://doi.org/10.1186/s13011-021-00369-8

Meyer, J., Faust, K., Faust, D., Baker, A., & Cook, A. (2013). Careless and random responding on clinical and research measures in the addictions. *International Journal of Mental Health &Addiction, 11*, 292-306. https://doi.org/10.1007/211469-012-9410-5

Miller, J. J. (1978). The Inverse of the Freeman – Tukey Double Arcsine Transformation. *The American Statistician, 32*(4), 138-138. doi:10.1080/00031305.1978.10479283

Morris, V., Patel, H., Vedelago, L., Reed, D. D., Metrik, J., Aston, E., MacKillop, J., & Amlung, M. (2018). Elevated Behavioral Economic Demand for Alcohol in Co-Users of Alcohol and Cannabis. *Journal of Studies on Alcohol and Drugs, 79(6),* 929–934. https://doi.org/10.15288/jsad.2018.79.929

Morris, V. L., Huffman, L. G., Naish, K. R., Holshausen, K., Oshri, A., McKinnon, M., & Amlung, M. (2020). Impulsivity as a mediating factor in the association between posttraumatic stress disorder symptoms and substance use. *Psychological trauma : theory, research, practice and policy, 12(6),* 659–668. https://doi.org/10.1037/tra0000588

Morris, V., Amlung, M., Kaplan, B. A., Reed, D. D., Petker, T., & MacKillop, J. (2017). Using crowdsourcing to examine behavioral economic measures of alcohol value and proportionate alcohol reinforcement. *Experimental & Clinical Psychopharmacology, 25*, 314-321. https://doi.org/10.1037/pha0000130

Moss, A. J., Rosenzweig, C., Robinson, J., & Litman, L. (2020). Demographic stability on Mechanical Turk despite COVID-19. *Trends in Cognitive Sciences, 24*(9), 678-680. https://doi.org/10.1016/j.tics.2020.05.014

Mullen, P. R., Fox, J., Goshorn, J. R., & Warraich, L. K. (2021). Crowdsourcing for online samples in counseling research. *Journal of Counseling & Development, 99*(2), 221-226. https://doi.org/10.1002/jcad.12369

Noel J. K. (2021). Using social media comments to reduce alcohol purchase intentions: An online experiment. *Drug and Alcohol Review, 40(6),* 1047–1055. https://doi.org/10.1111/dar.13262

Nichols, A. L., & Edlund, J. E. (2020). Why don't we care more about carelessness? Understanding the causes and consequences of careless participants. *International Journal of Social Research Methodology, 23*(6), 625-638. https://doi.org/10.1080/13645579.2020.1719618

O'Hara, R. E., Wang, W., & Troisi, J. D. (2020). Thanksgiving Day Alcohol Use: Associations With Expectations and Negative Affect. *Psychological Reports, 123(3),* 741–758. https://doi.org/10.1177/0033294119835763

Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). GOSH - a graphical display of study heterogeneity. *Res Synth Methods, 3*(3), 214-223. https://doi.org/10.1002/jrsm.1053

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance, 17,* 22-27. https://doi.org/10.1016/j.jbef.2017.12.004

Paulus, D. J., Rogers, A. H., Bakhshaie, J., Vowles, K. E., & Zvolensky, M. J. (2019). Pain severity and prescription opioid misuse among individuals with chronic pain: The moderating role of alcohol use severity. *Drug and Alcohol Dependence, 204,* 107456. https://doi.org/10.1016/j.drugalcdep.2019.02.036

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23,* 184-188. https://doi.org/10.1177%2F0963721414531598

Perski, O., Lumsden, J., Garnett, C., Blandford, A., West, R., & Michie, S. (2019). Assessing the Psychometric Properties of the Digital Behavior Change Intervention Engagement Scale in Users of an App for Reducing Alcohol Consumption: Evaluation Study. *Journal of Medical Internet Research, 21(11*), e16197. https://doi.org/10.2196/16197

Peterson, H., Simpson, S. L., & Laurienti, P. J. (2019). Wake Forest Alcohol Imagery Set: Development and Validation of a Large Standardized Alcohol Imagery Dataset. *Alcoholism, Clinical and Experimental Research, 43(12),* 2559–2567. https://doi.org/10.1111/acer.14214

Pettigrew, S., Jongenelis, M., Chikritzhs, T., Slevin, T., Pratt, I. S., Glance, D., & Liang, W. (2014). Developing cancer warning statements for alcoholic beverages. *BMC Public Health*, 14, 786. https://doi.org/10.1186/1471-2458-14-786

Pettigrew, S., Jongenelis, M. I., Glance, D., Chikritzhs, T., Pratt, I. S., Slevin, T., Liang, W., & Wakefield, M. (2016). The effect of cancer warning statements on alcohol consumption intentions. *Health Education Research, 31(1),* 60–69. https://doi.org/10.1093/her/cyv067

Phung, Q. H., Snider, S. E., Tegge, A. N., & Bickel, W. K. (2019). Willing to Work But Not to Wait: Individuals with Greater Alcohol Use Disorder Show Increased Delay Discounting Across Commodities and Less Effort Discounting for Alcohol. *Alcoholism, Clinical and Experimental Research, 43(5),* 927–936. https://doi.org/10.1111/acer.13996

Ramírez, J., Sayin, B., Baez, M., Casati, F., Cernuzzi, L., Benatallah, B., & Demartini, G. (2021). On the State of Reporting in Crowdsourcing Experiments and a Checklist to

Aid Current Practices. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW2), Article 387. doi:10.1145/3479531

Ranard, B. L., Ha, Y. P., Meisel, Z. F., Asch, D. A., Hill, S. S., Becker, L. B., Seymour, A. K., & Merchant, R. M. (2014). Crowdsourcing—Harnessing the masses to advance health and medicine, a systematic review. *Journal of General Internal Medicine*, *29*(1), 187-203. https://doi.org/10.1007/s11606-013-2536-8

Reynolds, J. P., Archer, S., Pilling, M., Kenny, M., Hollands, G. J., & Marteau, T. M. (2019). Public acceptability of nudging and taxing to reduce consumption of alcohol, tobacco, and food: A population-based survey experiment. *Social Science & Medicine, 236*, 112395. https://doi.org/10.1016/j.socscimed.2019.112395

Rea, S. C., Kleeman, H., Zhu, Q., Gilbert, B., & Yue, C. (2020). Crowdsourcing as a Tool for Research: Methodological, Fair, and Political Considerations. B*ulletin of Science, Technology & Society*, *40(*3-4), 40-53. doi:10.1177/02704676211003808

Reisner, S. L., Greytak, E. A., Parsons, J. T., & Ybarra, M. L. (2015). Gender minority social stress in adolescence: disparities in adolescent bullying and substance use by gender identity. *Journal of Sex Research*, *52(*3), 243–256. https://doi.org/10.1080/00224499.2014.886321

Reynolds, J. P., Archer, S., Pilling, M., Kenny, M., Hollands, G. J., & Marteau, T. M. (2019). Public acceptability of nudging and taxing to reduce consumption of alcohol, tobacco, and food: A population-based survey experiment. *Social Science & Medicine (1982)*, *236*, 112395. https://doi.org/10.1016/j.socscimed.2019.112395

Robinson, E., Gillespie, S., & Jones, A. (2020). Weight-related lifestyle behaviours and the COVID-19 crisis: An online survey study of UK adults during social lockdown. *Obesity Science & Practice*, *6(6)*, 735–740. https://doi.org/10.1002/osp4.442

Robinson, E., Smith, J., & Jones, A. (2021). The effect of calorie and physical activity equivalent labelling of alcoholic drinks on drinking intentions in participants of higher and lower socioeconomic position: An experimental study. *British Journal of Health Psychology, advanced online publication.* https://doi.org/10.1111/bjhp.12527

Rodriguez, L. M., Litt, D. M., & Stewart, S. H. (2020). Drinking to cope with the pandemic: The unique associations of COVID-19-related perceived threat and psychological distress to drinking behaviors in American men and women. *Addictive Behaviors, 110,* 106532. https://doi.org/10.1016/j.addbeh.2020.106532

Russell, A. M., & Barry, A. E. (2021). Psychometric Properties of the AUDIT-C within an Amazon Mechanical Turk Sample. *American Journal of Health Behavior, 45(4),* 695–700. https://doi.org/10.5993/AJHB.45.4.8

Schell, C., Godinho, A., & Cunningham, J. A. (2021). To thine own self, be true: Examining change in self-reported alcohol measures over time as related to socially desirable responding bias among people with unhealthy alcohol use. *Substance Abuse, 42(1),* 87–93. https://doi.org/10.1080/08897077.2019.1697998

Schroeders, U., Schmidt, C., & Gnambs, T. (2021). Detecting careless responding in survey data using stochastic gradient boosting. *Educational & Psychological Measurement,* 00131644211004708. https://doi.org/10.1177/00131644211004708

Scully, M., Brennan, E., Durkin, S., Dixon, H., Wakefield, M., Barry, C. L., & Niederdeppe, J. (2017). Competing with big business: a randomised experiment testing the effects of messages to promote alcohol and sugary drink control policy. *BMC Public Health, 17(1),* 945. https://doi.org/10.1186/s12889-017-4972-6

Siegel, M., Ayers, A. J., DeJong, W., Naimi, T. S., & Jernigan, D. H. (2015). Differences in alcohol brand consumption among underage youth by age, gender, and race/ethnicity

- United States, 2012. *Journal of Substance Use, 20(6),* 430–438. https://doi.org/10.3109/14659891.2014.942402

Siegel, M., Ramirez, R. L., Ross, C., DeJong, W., Albers, A. B., & Jernigan, D. H. (2014). Brand-specific consumption of flavored alcoholic beverages among underage youth in the United States. *The American Journal of Drug and Alcohol Abuse, 40(1),* 51–57. https://doi.org/10.3109/00952990.2013.841712

Skinner, K. D., & Veilleux, J. C. (2016). The Interactive Effects of Drinking Motives, Age, and Self-Criticism in Predicting Hazardous Drinking. *Substance Use & Misuse, 51(10),* 1342–1352. https://doi.org/10.3109/10826084.2016.1168448

Slater, M. D., Hayes, A. F., Goodall, C. E., & Ewoldsen, D. R. (2012). Increasing support for alcohol-control enforcement through news coverage of alcohol's role in injuries and crime. *Journal of Studies on Alcohol and Drugs, 73(2),* 311–315. https://doi.org/10.15288/jsad.2012.73.311

Silvana, C., & Kim, N. (2019). Conducting Survey Research Using MTurk. In A. Information Resources Management (Ed.), *Crowdsourcing: Concepts, methodologies, tools, and applications* (pp. 410-439). Hershey, PA, USA: IGI Global.

Spijkerman, R., Roek, M. A., Vermulst, A., Lemmers, L., Huiberts, A., & Engels, R. C. (2010). Effectiveness of a web-based brief alcohol intervention and added value of normative feedback in reducing underage drinking: a randomized controlled trial. *Journal of Medical Internet Research, 12(5),* e65. https://doi.org/10.2196/jmir.1465

Stautz, K., & Marteau, T. M. (2016). Viewing alcohol warning advertising reduces urges to drink in young adults: an online experiment. *BMC public health, 16,* 530. https://doi.org/10.1186/s12889-016-3192-9

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences, 21*(10), 736-748. https://doi.org/10.1016/j.tics.2017.06.007

Strickland, J. C., Alcorn, J. L., 3rd, & Stoops, W. W. (2019). Using behavioral economic variables to predict future alcohol use in a crowdsourced sample. *Journal of Psychopharmacology (Oxford, England), 33(7),* 779–790. https://doi.org/10.1177/0269881119827800

Strickland, J. C., & Bergeria, C. L. (2020). Contribution of alcohol- and cigarette-related cues to concurrent reinforcer choice in humans. *Behavioural Processes, 176*, 104124. https://doi.org/10.1016/j.beproc.2020.104124

Strickland, J. C., & Stoops, W. W. (2017). Stimulus selectivity of drug purchase tasks: A preliminary study evaluating alcohol and cigarette demand. *Experimental and Clinical Psychopharmacology, 25(*3), 198–207. https://doi.org/10.1037/pha0000123

Strickland, J. C., & Victor, G. A. (2020). Leveraging crowdsourcing methods to collect qualitative data in addiction science: Narratives of non-medical prescription opioid, heroin, and fentanyl use. *International Journal of Drug Policy, 75*, 102587. https://doi.org/10.1016/j.drugpo.2019.10.013

Strickland, J. C., & Stoops, W. W. (2019). Feasibility, acceptability, and validity of crowdsourcing for collecting longitudinal alcohol use data. *Journal of the Experimental Analysis of Behavior, 110*(1), 136-153. https://doi.org/10.1002/jeab.445

Strickland, J. C., & Stoops, W. W. (2019). The use of crowdsourcing in addiction science research: Amazon Mechanical Turk. *Experimental & Clinical Psychopharmacology, 27*(1), 1-18. https://doi.org/10.1037/pha0000235

Sun, L., & Feng, Y. (2019). Can results of meta-analysis with high heterogeneity provide any predictive values? *European Heart Journal*, *40*(38), 123-129.https://doi.org/10.1093/eurheartj/ehz530

Sunderland, M., Chalmers, J., McKetin, R., & Bright, D. (2014). Typologies of alcohol consumption on a Saturday night among young adults. *Alcoholism, clinical and Experimental Research*, *38*(6), 1745–1752. https://doi.org/10.1111/acer.12400

Tsai, J., Elbogen, E. B., Huang, M., North, C. S., & Pietrzak, R. H. (2021). Psychological distress and alcohol use disorder during the COVID-19 era among middle- and low-income U.S. adults. *Journal of Affective Disorders*, 288, 41–49. https://doi.org/10.1016/j.jad.2021.03.085

Vowles, K. E., Witkiewitz, K., Pielech, M., Edwards, K. A., McEntee, M. L., Bailey, R. W., Bolling, L., & Sullivan, M. D. (2018). Alcohol and Opioid Use in Chronic Pain: A Cross-Sectional Examination of Differences in Functioning Based on Misuse Status. *The Journal of Pain*, *19*(10), 1181–1188. https://doi.org/10.1016/j.jpain.2018.04.013

Wakefield, M. A., Brennan, E., Dunstone, K., Durkin, S. J., Dixon, H. G., Pettigrew, S., & Slater, M. D. (2017). Features of alcohol harm reduction advertisements that most motivate reduced drinking among adults: an advertisement response study. *BMJ open*, 7(4), e014193. https://doi.org/10.1136/bmjopen-2016-014193

Waites, S. F., & Ponder, N. (2016). *May I Have Your Attention Please? The Effectiveness of Attention Checks in Validity Assessment.* Paper presented at the Celebrating America's Pastimes: Baseball, Hot Dogs, Apple Pie and Marketing?, Cham.

Ward, M. K., Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of

personality traits and performance. *Computers in Human Behavior*, *76*, 417-430. https://doi.org/10.1016/j.chb.2017.06.032

Wardell, J. D., Kempe, T., Rapinda, K. K., Single, A., Bilevicius, E., Frohlich, J. R., Hendershot, C. S., & Keough, M. T. (2020). Drinking to Cope During COVID-19 Pandemic: The Role of External and Internal Factors in Coping Motive Pathways to Alcohol Use, Solitary Drinking, and Alcohol Problems. *Alcoholism, Clinical and Experimental Research*, *44*(10), 2073–2083. https://doi.org/10.1111/acer.14425

Wazny, K. (2017). "Crowdsourcing" ten years in: A review. *Journal of Global Health, 7*(2), 020602-020602. https://doi.org10.7189/jogh.07.020602

Weiss, N. H., Forkus, S. R., Raudales, A. M., Schick, M. R., & Contractor, A. A. (2020). Alcohol misuse to down-regulate positive emotions: A cross-sectional multiple mediator analysis among US military veterans. *Addictive Behaviors, 105*, 106322. https://doi.org/10.1016/j.addbeh.2020.106322

Wittleder, S., Kappes, A., Oettingen, G., Gollwitzer, P. M., Jay, M., & Morgenstern, J. (2019). Mental contrasting with implementation intentions reduces drinking when drinking is hazardous: An online self-regulation intervention. *Health Education, 46*(4), 666–676. https://doi.org/10.1177/1090198119826284

Wray, J. M., Funderburk, J. S., Gass, J. C., & Maisto, S. A. (2019). Barriers to and facilitators of delivering brief tobacco and alcohol interventions in integrated primary care settings. *The Primary Care Companion for CNS Disorders*, *21*(6), 19m02497. https://doi.org/10.4088/PCC.19m02497

Zamboanga, B. L., Napper, L. E., George, A. M., & Olthuis, J. V. (2019). Examining drinking game harms as a function of gender and college student status. *Psychology of Addictive Behaviors*, 33(8), 685–696. https://doi.org/10.1037/adb0000520

Zhang, M. W., Ward, J., Ying, J. J., Pan, F., & Ho, R. C. (2016). The alcohol tracker application: an initial evaluation of user preferences. *BMJ Innovations*, 2(1), 8–13. https://doi.org/10.1136/bmjinnov-2015-000087

Figure 1: PRISMA flow diagram for study selection

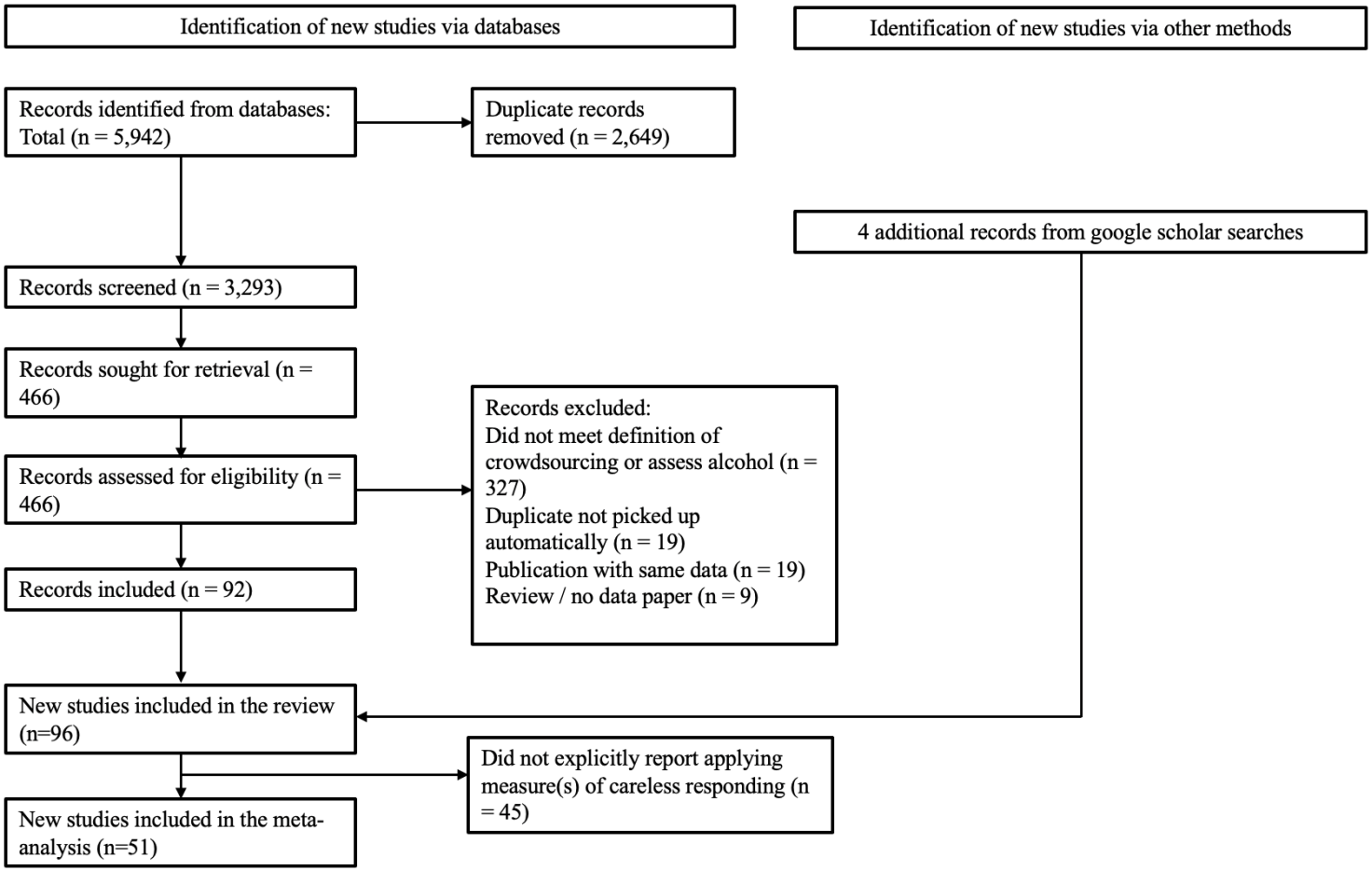

Figure 2: Forest plot for prevalence of careless responding in identified studies.

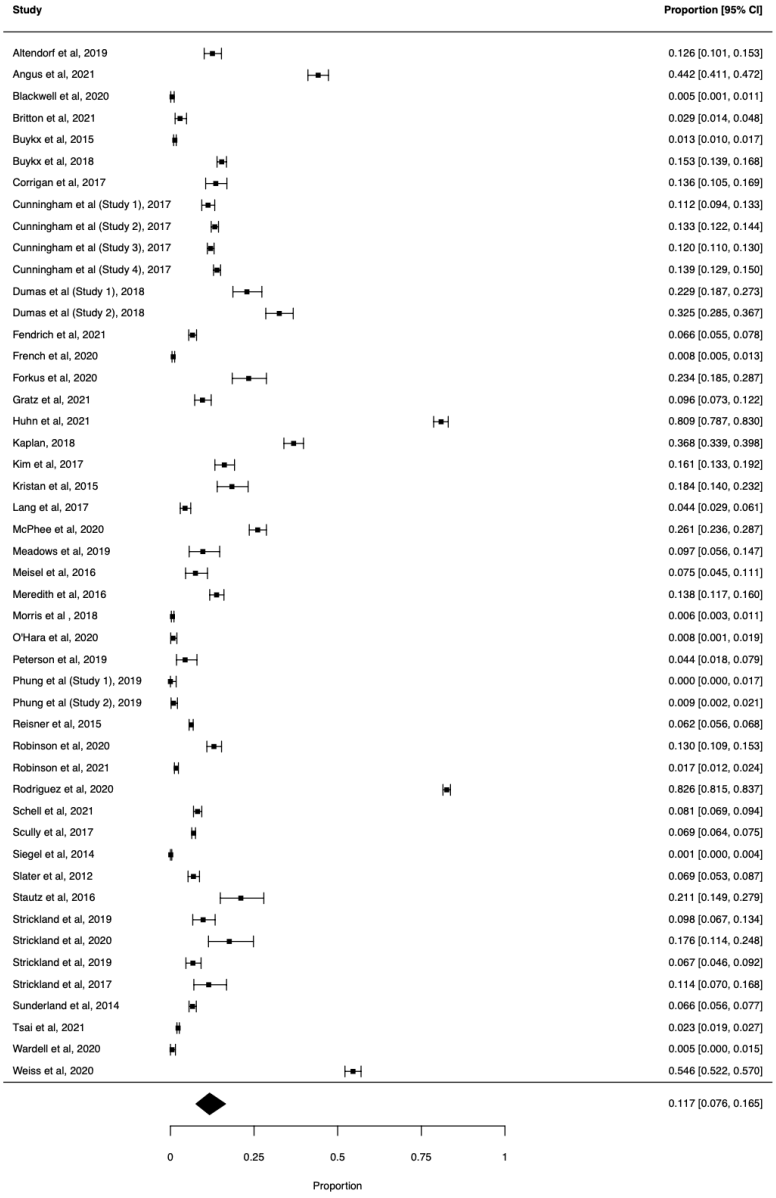

Figure 3: Meta-regression plot demonstrating the association between number of careless measures and prevalence of careless responding (values are raw proportions for ease of interpretation). Size of data points are reflective of sample sizes.

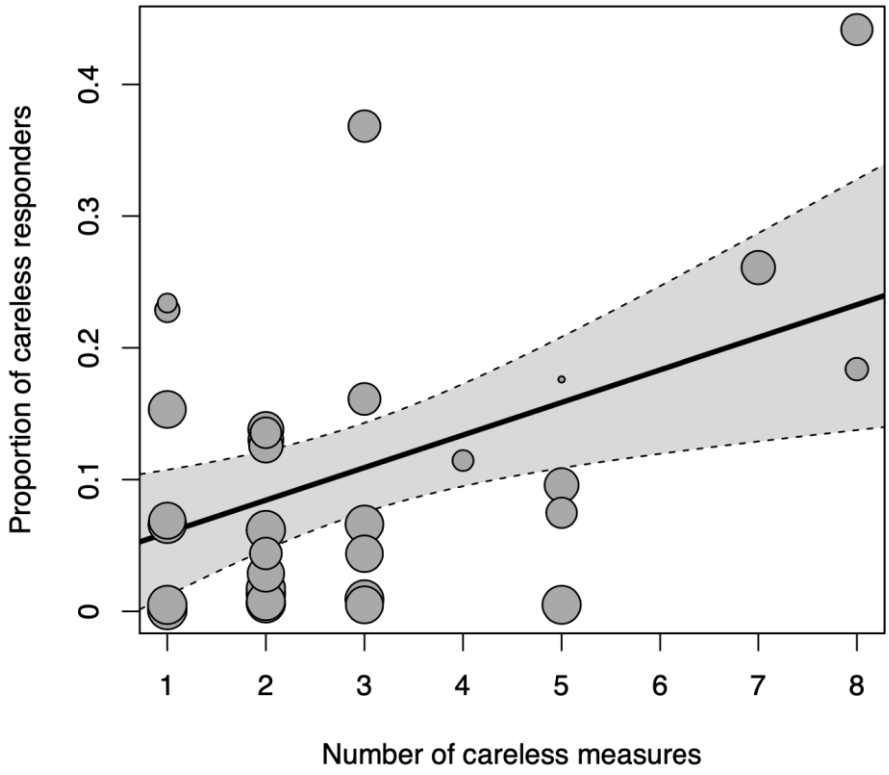

Table 1: Characteristics of studies included in prevalence meta-analysis.

Study ID	**Platform**	**Country**	**Sample demographics**	**Sample Size**	**Careless**	**Type of careless responding**	**Specific information**
Altendorf et al (2019)	PanelClix	Netherlands	Age_M = 46.6 Gender %f = 39.7	637	80	Implausible completion time (N = 1) Integrity check (N = 1)	"participated for too brief a period in the survey (i.e. had z scores >3 for participation time) were excluded." "respondents who gave contradictory responses to questions."
Angus et al (2021)	MTurk	USA	Age_M = 37.95 Gender %f = 59.74 AUDIT = 7.85 (alcohol group, only)	1010	446	Attention check (N = 4) Integrity check (N = 4)	"Three attention check items consisted of simple probe questions (e.g., "To continue, 'select 'strongly agree')." "Response integrity was assessed by examining missing responses for key variables; incomprehensible responses to open-text items; inconsistent responses to matched demographic items; inconsistent responses to paraphrased check items from the PGSI or AUDIT; failing to successfully complete a recaptcha checkbox".
Blackwell et al (2020)	Prolific	UK	Age_M = 38.2 Gender %f = 44 AUDIT = 9.65	812	4	Attention check (N = 1)	"An attention check was embedded within the questions post-randomisation: 'When was the last time you flew to Mars?' ('never'; 'a few days ago'; 'weeks ago'; 'months ago')".
Britton et al	Qualtrics	USA	Age_M = 33.2	383	11	Implausible	"To ensure valid responses, a speeding check was

(2021)	Panel		Gender %f = 59 AUDIT-C = 4.71			completion time (N = 1) Integrity check (N = 1)	included - one-half the median survey completion time – to screen out those who were not responding thoughtfully". [from Supplementary materials]: "Excluded for Data Quality Purposes • Qualtrics 'speeder' quality control (n=6) • Indicated impossible values for length of time living in US (n=5; 1 not unique)"
Buykx et al (2015)	Market Research Company	Australia	Age_M = 46.8 Gender %F = 50.8 AUDIT C = 3.98	3345	44	Implausible completion time (N = 1) Integrity check (N = 1)	"participants who completed the survey in less than a third of the median time (22 min) or exhibited low variability across pre-determined rating scale items."
Buykx et al (2018)	Vision One	UK	Age_M = 48 Gender %F 49.7 AUDIT C = 4.7	2480	380	Invalid responses (N = 1)	No information given other than "Following exclusion of 380 respondents who provided incomplete or invalid responses, a final sample of 2100 was obtained".
Corrigan et al (2017)	MTurk	USA	Age_M = 35.2 Gender %F = 45	450	61	Implausible completion time (N = 1) Attention check (N = 1)	"Please choose the number '4' for your answer below." "Additionally, people whose time on task was below minimal cutoff (3 minutes after viewing vignette) to complete the survey competently were excluded".
Cunningham et al. (2017, Study 1)	MTurk	USA/Canada	Age_M = 36.4 Gender %f = 52.1 AUDIT = 9.3	1023	115	Attention check (N = unclear) Honesty check (N = unclear)	No information given other than "Participants did not pass all attention checks" and "Participants who indicated that they did not respond honestly" in Table 1.
Cunningham et al. (2017, Study 2)	MTurk	USA/Canada	Age_M = 35.5 Gender %f = 53.2 AUDIT = 10.9	3740	496	Attention check (N = unclear) Honesty check (N = unclear)	Same as Study 1, above.
Cunningham et al. (2017, Study 3)	MTurk	USA/Canada	Age_M = 34.5 Gender %f =	4009	482	Attention check (N = unclear)	Same as Study 1, above.

			57.2 AUDIT = 10.8			Honesty check (N = unclear)	
Cunningham et al. (2017, Study 4)	MTurk	USA/Canada	Age_M = 33.5 Gender %f = 56.8 AUDIT = 10.1	4108	572	Attention check (N = unclear) Honesty check (N = unclear)	Same as Study 1, above.
Dumas et al (2018 Study 1)	MTurk	USA	Age_M = 26.3 Gender %f = 47.1	363	83	Attention check (N = 1)	"We excluded participants (n = 83) who did not adhere to validity questions (e.g., "please select the strongly agree option")."
Dumas et al (2018 Study 2)	MTurk	USA	Age_M = 25.45 Gender %f = 35.9	504	164	Attention check (N = unclear)	"Of the 504 original respondents, 164 (32.5%) were removed for not responding correctly to attention check items."
Fendrich et al. (2021)	MTurk	USA	Gender %f = 56	1742	115	Implausible completion time (N = 1)	"We eliminated surveys determined to be subpar due to quick responses (i.e., when respondent completion time was less than 10 min). … We screened out respondents who provided subpar responses due to quick completion times (n = 115)."
Forkus et al. (2020)	MTurk	USA	Age_M = 35.08 Gender %f = 22.7 AUDIT-C = 10.39	265	62	Attention check (N = unclear)	"attention checks were included throughout the survey to ensure participants were attentively reading and responding to the questions being asked (e.g., "Please select the color red from the options given")."
French et al. (2020)	MTurk	USA	Age_M = 41.41 Gender %f = 58.8	2040	17	Attention check (N = 2)	'Two items were randomly placed in the survey to provide an instructional manipulation check'
Gratz et al. (2021)	MTurk	USA	Age_M = 41.82 Gender %f = 50.4	553	53	Attention check (N = 4) Geolocation (N = 1)	"three explicit requests embedded within the questionnaires (e.g., "If you are paying attention, choose '2' for this question"), two multiple-choice questions (e.g., "How many words are in this sentence?"), a math problem (e.g., "What is 4 plus 2?"), and a free-response item (e.g., "Please briefly describe in a few sentences what you did in this study")".
Huhn et al. (2021)	MTurk	USA	Age_M = 33.6 Gender %f =	1257	1017	Data quality (N = unclear)	No information given other than "data quality issues."

			30				
Kaplan (2018)	MTurk	USA	Age_M = 35.5 Gender %f = 55.3	1040	383	Task specific measures (N = 3)	"Data were flagged for unsystematic patterns of responding by applying the three criteria proposed by Stein, Koffarnus, Snider, Quisenberry, and Bickel (2015)..."
Kim et al. (2017)	MTurk	USA	Gender %f = 40.89 AUDIT-C = 7.24	608	98	Honesty (N = 3), implausible completion time (N = 1), and unmatched T1/T2 responses (N = 1)	"Participants were asked to indicate on a 7-point Likert scale, from 1 (strongly disagree) to 7 (strongly agree), the following questions: "I answered the questions truthfully," "I paid close attention to the questions," and a question regarding the ease of answering sensitive questions on MTurk, "I find it easier to answer honestly to sensitive questions on MTurk, compared to an interview [in person or by phone]." "Participants were removed because their completion time was unusually fast"
Kristan et al. (2015)	MTurk	USA	Age_M = 23 Gender %f = 50	272	50	Attention check (N = 8)	"We included eight basic arithmetic questions interspersed throughout the 71 messages, with response options ranging from 0 to 7."
Lang et al. (2017)	MTurk	USA	Age_M = 34.28 Gender %f = 49	663	29	Attention check (N = 2), implausible completion time (N = 1)	"we excluded … participants who failed either of two checks of attention, …participants who took significantly longer than average (2.5 SDs) to complete the survey"
Linden-Carmichael et al. (2020, Study 1).	MTurk	USA	Age_M = 23 Gender %f = 53.5	323		Attention check (N = 3)	"attention checks were placed throughout the survey; work was approved if they answered two or more of the three attention checks correctly and if they provided plausible responses."
Linden-Carmichael et al. (2020, Study 2)	MTurk	USA	Age_M = 23 Gender % f = 46.4	289		Attention check (= 3)	"using the same eligibility criteria, attention checks [as Study 1]"
McPhee et al. (2020)	MTurk	USA	Age_M = 40.76 Gender %f = 35.3 AUDIT = 10.49	1127	294	Attention check (N = 5), Honesty (N = 2)	"Five attention-check questions were interspersed throughout the survey as a means of detecting random responding. Additionally, two questions appeared at the end of the survey asking the participant to confirm that they: (1) answered the questions honestly, and (2) paid attention to the questions."
Meadows et al. (2019)	MTurk	USA	Age_M = 38.6 Gender % f =	165	16	Attention check (N = unclear)	"Attention checks were included to identify nonsystematic, inconsistent, or inattentive

			66.4				responding"
Meisel et al. (2016)	MTurk	USA	Gender %f = 50.42	254	19	Attention check (N = 5)	"Accordingly, a 5-item internal consistency scale was developed for the current study to identify any unreliability in reporting. All questions on this scale were structured such that they should be answered with a zero. If participants entered a nonzero response to more than 2 of these items, then they were removed from subsequent analyses."
Meredith et al. (2016)	MTurk	USA	Age_M = 23.9 Gender %f = 62	1014	140	Attention check (N = 2)	"correctly answer two attention check questions (i.e., "trick" questions")."
Morris et al. (2018)	MTurk	USA	Age_M = 35.11 Gender % f = 54.9 AUDIT = 7.12	1643	10	Attention check (N = 1) Task Specific (N = 1)	"A single attention check item was presented following the instructional vignette to check for comprehension of APT instructions…responses on the APT were examined for non systematic data ."
O'Hara et al. (2020)	MTurk	USA	Age_M = 37.42 Gender % f = 55.7	392	3	Attention check (N = 2)	"We also embedded attention-check questions into the second and third surveys in order to remove participants whose responses may not be genuine. Any participant who missed both attention check questions … were removed from the analysis."
Peterson et al. (2019)	MTurk	USA	NA	182	8	Attention check (N = 2)	"The 2 validity checks asked, "What day falls between Tuesday and Thursday?" with multiple-choice responses of "Monday," "Wednesday," "Friday," and "Sunday" and "Please retype the sixth word from the following sentence: the quick brown fox jumped over the lazy red dogs."
Phung et al (2019, Study 1)	MTurk	USA	NA	100	0	Validation questions (N = 3)	"To be included in the final data set, participants needed to correctly answer at least 2 of 3 validation questions. These validation questions involved (i) typing out a 50-word sample prompt provided to them, (ii) recalling the number of words on a standard, double-spaced page, and (iii) answering for the immediate option when asked to choose between $100 now and $50 in 3 weeks."
Phung et al (2019, Study 2)	MTurk	USA	NA	432	4	Validation questions (N = 3)	Same as Study 1, above.
Reisner et al.	Harris Poll	USA	NA	5907	365	Implausible	"those who did not meet valid data requirements (e.g.,

(2015)	Online					completion time (N = 1) Inconsistent responses (N = 1)	time to complete survey was less than five minutes; self-reported age at the beginning and end of survey differed by more than one year) were dropped."
Robinson et al. (2020)	Prolific	UK	Age_M = 30.7 Gender %f = 67	907	118	Attention check (N = 2)	"Two attention checks were included in the survey (e.g., 'have you ever been to the planet Mars?') in order to identify any participants not completing questionnaire items as intended."
Robinson et al. (2021)	Prolific	UK	Age_M = 36 Gender %f = 51	1892	33	Attention check (N = 2)	"Participants were asked 'What planet do you live on?' (multiple choice, item included in demographic measure page) and 'I have blinked in the last 24 hours?' (5-point response scale ranging from very unlikely to very likely, item included in perceived behavioural effects of labelling page)."
Rodriguez et al. (2020)	Qualtrics Panels	USA	Age_M = 41.7 Gender %f = 50	4335	3581	Attention check (N = 2) Implausible completion time (N = 1)	"Two filter questions were included where participants were asked to select a specific answer. There was also one speeder check performed by Qualtrics."
Schell et al. (2021)	MTurk	USA	Age_M = 33.8 Gender %f = 46.1	1865	151	Attention check (N = unclear) and Honesty (N = unclear)	"Finally, attention checks and honesty questions were included with the survey tools at baseline and the latter at both time points in order to evaluate participant attention while completing the survey and improve the quality of the data."
Scully et al. (2017)	I-View	Australia	Gender %f = 57.8	8181	567	Data quality checks (N = unclear)	No information provided other than "standard quality control processes."
Siegel et al. (2014)	Knowledge Networks	USA	Gender %f = 58.5	1032	1	Inconsistent responses (N = 1)	"In an attempt to identify possible errant or implausible reports of the number of drinks consumed for particular brands, we examined the self-consistency of the survey responses by comparing the overall number of drinks per day that each respondent reported consuming in the past 30 days to the sum of the number of drinks that individual reported consuming of each brand."
Slater et al. (2012)	Knowledge Networks	USA	Age_M = 48.77 Gender %f = 50.02	843	58	Implausible completion time (N = 1)	"completed the entire study in less than 8 minutes (pretests indicated that it was impossible to actually read the news article and provide responses to each

							question so quickly)."
Stautz et al. (2016)	Youthsight	UK	Age_M = 21.47 Gender %f = 50 AUDIT = 8.8	152	32	Catch questions (N = unclear)	No information provided other than "'Catch' questions identified and screened out 32 participants who appeared not to be fully engaging with the study (leaving a study sample of 152)."
Strickland et al. (2017)	MTurk	USA	Gender %f = 54.7	166	19	Attention check (N =2) Honesty (N = 1) Data quality checks (N = 1)	"Several attention checks were used to identify inattentive or non-systematic participant data. These checks included: 1) comparison of age and sex responses at the start and end of the survey, 2) recall of a single digit number presented halfway through the survey that participants were instructed to remember and enter at the end of the survey, 3) an item that instructed participants to select a specific response (i.e., "Select 'A Little Bit'"), and 4) an item asking participants if they had been attentive and thought their data should be included."
Strickland et al. (2019a)	MTurk	USA	Age_M = 35.2 Gender %f = 52.9 AUDIT = 10.3	307	80	Data quality checks (N = unclear), Incomplete / Invalid Responding (N = 1) Task specific measures (N = 2).	"Thirty participants failed one or more data quality checks throughout the study, … "Fifty participants provided non-systematic data either violating these criteria (n=19) or reporting zero consumption at all prices (n=31) and four participants did not complete the alcohol purchase task data due to a technical error."
Strickland et al. (2019b)	MTurk	USA	Age_M = 34.3 Gender %f = 51.1 AUDIT = 12.7	476	32	Attention check (N = unclear) Data quality checks (N = unclear)	No information provided other than "Thirty-two of these participants failed 1 or more attention or data quality checks throughout the study and were removed from analysis".
Strickland et al. (2020)	MTurk	USA	Age_M = 35.2 Gender %f = 46.6 AUDIT = 6.6	125	22	Attention check (N = 2) Honesty (N = 1) Data quality check (N = 1) Incomplete	"Data were first evaluated for non-systematic or non-attentive responding. Checks included: 1) comparisons of age and gender at two points across the survey, 2) an item that instructed participants to select a particular response, 3) recall of a single digit number presented earlier in the survey that

						responding (N=1)	participants were instructed to remember, and 4) an item that asked if participants had been attentive and that their data should be used… An additional five participants were removed for failing to respond on more than 80 % of task trials indicative of either computer failure or inattention
Sunderland et al. (2014)	PureProfile	Australia	Age_M = 25	2013	133	Data Integrity / Task Specific (N = 3)	"Outliers were respondents who: (i) reported consuming more than 50 standard drinks on their last Saturday night (n = 101), as this would correspond to a blood alcohol level in the lethal range; (ii) paid more than 2 standard deviations above the mean for a standard drink (n = 20), or paid over 2 standard deviations above the mean for total alcohol consumed."
Tsai et al. (2021)	MTurk	USA	Gender %f = 47.9	6762	155	Validity Questions (N = 3)	"… failed three items from the validity scales from the Minnesota Multiphasic Personality Inventory-2 (MMPI-2), which included items: "It would be better if almost all laws were thrown away," ("yes" response was a validity failure) "Sometimes when I am not feeling well I am irritable," ("no" was validity failure) and "Once in a while I put off until tomorrow what I ought to do today" ("no" was validity failure)"
Wardell et al. (2020)	Prolific	Canada	Age_M = 31.99 Gender %f = 45.3	402	2	Attention check (N = 4) Implausible completion time (N = 1)	"We also included 4 attention check items … participants' data were automatically rejected from the study if they failed 2 or more attention check items and had a very fast completion time (defined as under 20 minutes in this study). Two participants were removed from the study based on these criteria."
Weiss et al. (2020)	MTurk	USA	Age_M = 37.74 Gender %f = 29.1 AUDIT = 9.74	1647	899	Attention / comprehension checks (N = 4)	"To improve data quality, we incorporated validity checks assessing attentive responding and comprehension (4 items; e.g., "I have never brushed my teeth")
Zhang et al. (2016)	No information provided.	Canada	Gender %f = 58	100		Implausible completion time (N = 1)	"A minimum time of 200 s was set for each questionnaire, and participants needed to spend a minimum of 200 s to fill up the survey."

Legend: AUDIT = Alcohol Use Disorders Identification Task.