Abstract

Objective: The prevalence of research conducted online in the addictions field has increased rapidly over the past decade. However, little focus has been given to careless responding in these online studies, despite the issues it may cause for statistical inference and generalisability. Our aim was to examine whether alcohol use is associated with careless responses. **Method:** Raw data was requested from online studies examining alcohol use and related problems which also addressed careless responding. We obtained 13 data sets of 12,237 participants (mean age = 42.16, *SD* = 15.65: 50.5% female). The sample had an average Alcohol Use Disorders Identification Test (AUDIT) score of 10.88 (*SD*=7.77). Predictors included demographic information (age, gender) and AUDIT total scores. The primary outcome was whether an individual was classed as a careless responder, for example by failing an explicit attention check question. **Results:** AUDIT total scores were associated with careless responding (OR=1.07 [95% CI: 1.06; 1.08], p < .001). Hazardous drinking or worse was associated with 2.21 greater odds (OR=2.21 [95% CI: 1.81; 2.71] of careless responding, whereas harmful drinking or worse was associated with 3.43 greater odds (OR=3.43 [95% CI: 2.83; 4.17]) and probable dependence was associated with 3.63 greater odds (OR=3.63 [95% CI: 2.95; 4.48]). **Conclusions:** Alcohol use and related problems are positively associated with careless responding in online research. Removal of individuals identified as careless responders may lead to issues of generalisability, and more care should be taken to identify and handle careless responder data.

Keywords: Alcohol Use Disorders Identification Task; Alcohol use; Attention; Careless responding; online research

Public health significance statements

This mega-analysis demonstrates a robust relationship between alcohol use (defined by the Alcohol Use Disorders Identification Test) and careless responding in online studies.

The findings suggest that the typical removal of careless responders from analyses in online alcohol studies is insufficient at best and at worst leads to issues with Statistical inference and generalisability.

Introduction

Conducting research online brings several benefits, including the recruitment of many participants quickly and efficiently, which greatly reduces the 'cost per observation' of studies. It allows for the recruitment of diverse and under-represented samples and overcomes geographical barriers (Jones et al., 2022). Studies assessing alcohol or other substance use online, as opposed to in-person, also benefit from the ability to measure consumption and behaviour without impression management concerns, or fear of stigmatisation (Groh, Ferrari, & Jason, 2009). These benefits have led to a massive increase in research conducted online in psychology, but also specifically addiction-related research (Strickland & Stoops, 2019).

One potential limitation of online research is the increased likelihood of careless responding (also known as 'insufficient effort responding'), which can be defined as intentional or unintentional responding that is not reflective of a participant's true nature. Careless responding by individuals can have detrimental consequences across studies, biasing effect size estimates, incorrectly categorising individuals with a psychiatric disorder (reducing specificity), and generally increasing noise within the data (Jones et al., 2022). In response to this, researchers have attempted to identify careless responders using 'attention checks', such as asking participants a question with one clear answer and several impossible answers (e.g. '*Which planet do you live on?*'). In this case, if participants chose anything other than 'Earth' they are classed as careless responders and likely removed from inferential analyses. Other methods also exist, such as infrequency scales, in which participants respond to statements such as '*I am answering a survey right now*' using a Likert scale with response options such as 'strongly disagree to strongly agree' (Kim, McCabe, Yamasaki, Louie, & King, 2018). For these methods, responses that occur infrequently ('strongly disagree' relative to 'strongly agree') are used to infer carelessness.

In a recent meta-analysis (Jones et al., 2022), we demonstrated that careless responding was prevalent in online studies examining alcohol use, with ~12% of participants (across 51 studies) being identified as careless. We examined various study-level predictors and demonstrated that only the number of careless response techniques used was a significant predictor of increased carelessness across studies, suggesting that the more attempts to identify carelessness increases the detection of carelessness. However, it is likely that various individual differences also contribute to carelessness within online studies, and it is important to isolate these to assess their influence on data quality.

Previous research has demonstrated that personality characteristics such as conscientiousness and agreeableness are negatively associated with carelessness (Bowling et al., 2016). Sociodemographic characteristics have also been identified as correlates of carelessness. For example, Berry et al., (Berry, Rana, Lockwood, Fletcher, & Pratt, 2019) demonstrated that male participants were more likely to be careless responders (but see (Ashley & Shaughnessy, 2021)), and Nichols and Edlund (Nichols & Edlund, 2020) demonstrated further that being male, younger, and college-educated were significant predictors of carelessness. However, there is a lack of evidence as to whether *individual differences* in alcohol-related variables are associated with careless responses. One USA study (Agley, Xiao, Nolan, & Golzarri-Arroyo, 2022) conducted via Amazon Mechanical Turk (MTurk) demonstrated significant differences in Alcohol Use Disorders Identification Test (AUDIT) scores from a sample-arm with no quality control (M_{AUDIT}=13.6, *SD* = 10.2) compared with a sample-arm with the addition of attention checks (M_{AUDIT} = 9.3, *SD* = 8.1: Cohen's *d* = 0.47). Assuming the initial randomisation was successful, these findings suggest that individuals who were removed by these attention checks had higher AUDIT scores. A substantial reduction in the proportion of participants meeting the cut-off for probable dependence was also identified between the two arms (30% vs 14.4%), and there was an increased negative skew in the arm with attention checks. The authors argued that individuals who fail quality control checks do not input random data but are more likely to report higher AUDIT scores.

However, it is also possible that careless responding may happen in several pseudo-random ways. First, individuals may respond uniformly (selecting each possible response with a similar probability); consistently (selecting the same response over several questions, known as long-string responding, which would be uniform if it was across the whole questionnaire) or even in a pattern (selecting 'a', then 'b', then 'c', and repeating this pattern; see (Kim et al., 2018)). Both uniform and long-string responding have been shown to cause an overinflation of associations between variables. To highlight this in the addictions field, King et al (King, Kim, & McCabe, 2018) used a large publicly available data set and replaced varying amounts of data (2.5%, 5%, and so on) with uniform or long-string random responses. They demonstrated even small amounts of random data could inflate a 'true' correlation between past-year alcohol use and closeness to their mother from *r* = .012 to .24 with long string responding and to *r* = .18 with uniform responding. Similar findings were also shown by Crede (Credé, 2010) who demonstrated even 5% of random responding can substantially inflate correlations.

It is therefore important to determine the cause and consequence of careless responses in relation to individual's alcohol use. Should carelessness be non-random (e.g., a function of increased alcohol-use) this raises concerns about excluding these individuals from alcohol-related research, as it creates data missing not at random, which can bias predictive models, excludes the very population of interest in many alcohol studies, and, in turn, reduces the generalisability of any findings. Should carelessness be random, this can potentially inflate estimates or scores on diagnostic tests (Meyer, Faust, Faust, Baker, & Cook, 2013) especially if the true distribution is positively skewed (larger number of lower values (King et al., 2018), as well as the correlations between alcohol and other variables of interest.

Therefore, the aim of this mega-analysis was to examine the predictors of careless responses in online studies examining alcohol use. Specifically, we preregistered two main research questions: 1) do increased AUDIT scores predict an increased likelihood of careless responding; 2) do demographic variables (e.g. age, gender) increase the likelihood of careless responding within alcohol-related studies[1]. Our preregistration can be found here: https://aspredicted.org/8mx7y.pdf .

Method

Participants and statistical power

We aimed to obtain individual data from studies that were conducted online, measured alcohol consumption, and implemented a measure of careless responding. As a formal systematic

[1] Note – we also hypothesised testing for mental health problems but too few studies had this information for it to be estimated reliably.

review was not feasible, we first extracted data from studies conducted in our own laboratory. Second, we emailed corresponding authors from our recent meta-analysis of careless responses in alcohol use (Jones et al., 2022) and requested the raw data. Finally, we conducted further scoping searches via Google Scholar searching the first 100 hits for (‘careless responding’ OR ‘attention check’) AND (alcohol OR ‘Alcohol Use Disorder Identification Task’) AND (online OR MTurk OR Qualtrics OR Prolific).

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. We conducted a post-hoc power calculation using a subset of the data used for the final analyses with the ‘simR’ package (Green & MacLeod, 2016) in R. Using data from 5 studies (1860 participants total) we observed a significant effect of AUDIT scores on careless responding (*b* = .07 [95% CI: .05 to .09]). With this information, we determined that we had 95% power [95% CI: 88.7 to 98.4] to detect this effect with a = .05. However, given the issues with typical post-hoc power calculations (Heckman, Davis & Crowson, 2022), we examined the sensitivity of our statistical power by simulating the inclusion of data from another 5 studies and re-estimating the statistical power. Doing so increased the statistical power to 99.9% [96.4 to 100%], assuming similar sample sizes of these studies. Therefore, we aimed to include 10 studies at a minimum for our data analyses. We searched in December 2021-January 2022, and then reran searches in December 2022 after the peer review process, identifying 2 further articles (Davies et al 2022, and Copeland et al, 2022). The analysis script for our power calculation can be found here: https://osf.io/49e5x/

In total, we obtained 13 data sets with 12,237 participants (see table 1). On average, participants were 42.16 years old (*SD*=15.65), 50.5% female and had an average AUDIT score of 10.88 (*SD* = 7.77). All studies were recruited in the UK/USA.

Measures

Demographic questions:

Where available, we extracted age, gender, education level and ethnicity of each participant within each study. With regards to gender, a small number of individuals identified as non-binary (N = 29, < 0.1%), however, this was not enough to create a statistically meaningful comparison group, and these were therefore removed from all our primary analyses (however, in online supplementary materials we included them in a male vs other category – notably the pattern of results was unaffected). For education level, there was considerable heterogeneity across studies in how this was measured, therefore we manually coded this to reflect higher (university/college degree or above) vs lower (educated to less than degree level), similar to previous research (Robinson, Smith, & Jones, 2022). For our outcome of carelessness, we created a binary variable (non-careless responder vs careless) based on whether participants had been identified as careless in the original studies (e.g., failed an attention check).

Data reduction and analysis

To maximise the sample size, we applied several models to examine our hypotheses. First, in model 1 we analysed age and gender as predictors of careless responding. In model 2, we included education (below undergraduate degree vs degree and above) and ethnicity (white vs non-white). In model 3 (testing our confirmatory hypothesis) we included age, gender, and total AUDIT scores. In exploratory models (models 4 – 6) we used AUDIT cut-offs to examine whether there was greater odds of careless responding in hazardous drinking or worse (AUDIT >7: Model 4) vs not; harmful drinking or worse (AUDIT >15: Model 5) vs not; and probable dependence (AUDIT >19: Model 6) vs not. We removed ethnicity and education from models 3 – 6 as their inclusion greatly reduced the available data.

Each model was analysed using a multilevel logistic regression, with a random intercept for study to adjust for dependent data points within individual studies. Across all models, there was limited evidence of multicollinearity (VIFS < 1.05). Intra-class correlation coefficients were calculated as Level 2 variance / (Level 2 variance + 3.29), and interpreted as the proportion of variance that is attributable to systematic differences between studies (Sommet & Morselli, 2017). We also computed the marginal R^2 of each model (variance explained by the fixed effects) using the ‘sjPlots’ package (Lüdecke, 2022).

In exploratory analyses, we visually examined the distribution of AUDIT scores separately for careless and non-careless responders, but also compared the distribution of careless and non-careless responders using a Kolmogorov-Smirnov test. Where question-level data was available for the AUDIT, we computed the long-string index for the first 8 items (which all have similar response options 0 – 4). Long-string index (Johnson, 2005) is the longest consecutive number of the same response, for example in the sequence of responses ‘1’,’2’,’1’,’3’,’3’,’3’,’3’,1’, the long-string index would be 4 (four consecutive ‘3’ responses). A rule of thumb is that individuals who have a long-string response greater than half the length of the scale are considered careless. We also calculated the intra-individual response variability

(IRV: (Dunn, Heggestad, Shanock, & Theilgard, 2018)), which is the within-person standard deviation of the raw scores. A small IRV is indicative of consistent responding (similar to long-string responding), however, a larger IRV may also be considered as highly random responding. We used the 'careless' R package (Yentes & Wilhelm, 2021) to compute these scores. Here, we removed any participants who only provided a positive score on the final 2 items of the AUDIT as these may reflect individuals who no longer drink but have been injured or had advice to cut back in the past. In this case, they would have a maximum long-string score (=8) but be a truthful responder[2]. In each case, we compared careless vs non-careless responders on the long-form and IRV scores, but also correlated these scores with total AUDIT scores.

Analysis scripts and data can be found here: https://osf.io/49e5x/ .

Results

Across all models, age was a negative predictor of careless responding, suggesting that younger participants were more likely to carelessly respond. In models 1 and 2 gender was a significant predictor; male participants had greater odds of careless responding compared to female participants. When AUDIT scores were included in the model, gender was no longer a significant predictor. Male participants had significantly higher AUDIT scores than female participants (Male=12.12 SD=8.15 , Female=9.55, SD=7.66, t(4684)=11.53, p<.001, d=0.33 [95% CI: 0.28 to 0.39).

[2] Indeed, Copeland et al (2022) specifically recruited individuals who had reduced their drinking in the previous months.

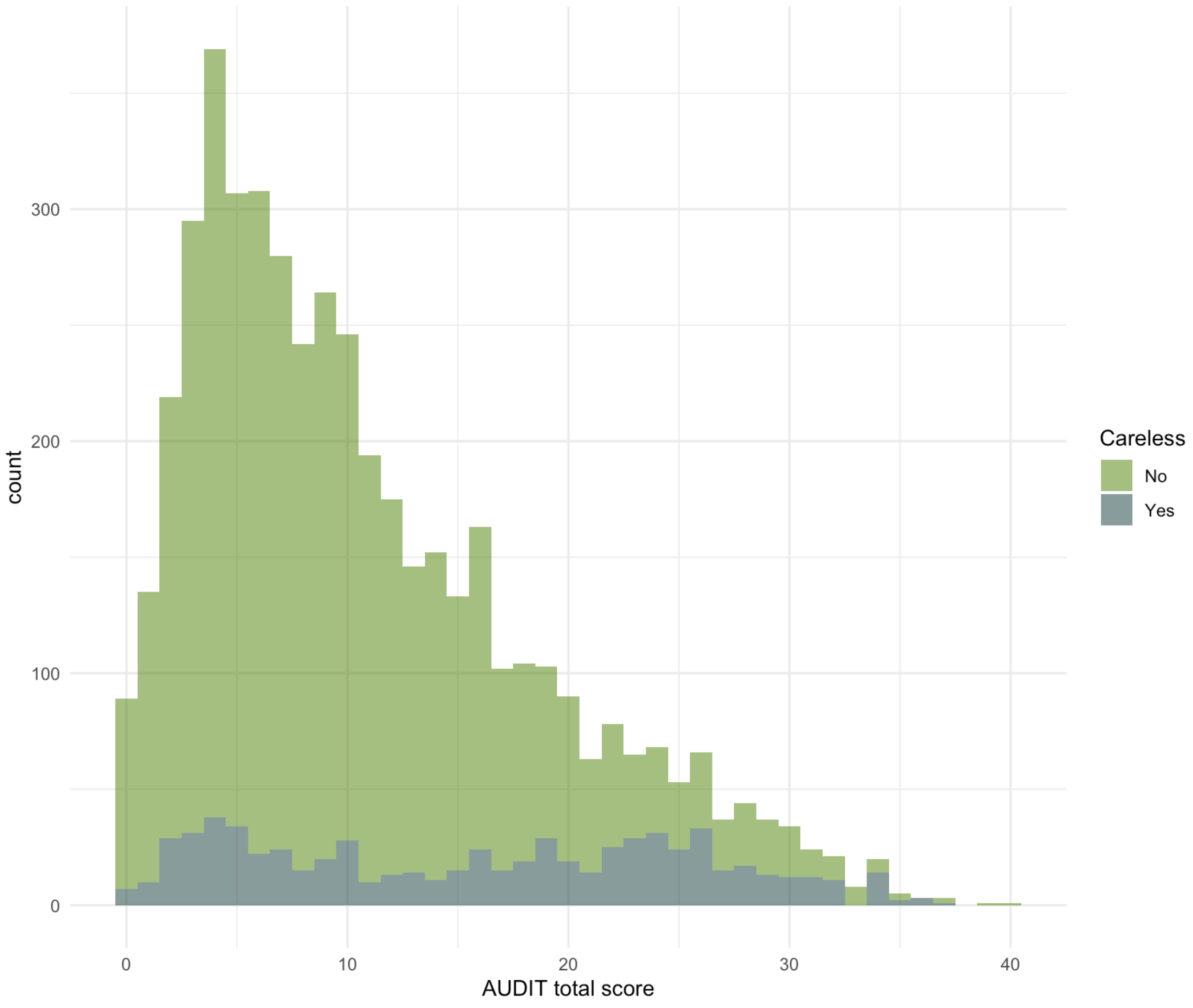
Overall, total AUDIT scores were associated with careless responding (OR = 1.07: model 3). Hazardous drinkers or worse (AUDIT>7) had 2.21 greater odds of careless responding (model 4). Harmful drinkers or worse (AUDIT>15) had 3.43 greater odds of careless responding (model 5). Individuals with probable dependence (AUDIT>19) had 3.63 greater odds of careless responding (model 6).

Across the whole sample which included demographic information, the removal of careless responders reduced the AUDIT score from 10.87 (*SD* = 7.76) to 10.03 (*SD* = 7.06). The AUDIT score for individuals identified as careless responders was 15.82 (*SD* = 9.69).

Exploratory analyses of carelessness

In line with the prediction that careless responding may follow a uniform distribution (e.g. equal likelihood of responding of AUDIT total scores from 0 – 40) we examined the distribution of AUDIT total scores in careless vs non-careless responders (see Figure 1). Distributions for the AUDIT total scores were visibly different, with careless responders having a more uniform distribution and non-careless responders having a skewed positive distribution as expected (King et al., 2018). A Kolmogorov-Smirnov test demonstrated a significant difference between the distributions (*D* = .33, *p* < .001). Note, that a Kolmogorov-Smirnov test was also significant when comparing the careless responders to several randomly simulated uniform distributions (all ps < .001).

Figure 1: Histogram of AUDIT total scores for careless vs non careless responders.



Exploratory analysis: Long-string responding

We computed long string scores for the first 8 items of the AUDIT. There was a significant difference in that non-careless responders had longer long-string scores (mean = 3.75, *SD* = 1.78) compared to careless responders (mean = 3.16, *SD* = 1.90: t(884) = 7.53, p < .001, d = 0.33 [95% CI: 0.25 to 0.41]). Across the complete sample, the correlation between long string score and total AUDIT was significant, *r*(4427) = -.59, p < .001. This suggests that increased AUDIT scores were associated with less consistent responses. Finally, we examined a cut-off of 4 (half the scale length) as a measure of carelessness. Those measured as careless had lower odds of long-string responding greater than the cut-off (OR = 0.607 [95% CI: 0.511 to 0.722], p < .001), compared to non-careless responders.

Exploratory analysis: Intra-individual response variability

We computed the intra-individual response variability for the first 8 items of the AUDIT. There was a significant difference in that careless responders had lower IRV scores (mean = 0.85, *SD* = 0.33) than non-careless responders (mean = 1.03, *SD* = 0.34: t(949) = 13.56, p < .001, d = 0.55 [95% CI: 0.46 to 0.63]). Across the complete sample, the correlation between IRV and total AUDIT was significant, *r*(4427) = .07, p < .001. This suggests that increased AUDIT scores were associated with increased variability in responses across the individual AUDIT questions, however the size of this association was small.

Discussion

We conducted a mega-analysis on online studies examining alcohol use and related problems which also addressed careless responding. Across 13 studies with >12,000 participants, we demonstrated a robust association between careless responding and alcohol use and related problems (AUDIT scores). We were also able to replicate previous findings suggesting that male (vs female) and younger participants are more likely to be careless responders (see (Berry et al., 2019; Nichols & Edlund, 2020)).

We demonstrated a reliable association between careless responding and AUDIT scores across multiple models. Specifically, hazardous drinkers had > 2 odds increase of being a careless responders, while harmful drinkers and those with probable dependence had > 3 times the odds. In line with data from Agley et al, (Agley et al., 2022) this may be explained in one of two ways; (1.) individuals with higher levels of alcohol use are more inattentive during online surveys, or (2.) careless respondents do not add random noise to the data, but instead, bias estimates of alcohol consumption upwards. In support of the former explanation, there is evidence to suggest that higher levels of alcohol use are associated with poor attention, impulsivity, and a general lack of cognitive abilities (Martins, Bartholow, Cooper, Von Gunten, & Wood, 2018), and general cognitive abilities are negatively associated with careless responding (*r*=-.38 :(Huang & DeSimone, 2020)) – however, this casual pathway needs testing directly in future studies.

Our exploratory analyses provide some tentative support for the latter explanation, however. First, we observed clear differences in the distribution of total AUDIT scores between careless and non-careless responders, with careless responders having a much more uniform distribution, and non-careless responders having a somewhat positively skewed distribution, as expected (Kehoe, Gmel, Shield, Gmel, & Rehm, 2012). Whilst a completely uniform distribution should lead to an average AUDIT total score ~ 20, the AUDIT score of our careless responders (~16) was lower than that but significantly higher than non-careless responders. This supports observations across other studies suggesting that careless responders bias survey scales upwards (Meyer et al., 2013), but also inflate correlations between measures (Credé, 2010; King et al., 2018).

The exploratory data-driven estimates of carelessness (long-string responding and inter-individual response variability) led to somewhat different conclusions. We observed individuals identified as careless via individual study methods (attention checks) were much less likely to respond consistently but also had lower variability in responses. These counterintuitive observations suggest that neither measure in isolation is particularly useful in identifying careless responders (Hong, Steedle, & Cheng, 2020), or that the AUDIT, being fairly short and not having any negatively worded items is not suitable for such methods (Curran, 2016; Schroeders, Schmidt, & Gnambs, 2021).

The identification of the robust relationship between careless responses and alcohol use has wider ramifications for online alcohol-related research. In most studies with measures of careless responding, identified individuals are removed from subsequent analyses (Jones et al., 2022). In this case, researchers who assume carelessness is randomly distributed throughout their sample may be inadvertently constraining their analytic sample to individuals with lower alcohol consumption. This has been described as 'tantamount to survey nonresponse' (Dunn et al., 2018) and has implications for the interpretation of data, which would appear unlikely to be missing at random. In addition, heavier drinkers are the population of interest in many alcohol studies and their exclusion can therefore impact the generalisability of study findings. In these instances, we reiterate calls from researchers to thoroughly and transparently examine both the causes of missing data, but also to discuss how the exclusion of this data might influence both descriptive and inferential analyses (Agley et al., 2022; Curran, 2016; Jones et al., 2022). Indeed, future research could also examine whether the inclusion or removal of careless responders has amplified or attenuated reported effects in previous studies.

Strengths and weaknesses

Strengths of these analyses are that we included data from several different studies using different sampling techniques from different online platforms and different countries (see table 1). We had high statistical power (>95%) to detect the effects, suggesting these findings are robust. For our confirmatory model we had a similar percentage of carelessness (14.3%) to our recent meta-analysis which included 48 studies and more than 75,000 participants (11.7% [95% CI: 7.6% to 16.5%]: Jones et al., 2022). We examined multiple forms of carelessness, including individual question failures, long-string index and intra-individual response. However, as this was a secondary analysis, we were unable to determine the precise careless measures used and included in the individual studies (which were all 'attention check' failures). Given discrepancies between different methods, true careless responding may be difficult to diagnose and failure on a single item (a zero-tolerance approach) is perhaps too conservative (Kim et al., 2018). This would be true if carelessness was akin to a lapse in attention which may be momentary, rather than across the duration of a study. Furthermore, not all measures of carelessness are equal; it has been shown that some measures of careless responding perform better than others, and some measures may inappropriately categorise an individual as careless (Curran & Hauser, 2019). For example, statements such as '*All my friends say I would make a great poodle'* lead to high false positive rates of careless responding, as conscientious responders can provide rational answers ('*Friends say I share a dog-like personality*'). Researchers are now moving beyond the individual item(s) approach to more sophisticated approaches (e.g. latent profiles of carelessness across multiple methods: (Brühlmann, Petralito, Aeschbach, & Opwis, 2020)). Finally, data from our own laboratory made up a large proportion of the overall data. However, we made multiple attempts to obtain data from elsewhere to attempt to overcome this, with limited success (Wicherts, Borsboom, Kats, & Molenaar, 2006).

In conclusion, careless responding presents a significant challenge in online alcohol research. Here we have demonstrated that careless responses and heavier alcohol use are positively associated; however, the causal pathway remains unknown. Increased alcohol use may lead to more careless responding, but alternatively careless responding may bias estimates of alcohol use upwards. Regardless of this causal path, researchers should carefully consider how to measure carelessness and the ramifications of removing careless responders for the statistical inferences and the generalisability of their findings.

Acknowledgements: We thank the authors who shared their data for these analyses. We also thank an anonymous reviewer for helpful comments and improvements on our analysis code.

References

Agley, J., Xiao, Y., Nolan, R., & Golzarri-Arroyo, L. (2022). Quality control questions on Amazon's Mechanical Turk (MTurk): A randomized trial of impact on the USAUDIT, PHQ-9, and GAD-7. *Behavior research methods, 54*(2), 885-897. doi:10.3758/s13428-021-01665-8

Ashley, M., & Shaughnessy, K. (2021). Predicting insufficient effort responding: The relation between negative thoughts, emotions, and online survey responses. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, No Pagination Specified-No Pagination Specified. doi:10.1037/cbs0000308

Berry, K., Rana, R., Lockwood, A., Fletcher, L., & Pratt, D. (2019). Factors associated with inattentive responding in online survey research. *Personality and Individual Differences, 149*, 157-159. doi:https://doi.org/10.1016/j.paid.2019.05.043

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *J Pers Soc Psychol, 111*(2), 218-229. doi:10.1037/pspp0000085

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. doi:https://doi.org/10.1016/j.metip.2020.100022

Credé, M. (2010). Random Responding as a Threat to the Validity of Effect Size Estimates in Correlational Research. *Educational and Psychological Measurement, 70*(4), 596-612. doi:10.1177/0013164410366686

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology, 66*, 4-19. doi: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. doi:https://doi.org/10.1016/j.jrp.2019.103849

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual Response Variability as an Indicator of Insufficient Effort Responding: Comparison to Other Indicators and Relationships with Individual Differences. *Journal of Business and Psychology, 33*(1), 105-121. doi:10.1007/s10869-016-9479-0

Groh, D. R., Ferrari, J. R., & Jason, L. A. (2009). Self-reports of Substance Abusers: The Relation between Social Desirability and Social Network Variables. *J Groups Addict Recover, 4*(1-2), 51-61. doi:10.1080/15560350802712397

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of Detecting Insufficient Effort Responding: Comparisons and Practical Recommendations. *Educ Psychol Meas, 80*(2), 312-345. doi:10.1177/0013164419865316

Huang, J. L., & DeSimone, J. A. (2020). Insufficient effort responding as a potential confound between survey measures and objective tests. *Journal of Business and Psychology*, No Pagination Specified-No Pagination Specified. doi:10.1007/s10869-020-09707-2

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*(1), 103-129. doi:https://doi.org/10.1016/j.jrp.2004.09.009

Jones, A., Earnest, J., Adam, M., Clarke, R., Yates, J., & Pennington, C. R. (2022). Careless responding in crowdsourced alcohol research: A systematic review and meta-analysis of practices and prevalence. *Exp Clin Psychopharmacol*. doi:10.1037/pha0000546

Jones, A (2023). Data and analysis scripts for careless-responding mega-analysis. Retrieved from osf.io/49e5x 23-01-2023.

Kehoe, T., Gmel, G., Shield, K. D., Gmel, G., & Rehm, J. (2012). Determining the best population-level alcohol consumption model and its impact on estimates of alcohol-attributable harms. *Population Health Metrics, 10*(1), 6. doi:10.1186/1478-7954-10-6

Kim, D. S., McCabe, C. J., Yamasaki, B. L., Louie, K. A., & King, K. M. (2018). Detecting random responders with infrequency scales using an error-balancing threshold. *Behav Res Methods, 50*(5), 1960-1970. doi:10.3758/s13428-017-0964-9

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

Martins, J. S., Bartholow, B. D., Cooper, M. L., Von Gunten, C. D., & Wood, P. K. (2018). Associations between executive functioning, affect-regulation drinking motives, and alcohol use and problems. *Psychology of Addictive Behaviors, 32*(1), 16-28. doi:10.1037/adb0000324

Meyer, J. F., Faust, K. A., Faust, D., Baker, A. M., & Cook, N. E. (2013). Careless and Random Responding on Clinical and Research Measures in the Addictions: A Concerning Problem and Investigation of their Detection. *International Journal of Mental Health and Addiction, 11*(3), 292-306. doi:10.1007/s11469-012-9410-5

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. doi:10.1080/13645579.2020.1719618

Robinson, E., Smith, J., & Jones, A. (2022). 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. *Br J Health Psychol, 27*(1), 30-49. doi:10.1111/bjhp.12527

Schroeders, U., Schmidt, C., & Gnambs, T. (2021). Detecting Careless Responding in Survey Data Using Stochastic Gradient Boosting. *Educational and Psychological Measurement, 82*(1), 29-56. doi:10.1177/00131644211004708

Sommet, N., & Morselli, D. (2017). Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS. *International Review of Social Psychology, 30*, 203-218.

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

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *Am Psychol, 61*(7), 726-728. doi:10.1037/0003-066x.61.7.726

Table 1: Description of studies and data that were used within our analyses.

Study ID	Description	Sample	Measure(s) of carelessness	N# careless (% of sample)

Angus et al (2021)	Study examined whether framing of the research on a MTurk influenced self-reported problem drinking or gambling severity in participants.
	USA sample
		N = 1010 recruited
		Age Mean = 36.1 (11.6)
		M:F = 520:485
		Ethnicity = NA
		Education = 668 degree or above, 342 below degree
		AUDIT Mean = 8.4 (7.9)
			'Three attention check items consisted of simple probe questions (e.g., "To continue, select 'strongly agree'")	N = 89 (8.81%)

Baines et al (2020) exp 1	Study examined the relationship between cognitive processes and alcohol use, using an online convenience sample
	UK sample recruited via opportunity sampling
		N = 108 recruited
		Age Mean = 24.1 (8.5)
		M:F = 26:82
		Ethnicity = NA
		Education = NA
		AUDIT Mean = 10.4 (5.7)
			'If you are paying attention leave this question blank': with the answers No, Yes but not in the last year and Yes during the last year	N = 3 (2.78%)

Baines et al (2020) exp 2	Study examined the relationship between cognitive processes and alcohol use, using an online convenience sample
	UK sample recruited via opportunity sampling
		N = 116 recruited
		Age Mean = 22.00 (6.1)
		M:F = 53:63
		Ethnicity = NA
		Education = NA
		AUDIT Mean = 13.0 (6.2)
			'If you are paying attention leave this question blank': with the answers No, Yes but not in the last year and Yes during the last year	N = 3 (2.59%)

Blackwell et al (2020)	Study was a randomised controlled trial examining the impact of availability on alcoholic drink selection.
	UK sample, recruited via Prolific
		N = 812 recruited
		Age Mean = 37.9 (12.3)
		M:F = 607:533
		Ethnicity = NA
		Education = 757 degree or above, 390 below degree
		AUDIT Mean = 9.7 (5.4)
			'When was the last time you flew to Mars?' ('never'; 'a few days ago'; 'weeks ago'; 'months ago')'	N = 4 (0.49%)

Clarke et al (2020)	Study was a factorial experimental design examining the effect of warning labels on alcohol selection. UK sample recruited via Qualtrics	N = 6198 recruited Age Mean = 49.1 (15.5) M:F = 3131:3059	Inattentive participants were screened out via an attention check embedded in the study (those not answering 'never' to the question:	N = 174 (2.81%)

		Ethnicity = 5784 white, 414 other
		Education = 3056 degree or above, 3126 below degree
		AUDIT Mean = NA
			‘When did you last fly to Mars?’	

Copeland et al (2022)	Study examined the behavioural economic differences in heavy drinkers and people who have reduced their consumption.	N = 120 recruited
		Age mean = 36.56 (13.05)
		M:F = 60:59
		Ethnicity = 112 white, 8 other
		Education = 57 degree or above, 63 below degree
		AUDIT = 15.00 (6.83)
			‘8 Attention checks in total, including; This is an attention check question. Please select “Can’t Say True or False”, and “This is an attention check question. Please select “Monthly”	N = 14 (11.67%)

Davies et al (2022)	Study examined the framing of messages on alcohol labels (positive, negative, neutral) on drinking intentions.
	UK Sample recruited via a university course.
	Note – data from only ‘University 2’ is included.
		N = 302 recruited
		Age mean = 30.10 (15.80)
		M:F = 74:227
		Ethnicity = 277 white, 25 other
			‘Two attention check questions were included in the version of the survey implemented at University 2’	N = 26 (8.61%)

		Education = 88 degree or above, 214 below degree
		AUDIT = 10.56 (6.22)		

Jones et al (2020)	Study examined the prevalence of negative outcomes experienced following alcohol use.
	UK Sample recruited via university course credits and opportunity sampling.
		N = 299 recruited
		Age mean = 24.3 (10.7)
		M:F = 87:211
		Ethnicity = NA
		Education = NA
		AUDIT Mean = 11.3 (5.9)
			'To ensure you are paying attention leave this question blank' With four response options.	N = 9 (3.01%)

McPhee et al (2020)	Study examined the changes in alcohol use and outcomes after the introduction of COVID-19 social distancing.
	US sample recruited via MTURK
		N = 1127 recruited
		Age Mean = 40.2 (10.3)
		M:F = 739:381
		Ethnicity = 737 white, 386 other
		Education = NA
		AUDIT Mean = 13.3 (9.4)
			'Five attention-check questions were interspersed throughout the survey as a means of detecting random responding.'	N = 481 (42.68%)

Robinson et al (2020)	Study examined lifestyle related changes following the introduction of COVID-19 lockdowns.
	UK Sample recruited via Prolific
		N = 902 recruited
		Age Mean = 30.6 (9.7)
		M:F = 296:587
		Ethnicity = 705 white, 176 other
		Education = 558 degree or above, 322 below degree
		AUDIT = NA
			‘Two attention checks were included in the survey (e.g., ‘have you ever been to the planet Mars?’)’	N = 33 (3.66%)

Strickland et al (2019)	Study tested the feasibility and acceptability of delivering cognitive training interventions via crowdsourcing.
	US sample recruited via MTURK
		N = 476 recruited
		Age Mean = 34.1 (9.8)
		M:F = 236:240
		Ethnicity = 370 white, 106 other
		Education = 244 degree or above, 232 below degree
		AUDIT Mean = 12.6 (7.3)
			One or more attention checks were used throughout the study.	N = 32 (6.72%)

Strickland et al (2019b)	Study examined the predictive relationship between behavioural economic demand,	N = 307 recruited Age Mean = 35.5 (10.7)	One or more attention checks were used throughout the study.	N = 30 (9.77%)

	delay discounting and alcohol reinforcement and alcohol use.
	US Sample recruited via MRUTK
		M:F = 136:171
		Ethnicity = 251 white, 56 other
		Education = 155 degree or above, 152 below degree
		AUDIT Mean = 10.4 (7.8)		

Strickland et al (2020)	Study aimed to examine the association between concurrent choice tasks and alcohol.
	US Sample recruited via MTURK
		N = 125 recruited
		Age Mean = 34.8 (11.1)
		M:F = 64:61
		Ethnicity = 94 white, 31 other
		Education = 73 degree or above, 52 below degree.
		AUDIT Mean = 6.5 (6.0)
			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 participants were instructed to remember, and 4) an item that asked if participants had been attentive and that their data should be used	N = 17 (13.60%)

Table 2: Multilevel binomial regression models examining the association between sociodemographic characteristics, alcohol use and careless responding.

Variables	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
	OR [95% CI]	OR [95% CI]	OR [95% CI]	OR [95% CI]	OR [95% CI]	OR [95% CI]
Age	0.969 [0.962, 0.975]**	0.959 [0.950, 0.968]**	0.991 [0.982, 1.000]	0.989 [0.980, 0.998]*	0.989 [0.980, 0.998]*	0.988 [0.979, 0.997]**
Gender	0.674 [0.575, 0.789]**	0.611 [0.485, 0.769]**	0.952 [0.780, 1.162]	0.895 [0.735, 1.088]	0.894 [0.734, 1.089]	0.879 [0.722, 1.070]
Ethnicity		0.996 [0.721, 1.377]	-	-	-	-
Education		1.376 [1.088, 1.738]**	-	-	-	-
AUDIT total			1.071 [1.060, 1.083]**	-	-	-
AUDIT hazard				2.211 [1.807, 2.713]**	-	-
AUDIT harmful					3.434 [2.829, 4.172]**	-
AUDIT prob dep						3.633 [2.946, 4.481]**
Model Information						
N total / Study Total	11756 / 13	8353 / 7	4660 / 11	4660 / 11	4660 / 11	4660 / 11
AIC	4902	2569	2908	3017	2914	2923
ICC	0.29	0.06	0.30	0.32	0.30	0.29
Marginal R^2	.056	.132	.063	.040	.065	.052
Careless N (%)	909 (7.7%)	324 (3.9%)	678 (14.3%)	678 (14.3%)	678 (14.3%)	678 (14.3%)

*Legend: AUDIT = Alcohol Use Disorders Identification Task; prob dep = Probable Dependence ;*p< .05; **p< .01*