Published in final edited form as:

Drug Alcohol Depend. 2018 February 01; 183: 102–110. doi:10.1016/j.drugalcdep.2017.10.033.

Random responses inflate statistical estimates in heavily skewed addictions data*

Kevin M. King^a, Dale S. Kim^b, and Connor J. McCabe^a

^aUniversity of Washington, Seattle, WA, 98195

^bUniversity of California, Los Angeles, Los Angeles, CA 90095

Abstract

Background—Some respondents may respond at random to self-report surveys, rather than responding conscientiously (Meade et al., 2007; Meade and Craig, 2012), and this has only recently come to the attention of researchers in the addictions field (Godinho et al., 2016). Almost no research in the published addictions literature has reported screening for random responses. We illustrate how random responses can bias statistical estimates using simulated and real data, and how this is especially problematic in skewed data, as is common with substance use outcomes.

Method—We first tested the effects of varying amounts and types of random responses on covariance-based statistical estimates in distributions with varying amounts of skew. We replicated these findings in correlations from a real dataset (Add Health) by replacing varying amounts of real data with simulated random responses.

Results—Skew and the proportion of random responses influenced the amount and direction of bias. When the data were not skewed, uniformly random responses deflated estimates, while long-string random responses inflated estimates. As the distributions became more skewed, all types of random responses began to inflate estimates, even at very small proportions. We observed similar effects in the Add Health data.

Conclusions—Failing to screen for random responses in survey data produces biased statistical estimates, and data with only 2.5% random responses can inflate covariance-based estimates (i.e., correlations, Cronbach's alpha, regression coefficients, factor loadings, etc.) when data are heavily

*Supplementary material can be found by accessing the online version of this paper at <http://dx.doi.org> and by entering doi:...

Correspondence: Kevin M. King, Department of Psychology, University of Washington, Seattle, WA, 98195, Box 351525, kingkm@uw.edu.

Contributors

Kevin M. King, University of Washington. Dale S. Kim, University of California, Los Angeles. Connor J. McCabe, University of Washington. Kevin King generated the idea, wrote the majority of the manuscript and conducted the analyses. Dale Kim aided with the simulations and provided critical edits to the manuscript. Connor McCabe provided critical edits to the manuscript. All authors have approved the final article.

Conflict of Interest

No conflict declared.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

skewed. Screening for random responses can substantially improve data quality, reliability and validity.

Keywords

statistical bias; online surveys; invalid responses; skewed data

1. Introduction

Attending to sources of bias in self-report surveys of substance use is important for obtaining accurate estimates of behavior (Harrison and Hughes, 1997; Johnson and Fendrich, 2005). One important source of bias in self-reports is invalid survey responses, which are responses that do not accurately reflect a participants' true level of a given survey item. To date, most research on invalid survey responses in the addictions literature has focused on intentionally distorted responses, such as misreporting due to social desirability (Crowne and Marlowe, 1960), faking (Nichols et al., 1989), mischievous responding (Robinson-Cimpian, 2014), or misrepresentation (Johnson, 2005). Efforts to detect intentional distortions in substance use survey research have ranged from directly measuring participants' social desirability (Nederhof, 1985), using external stimuli to promote truthful responding, such as the bogus pipeline technique (Roese and Jamieson, 1993), or including fake drug names to detect over-reporting of drug use (Pape and Storvoll, 2006). A second type of invalid response is termed random responding, in which participants respond to survey items without regard to their content (Nichols et al., 1989). Random responding (more accurately, pseudo-random responding, Mosimann et al., 1995) reflects a response pattern where all responses have a roughly equal chance of being chosen in the long run. These response patterns could be generated by participants choosing responses essentially at random or "long-string responding" where a participant randomly chooses a response which is then repeated across a large number of items (Johnson, 2005). Figure 1 illustrates examples of these invalid responses, compared to valid responses.

It is important to note that invalid responses of either kind cannot be detected with certainty by merely inspecting response patterns (Meade and Craig, 2012). For example, some individuals may genuinely respond with a single response (such as all "Agree") across a set of items because it reflects their true level of a construct. On the other hand, invalid responders may provide identical responses because they are attempting to move quickly through a survey. Other kinds of over or under-reporting are similarly impossible to detect without a secondary indicator of the individual's propensity for inaccurate reporting, such as a social desirability scale (Crowne and Marlowe, 1960). The goal of this study is to illustrate how random responding might affect statistical procedures and the inferences researchers make.

Invalid responses generally, and random responding specifically, are a potential problem for all self-report surveys, regardless of mode of administration. Although some research has suggested in-person versus online data collection methods produce largely similar estimates for certain health behavior and personality measures (Hewson and Charlton, 2005; Lara et al., 2012; Vosylis et al., 2012), the absence of direct researcher interaction and oversight for

mass in-person or online surveys may increase the chances of random responding, relative to interview methods (Godinho et al., 2016; Meyer et al., 2013). This may be due to participants having lower incentives to carefully consider survey content when they are not directly interacting with an interviewer. One large study reported that rates of random responding tripled in online respondents relative to paper and pencil surveys (Johnson, 2005). Online surveys provide a cost- and time-effective method to accrue large samples (Wright, 2005) that can also better represent the broader population than typical convenience samples used in psychological research (Henrich et al., 2010), and as such, have been increasingly adopted by researchers (Gosling and Mason, 2015). Thus, it is important that methodological attention is drawn to issues of data quality.

There are several important reviews on the types and detection of random responding (Curran, 2016; Meade and Craig, 2012). Formal detection methods (such as reverse coded items, infrequency items, or “bogus flag items”) are important because data-based approaches (such as searching for long strings of constant responses) have been shown to be poor predictors of invalid responding (Meade and Craig, 2012).

Random responding has received considerably less attention in addiction research (Godinho et al., 2016; Meyer et al., 2013), and it is unclear if researchers in the addictions field are screening for random responding. For example, only four manuscripts published in addiction journals referenced the four most highly cited manuscripts on random responding (Huang et al., 2012; Johnson, 2005; Maniaci and Rogge, 2014; Meade and Craig, 2012). Two of these manuscripts described the issue of random responding (Godinho et al., 2016; Meyer et al., 2013), two others reported screening for random responding (Blevins and Stephens, 2016; Hershberger et al., 2016), and a fourth studied the issue directly (Wardell et al., 2014). To confirm this, we conducted a broader literature search in the full text of every research article published in 2016 in 14 journals (randomly selected from among all journals that primarily publish research on addictions: *Addiction*, *Addiction Biology*, *Addiction Research and Theory*, *Addictive Behaviors*, *Alcohol and Alcoholism*, *Alcoholism: Clinical and Experimental Research*, *Drug and Alcohol Dependence*, *Drugs: Education, Prevention and Policy*, *Journal of Child and Adolescent Substance Abuse*, *Journal of Drug Education*, *Journal of Studies on Alcohol and Drugs*, *Nicotine and Tobacco Research*, *Psychology of Addictive Behaviors*, *Substance Use and Misuse*) for keywords related to random responding (*invalid*, *invalid response*, *random respond*, *random response*, *bogus*, *mischievous*, and *long-string*). Only 11 of 2,079 (0.0053%) research studies reported any method of screening for random responding when the data were collected online, while 9 studies (0.0043%) reported screening for random responses with data collected in person (all large classroom surveys). Most studies that did screen for random responding used an “eyeball” method, where data were visually inspected for constant or scattered responding, and only two studies reported using formal methods (such as invalid response screening items) for detecting random responding (Buckner et al., 2015; Lindgren et al., 2015). Although it may be that some studies do screen for random responding but do not report it, this suggests that formal screening methods for random responding are extraordinarily rare, regardless of the mode of administration. Thus, there is little appreciation of the degree to which findings in addiction research may be vulnerable to issues of random responding.

If these respondents only introduced random error, (e.g., only reducing associations between variables), the solution would be to merely increase power by collecting more subjects, and relatively small proportions of random responses might be ignorable. Most researchers will be familiar with the correction for attenuation due to measurement error, which inflates correlations based on the (un)reliability of the measured variables. It is thus natural to assume that when random responses are introduced into a data set, that an observed correlation will always decline because random responses decrease the reliability of the observations. Indeed, studies have reported that when random respondents comprised as little as 10% of a sample, statistical power was reduced (Maniaci and Rogge, 2014), experimental effects were attenuated (Oppenheimer et al., 2009), correlations were reduced between psychopathology constructs (Fervaha and Remington, 2013) and caused one factor models to be more likely to be rejected when the proportion of random responses was greater than 10% (Woods, 2006).

However, measurement error can sometimes inflate correlations when the sources of measurement error (such as reporter bias) are correlated across variables (Schmitt, 1996). Similarly, we hypothesize that random responders do not always reduce correlations; they may sometimes inflate correlations and other covariance-based statistics (e.g., factor loadings, regression coefficients, etc.). We also hypothesize that this will be especially likely in substance use data, where response variables (such as the number of drinks, or number of substance use disorder symptoms) are often skewed and zero inflated (Atkins et al., 2012).

Prior studies support this assertion. Credé (Credé, 2010) and Huang, Liu and Bowling (2015b) presented both simulated and real data with various amounts of random observations, and showed that the presence of random responses could lead to *either* inflated or deflated correlations across scales. These studies attributed these findings to a difference of means between the valid and random responses to a measurement scale. If the valid responses on two measures both had high (or low) means, random responses would inflate the correlation between the substantive variables.

The reason for this selective inflation/deflation is based on the law of total covariance, which describes the decomposition of the covariance between any two variables conditional on a third. The covariance between any two variables can be expressed in Equation 1 (Cohen et al., 2003):

$$Cov(X, Y) = E[XY] - E[X]E[Y]. \quad (1)$$

When a third variable is involved, such as whether a response is valid or invalid, this covariance formula can be conditioned on a third variable as in Equation 2:

$$Cov(X, Y) = E[Cov(X, Y|Z)] + Cov(E[X|Z], E[Y|Z]). \quad (2)$$

This provides a way to understand how the covariance between two variables (X and Y) may be influenced by mean differences across a third variable (Z). For example, Z could indicate

a mixture of group membership, such as treatment and control groups, or conscientious and random responders. The first term of this decomposition, $E[Cov(X, Y|Z)]$, refers to the covariances between X and Y averaged across each component of the mixture (Z). The second term, $Cov(E[X|Z], E[Y|Z])$, states that the covariance between all the pairwise expectations of the mixtures (i.e., the bivariate mean of X and Y for each Z) is also summed into the total covariance. This formula shows that not only is the covariance of X and Y influenced by the covariances within the groups, it is influenced by the *mean* of both X and Y between those groups as well.

This “problem of the means” is important in the context of random responding, as it shows that when unobserved subgroups in a population have substantially different means, the covariance *between* two variables in that population will be influenced regardless of the covariances *within* those groups. Thus, variables with particularly strong floor or ceiling effects will be especially susceptible to even small percentages of random responding in a sample. Since random responses may be expected to have a different mean than the true responses, the covariance may be expected to increase, rather than decrease, with their inclusion.

Thus, the goal of the current study was to demonstrate how random responding, reflected by both uniform (i.e. fully random) and long-string responding (i.e. randomly choosing a response that is then repeated across all items), introduce bias into covariance-based statistical estimates, and to illustrate how these will be particularly influential in distributions typical to the addictions field. Because covariances are the basis of most inferential statistics used in psychology, we hypothesized that random responding would bias estimates of most covariance-based statistics, including those used to assess measurement (such as average inter-item correlations, coefficient alpha, and factor loadings), and associative statistics (such as regression coefficients and correlations). We tested our hypotheses first in simulated data, then we replicated our findings using data from a nationally representative epidemiological survey.

2. Study 1

2.1. Method

We designed a simulation study where covariance-based statistical estimates (described below) were estimated from valid responses simulated from a pre-specified true distribution that were mixed with simulated invalid responses of varying proportions and types. We tested the effects of three factors: (1) the distributional shape of the valid responses, (2) the proportion of uniform responses, and (3) the proportion of long-string responses. We chose five distributional shapes for the valid responses to reflect varying degrees of floor and ceiling effects found in real data (see Figure 2). The proportion of both uniform and long-string responses ranged from 0% to 30%, increasing in 2.5% increments. This resulted in a $5 \times 13 \times 13$ design for a total of 845 conditions.

For each condition, we generated one dataset consisting of $p = 6$ variables (X_1 through X_6) with $n = 10,000$ responses. In every condition, the covariance among the six variables were fixed for the valid responses, with covariances set to 0.36 and variances set to 1. The

uniform responses were generated from a multivariate uniform (i.e., flat) distribution with range parameters of -3 to 3 , with covariances fixed to zero among the variables. The long-string responses were generated with the same parameters as the uniform responses, with the caveat that one value was drawn per observation, which was then repeated for each of the six variables. All distributions were continuous.

For each dataset, we computed five statistical estimates (their expected values for valid responses are listed in parentheses). These were the correlation coefficient between X_1 and X_2 ($r = 0.36$), the linear regression slope coefficient¹ by regressing X_1 onto X_2 ($\beta = 0.36$), the coefficient alpha among the six variables ($\alpha = 0.77$), the mean inter-item correlation among the six variables ($\bar{r} = 0.36$) and the factor loading of X_1 from fitting unidimensional factor model among the six variables ($\lambda = 0.6$). The regression model was estimated via ordinary least squares (OLS) and the factor model was estimated with maximum likelihood using a standardized latent variable. The final analytic dataset contained 5 statistical estimates derived from each of the 845 datasets for each condition, which varied across the 3 experimental factors.

2.1.1. Analytic Plan—We used an OLS regression model to analyze the effect of the three experimental factors (distribution of the valid responses, and the proportion of uniform and long-string responses) on the statistical estimates. The regression coefficients represented the empirical bias produced by each factor. For example, a β coefficient of -0.05 for the effect of uniform responses on correlation coefficients would indicate that for every 10% increase in uniform responses in a dataset, the correlation coefficient would be underestimated by 0.05. Because these factors may appear in various combinations in real data, we estimated three-way interactions with all associated two-way interactions and main effects for all models. There were five analytic models, one with each statistical estimate as an outcome.

2.2. Results

2.2.1. Correlation Coefficient—Figure 3 illustrates these findings, while Table 1 provides the main effects. There was a three-way interaction between the distribution of true responses and the proportion of random and long-string responses, and all lower level two-way interactions were significant. Therefore, we estimated simple main effects for each predictor by fixing the other predictors to values of interest.

When the distribution was mean-centered and had no skew, correlations were reduced by 0.071, on average, for every 10% increase in uniform responses when long-string responses were not present in the data. Conversely, a 10% additive increase in long-string responses (in the absence of uniform responses) was related to a linear increase in the correlation of 0.119.

As the data became more skewed, these effects changed. For uniform responses (in the absence of long-string responses), the main effect on correlations turned positive as distributions became more skewed, ranging from 0.007 to 0.011 in the most negatively and

¹Regression intercepts were estimated in these models, but not reported. This is because invalid responders affect regression intercepts mostly through raw means, rather than the covariances. Any effect that covariance would have on the intercept occurs via through the slope coefficient, which is already reported and analyzed. As such, the regression intercept analysis is not reported for parsimony.

positively skewed distributions, respectively. For long-string responses (in the absence of uniform responses), skewness increased the observed inflation in the correlations, such that at the most negative and positively skewed distributions, a 10% increase in long-string responses produced a 0.133 and 0.138 linear increase in correlations, respectively.

There was also an interaction between the proportion of uniform and long-string responses and skew (Figure 3). In general, as there were more uniform responses in a dataset, the more correlations were under-estimated, and as skew increased in the data, including more uniform responses produced greater decreases in correlations when more long-string responders were present in the data.

2.2.2. Generalizing to Other Correlation Based Statistics—The effects of uniform and long-string responses and their interaction on correlations was nearly identical to that of all other correlation based statistics: regression slopes, mean inter-item correlations, coefficient alpha, and factor loadings (see supplementary table S1²). The one exception was coefficient alpha, where long-string responses inflated this statistic more so with any amount of skew, but no observable trend was detected.

3. Study 2

3.1. Method

We used the publicly available Add Health dataset (Harris and Udry, 2008) to demonstrate the effects of invalid responses on correlations in real data. We used data from the Wave 1 in-home interview, which were collected using interviewer administered questionnaires. We assumed that both the probability of random responding was very low and the coefficient estimates in the data were as close as possible to their true parameter values. Thus, our simulations reflect the counterfactual scenario where the Wave 1 Add Health data were collected in an alternative administration mode that increased the proportion of invalid respondents (such as online, or large-scale paper and pencil surveys).

We selected five variables that are common in alcohol research with adolescents. Two items were related to the maternal-child relationship, maternal closeness and maternal care (Add Health items H1WP9 and H1WP10), with 5 response options ranging from 1 “Not at all” to 5 “Very Much”. These items were moderately skewed (range -1.84 to -4.17), and moderately to very kurtotic (range 3.06 to 20.35), with median scores of 5 on a scale from 1 to 5. Three items were related to alcohol use – past year frequency of drinking, past year frequency of getting drunk, and past year frequency of drinking 5 drinks or more (Add Health items H1TO15, H1TO17, and H1TO18) – with response options ranging from 1 “Every day/almost every day” to 7 “Never”. The alcohol use items were moderately skewed (range -1.35 to -2.21), and kurtotic (range 0.93 to 4.32), with median scores of 7 on a scale from 1 to 7, indicating a large majority of the sample reported no alcohol use.

3.1.1. Analytic Plan—We first computed the raw correlation between each maternal-child relationship variable and each alcohol use variable. Then, we conducted bootstrap simulation

²Supplementary material can be found by accessing the online version of this paper at <http://dx.doi.org> and by entering doi:...

analyses where we randomly selected varying proportions of the data (2.5%, 5%, 10%, 20% and 30%) and replaced those valid data proportions with random responses. For the simulated random uniform data, we followed the methods outlined above to generate a random item response on the same scale as the original item. For the random long-string responses, we randomly generated a response to the first maternal warmth item (which had a 1 to 5 scale) and copied it to all remaining items. We then re-estimated the correlations. This procedure was repeated 10,000 times for each of the 5 proportions of long-string and uniform responders. Because we assumed that the simulation results from Study 1 would hold across other parameter estimates and conditions, and because other studies have reported that the prevalence of long-string respondents tends to be low (Huang et al., 2012), we focused on the main effects of uniform and long-string responders on estimates of correlations.

3.2. Results

Table 2 presents the original and contaminated correlations observed in the data. As we expected given the results of Study 1, with the inclusion of even 2.5% of uniform or long-string data, correlations between all variables modeled were inflated beyond their original values. This inflation increased as the proportion of invalid responses increased, although the effect was much more dramatic for long-string responses than for uniform responses. The magnitude of this inflation was large: when only 2.5% of the responses were replaced with invalid responses, the correlation estimate was over-estimated by 0.12 to 0.26 for long-string responses, and by 0.06 to 0.17 for uniform responses. Figure 4 illustrates this effect, showing how a very small number of invalid cases can skew the correlation between two variables.

4. Discussion

Our findings suggest that even small proportions of random responses bias all forms of covariance-based statistics, and that the types of highly skewed data collected in addiction research are particularly vulnerable to even small amounts of these kinds of data. Importantly, the direction of the bias depended on the proportion of uniform versus long-string responses and the amount of skewness in the data. Although our findings are agnostic to administration mode, we believe that random responses are more likely when there are fewer incentives for attentive or careful responding. Almost no studies in the addiction literature reported screening for random respondents, regardless of administration mode. This is particularly problematic now as access to the internet has become nearly ubiquitous (Lenhart, 2015) and online surveys are adopted in psychology (Casler et al., 2013), and thus represent an increasing proportion of research studies.

In highly skewed data, the effects of random responses were especially pernicious. Even uniform responses, typically thought to deflate associations between variables, began to inflate associations between variables when the data were highly skewed, although their effects were somewhat offset when there were also long-string responses. The consequences of ignoring random respondents, or assuming that they universally decrease associations and produce conservative estimates of effects, may be more serious than previously understood. In both our most heavily skewed simulated data, as well as our sample of real

epidemiological data, the inclusion of just 2.5% of uniform responses inflated correlation based statistical estimates, sometimes dramatically.

Inflating parameters increases the Type I error rate, making the risk of false positives higher, especially when the samples are larger or parameter inflation is larger. Although we observed bias in correlations of around .12 when the proportion of long-string respondents was only 10% in our simulated data, we observed inflation of .40 to .50 in the Add Health data (which were more heavily skewed than our simulated data) for the same proportion of long-string respondents. Figure 5 illustrates how parameter bias and sample size interact to inflate the risk of false positives: as either sample sizes or parameter bias increases, the false positive rate increases, often dramatically. Given how much research in addictions is conducted with heavily zero-inflated and skewed outcomes, and the apparent lack of attention to these issues, it is critical to increase attention to this issue among addiction researchers.

4.1. Generalizing to Other Analytic Approaches

Because random responses influence the estimation of covariance, the bias introduced by random responses generalizes to all covariance-based analyses. Most common statistical practices cannot account or correct for the bias in parameter estimates introduced by invalid responses if those responses are not identifiable in the data (such as with infrequency scales). Outliers, for example, are more likely to represent random responses because they tend to have large deviations from sufficient statistics (such as the mean), but classifying any single case as a random response versus an outlier with certainty is impossible without additional information. Researchers often attempt to transform data for OLS to reduce heteroskedasticity of the residuals, however, this will not change the bias in the regression parameter estimates, which are known to be unbiased otherwise (Cohen et al., 2003). We have also shown that this problem extends to methods that are commonly estimated with maximum-likelihood, such as factor analysis, and by extension, structural equation modeling. Again, because random responses bias the covariance matrix, which is often the starting point for maximum likelihood estimation, bias in the covariance matrix will simply be perpetuated to the parameters. Because robust estimators for maximum likelihood in most software packages typically correct only standard errors or fit statistics (Satorra and Bentler, 2000; Savalei, 2014), they will not correct the bias introduced by random responses. The effects of random responses will also be pernicious when data are missing; because covariances are the core of regression based imputation, such as multiple imputation (Graham, 2009), studies that utilize regression imputation approaches without accounting for random responses will likely misestimate imputed values in their data.

Moreover, because the biasing effects were caused by mean differences between the random and valid responses, our findings will also be expected to hold for other forms of invalid responses. Mischievous responding, under- or over-reporting, “faking good”, or “faking bad”, will all produce response distributions that bias coefficients away from their true values, if they have a mean response that differs from the mean of valid responses. This bias will generalize to all covariance-based coefficients that we illustrated in the current study. Thus, researchers should work to not only reduce and detect random responding in survey

data, but to reduce response bias in self-report data in general to obtain more accurate coefficient estimates.

4.2. Recommendations

Unfortunately, although there is much excellent work on detecting careless respondents in surveys (Credé, 2010; Curran, 2016; Huang et al., 2015a; Meade and Craig, 2012), there are few concrete recommendations for precise and reliable screening methods and decision rules to detect and eliminate invalid responses. First, prior authors have recommended using active screening methods that have clearly correct responses, such as forced choice items (such as “Respond to the following item with “Strongly Agree””; (Meade and Craig, 2012) rather than bogus flag items (i.e., “unusual” items), as these items are likely to be less confusing to participants and thus serve to better differentiate conscientious from careless respondents. However, few recommendations for detection rules have been provided. Recent research (Kim et al., 2017) has shown that, using these screening methods, zero-tolerance cutoffs (where respondents are eliminated if they failed even a single screening item) performed very poorly and were likely to screen out far too many respondents who were responding conscientiously. We have demonstrated that an error-balancing cutoff technique that balances the rates of true and false positives as a function of the probability of a subject randomly failing a screening item, outperforms zero-tolerance approaches (Kim et al., 2017). However, there is much more work to be done to help researchers identify and screen invalid respondents in surveys. For example, dynamic online surveys might be able to use reminders after failed attention checks to help respondents “get back on track” or to boost, motivation and attention to surveys. Some research has suggested that design features of surveys can improve subjects’ attention to survey items and improve data quality (Kapelner and Chandler, 2010). Alternatively, estimation methods might be developed that could down-weight respondents according to an index of invalidity.

4.3. Limitations and Future Directions

The current study has multiple strengths, including the simulation of multiple combinations of types of invalid respondents combined with multiple conditions of skew, which are likely to be found in addictions data, and validation of our findings in real data. However, there are several limitations that should be considered. Perhaps most importantly, our models were idealized scenarios that made certain assumptions about the nature of invalid respondents: that their responses are trait-like, and thus will affect all items in the same way. First, humans generally do not produce sequences of random numbers even when they try to (Mosimann et al., 1995); thus these response generation processes may best be considered as “pseudo-random” rather than truly random. Second, long-string responses may reflect two processes: one where a response is chosen pseudo-randomly and then repeated, and another (terms “mischievous responding” or “response distortions”) where participants intentionally choose extreme responses. Although we modeled the former, the latter process is also found in surveys (Robinson-Cimpian, 2014) and will even more greatly exaggerate the effects we illustrate in the current study because the mean differences between the conscientious and mischievous responses will on average be larger.

We also did not test our models against real data where respondents were either induced to respond invalidly (i.e. Downes-Le Guin et al., 2012) or where we had a method of screening for invalid responding. It may be that “real” inattentive respondents respond less randomly than truly “random” data, and may have a different effect on correlations than we estimated. Much substance use data reflects counts of behaviors (such as the number of dependence symptoms endorsed, or the number of past-month heavy drinking episodes), and models for count data are increasingly common in the substance use literature (Atkins et al., 2012). However, analysis of count data rely on maximum likelihood estimation in the generalized linear model, which relies on a non-linear transformation of covariance-based statistics to compute parameter estimates (Hilbe, 2011). As such, although it is likely that invalid responses will bias count data, it is unclear to what degree the effects will be similar. Moreover, it is likely that mischievous respondents only selectively respond to items that trigger extreme responses, and may respond to other items attentively (Robinson-Cimpian, 2014). Thus, future research should consider the effect on covariances when invalid responding behaves more as a state-like variable, rather than a trait (as is usually assumed). Finally, incentives such as money, participation appreciation, or perceived importance of participation may reduce or eliminate invalid responding, but this has yet to be tested to our knowledge.

There is increasing attention to improve the quality of research findings across domains of psychology and other sciences (Ioannidis, 2014). This has encompassed using more open access research methods and protocols, pre-registering data analyses, avoiding methods that may inflate Type I error, as well as emphasizing estimation of effect sizes and confidence intervals rather than p-values (Asendorpf et al., 2013; Bakker et al., 2012; Cumming, 2014; Miguel et al., 2014; Nosek et al., 2015; Patil et al., 2016). However, one method by which all data analysis may be improved is by improving the quality of the data before analyses have even begun. Improving data collection practices to ensure high quality data collection may be one important step towards increasing replicability and validity across the addictions field.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Role of Funding Source

This research was partially supported by a grant from the National Institute on Drug Abuse (DA040376) to Connor J. McCabe. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The authors declare no conflicts of interest.

References

- Asendorpf JB, Conner M, De Fruyt F, De Houwer J, Denissen JJA, Fiedler K, Fiedler S, Funder DC, Kliegl R, Nosek BA, Perugini M, Roberts BW, Schmitt M, Van Aken MAG, Weber H, Wicherts JM. Recommendations for increasing replicability in psychology. *Eur J Pers.* 2013; 27:108–119. DOI: 10.1002/per.1919

- Atkins DC, Baldwin Sa, Zheng C, Gallop RJ, Neighbors C. A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychol Addict Behav.* 2012; 27:379.doi: 10.1037/a0029508
- Bakker M, van Dijk a, Wicherts JM. The rules of the game called psychological science. *Perspect Psychol Sci.* 2012; 7:543–554. DOI: 10.1177/1745691612459060 [PubMed: 26168111]
- Blevins CE, Stephens RS. The impact of motives-related feedback on drinking to cope among college students. *Addict Behav.* 2016; 58:68–73. DOI: 10.1016/j.addbeh.2016.02.024 [PubMed: 26914263]
- Buckner JD, Henslee AM, Jeffries ER. Event-specific cannabis use and use-related impairment: The relationship to campus traditions. *J Stud Alcohol Drugs.* 2015; 76:190–194. doi:<https://doi.org/10.15288/jsad.2015.76.190>. [PubMed: 25785793]
- Casler K, Bickel L, Hackett E. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Comput Human Behav.* 2013; 29:2156–2160. DOI: 10.1016/j.chb.2013.05.009
- Cohen, J., Cohen, P., West, SG., Aiken, L. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* 1. Lawrence Erlbaum Associates; Hillsdale NJ: 2003.
- Credé M. Random responding as a threat to the validity of effect size estimates in correlational research. *Educ Psychol Meas.* 2010; 70:596–612. DOI: 10.1177/0013164410366686
- Crowne DP, Marlowe D. A new scale of social desirability independent of psychopathology. *J Consult Psychol.* 1960; 24:349–354. DOI: 10.1037/h0047358 [PubMed: 13813058]
- Cumming G. The new statistics: Why and how. *Psychol Sci.* 2014; 25:7–29. DOI: 10.1177/0956797613504966 [PubMed: 24220629]
- Curran PG. Methods for the detection of carelessly invalid responses in survey data. *J Exp Soc Psychol.* 2016; 66:4–19. DOI: 10.1016/j.jesp.2015.07.006
- Downes-Le Guin T, Baker R, Mechling J, Ruylea E. Myths and realities of respondent engagement in online surveys. *Int J Mark Res.* 2012; 54:1–21. DOI: 10.2501/IJMR-54-5-000-000
- Fervaha G, Remington G. Invalid responding in questionnaire-based research: Implications for the study of schizotypy. *Psychol Assess.* 2013; 25:1355–1360. DOI: 10.1037/a0033520 [PubMed: 23815113]
- Godinho A, Kushnir V, Cunningham JA. Unfaithful findings: Identifying careless responding in addictions research. *Addiction.* 2016; 111:955–956. DOI: 10.1111/add.13221 [PubMed: 26662631]
- Gosling SD, Mason W. Internet research in psychology. *Annu Rev Psychol.* 2015; 66:877–902. DOI: 10.1146/annurevpsych-010814-015321 [PubMed: 25251483]
- Graham JW. Missing data analysis: Making it work in the real world. *Annu Rev Psychol.* 2009; 60:549–76. DOI: 10.1146/annurev.psych.58.110405.085530 [PubMed: 18652544]
- Harris KM, Udry JR. National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994–2008 [Public Use] 1994–2008. 2008; doi: 10.3886/ICPSR21600.v15
- Harrison, L., Hughes, A. NIDA research monograph. National Institutes of Health; Rockville, MD: 1997. The validity of self-reported drug use: improving the accuracy of survey estimates.
- Henrich J, Heine SJ, Norenzayan A. The weirdest people in the world? *Behav Brain Sci.* 2010; 33:61–83-135.doi: 10.1017/S0140525X0999152X [PubMed: 20550733]
- Hershberger AR, VanderVeen JD, Karyadi KA, Cyders MA. Transitioning from cigarettes to electronic cigarettes increases alcohol consumption. *Subst Use Misuse.* 2016; 51:1838–1845. DOI: 10.1080/10826084.2016.1197940 [PubMed: 27653988]
- Hewson C, Charlton JP. Measuring health beliefs on the Internet: A comparison of paper and Internet administrations of the Multidimensional Health Locus of Control Scale. *Behav Res Methods.* 2005; 34:691–702. DOI: 10.3758/BF03192742
- Hilbe, JM. *Negative Binomial Regression.* Cambridge University Press; Cambridge, UK: 2011.
- Huang JL, Bowling NA, Liu M, Li Y. Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *J Bus Psychol.* 2015a; 30:299–311. DOI: 10.1007/s10869-014-9357-6
- Huang JL, Curran PG, Keeney J, Poposki EM, DeShon RP. Detecting and deterring insufficient effort responding to surveys. *J Bus Psychol.* 2012; 27:99–114. DOI: 10.1007/s10869-011-9231-8

- Huang JL, Liu M, Bowling NA. Insufficient effort responding: Examining an insidious confound in survey data. *J Appl Psychol*. 2015b; 100:828–845. DOI: 10.1037/a0038510 [PubMed: 25495093]
- Ioannidis JPA. How to make more published research true. *PLoS Med*. 2014; :11.doi: 10.1371/journal.pmed.1001747
- Johnson JA. Ascertaining the validity of individual protocols from Web-based personality inventories. *J Res Pers*. 2005; 39:103–129. DOI: 10.1016/j.jrp.2004.09.009
- Johnson T, Fendrich M. Modeling sources of self-report bias in a survey of drug use epidemiology. *Ann Epidemiol*. 2005; 15:381–389. DOI: 10.1016/j.annepidem.2004.09.004 [PubMed: 15840552]
- Kapelner A, Chandler D. Preventing satisficing in online surveys: A “kapcha” to ensure higher quality data. *Inf Syst J*. 2010:1–10.
- Kim, DS., McCabe, CJ., Yamasaki, BL., Louie, KA., King, KM. Detecting random responders with infrequency scales using an error balancing threshold. *Behav Res Methods*. 2017. doi:<https://doi.org/10.3758/s13428-017-0964-9>
- Lara DR, Ottoni GL, Brunstein MG, Frozi J, De Carvalho HW, Bisol LW. Development and validity data of the Brazilian Internet Study on Temperament and Psychopathology (BRAINSTEP). *J Affect Disord*. 2012; 141:390–398. DOI: 10.1016/j.jad.2012.03.011 [PubMed: 22460054]
- Lenhart, A. Pew Res Cent. 2015. Teens, Social Media and Technology Overview 2015.
- Lindgren KP, Neighbors C, Teachman BA, Gasser ML, Kaysen D, Norris J, Wiers RW. Habit doesn't make the predictions stronger: Implicit alcohol associations and habitualness predict drinking uniquely. *Addict Behav*. 2015; 45:139–145. DOI: 10.1016/j.addbeh.2015.01.003 [PubMed: 25665917]
- Maniaci MR, Rogge RD. Caring about carelessness: Participant inattention and its effects on research. *J Res Pers*. 2014; 48:61–83. DOI: 10.1016/j.jrp.2013.09.008
- Meade AW, Craig SB. Identifying careless responses in survey data. *Psychol Methods*. 2012; 17:437–455. DOI: 10.1037/a0028085 [PubMed: 22506584]
- Meade AW, Michels LC, Lautenschlager GJ. Are internet and paper-and-pencil personality tests truly comparable?: An experimental design measurement invariance study. *Organ Res Methods*. 2007; 10:322–345. DOI: 10.1177/1094428106289393
- Meyer JF, Faust KA, Faust D, Baker AM, Cook NE. Careless and random responding on clinical and research measures in the addictions: A concerning problem and investigation of their detection. *Int J Ment Health Addict*. 2013; 11:292–306. DOI: 10.1007/s11469-012-9410-5
- Miguel E, Camerer C, Casey K, Cohen J, Esterling KM, Gerber A, Glennerster R, Green DP, Humphreys M, Imbens G, Laitin D, Madon T, Nelson L, Nosek BA, Petersen M, Sedlmayr R, Simmons JP, Simonsohn U, Van der Laan M, Castro MF, Nosek BA, Gerber AS, Malhotra N, Casey K, Simmons JP, Ioannidis JPA, Stroebe W, Humphreys M, Lane DM, Dunlap WP, Rosenblum M, Van der Laan MJ, Rubin D, Gerber AS, Simonsohn U, Mathieu S, Zarin DA, Tse T, Neumark D. Social science. Promoting transparency in social science research. *Science*. 2014; 343:30–1. DOI: 10.1126/science.1245317 [PubMed: 24385620]
- Mosimann JE, Wiseman CV, Edelman RE. Data fabrication: Can people generate random digits? *Account Res*. 1995; 4:31–55. DOI: 10.1080/08989629508573866
- Nederhof AJ. Methods of coping with social desirability bias: A review. *Eur J Soc Psychol*. 1985; 15:263–280. DOI: 10.1002/ejsp.2420150303
- Nichols DS, Greene RL, Schmolck P. Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *J Clin Psychol*. 1989; 45:239–250. DOI: 10.1002/1097-4679(198903)45:2<239::aid-jclp2270450210>3.0.co;2-1 [PubMed: 2723079]
- Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, VandenBos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T. Promoting an open research culture. *Science*. 2015; 348:1422–1425. DOI: 10.1126/science.aab2374 [PubMed: 26113702]
- Oppenheimer DM, Meyvis T, Davidenko N. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J Exp Soc Psychol*. 2009; 45:867–872. DOI: 10.1016/j.jesp.2009.03.009

- Pape H, Storvoll EE. Teenagers' "use" of non-existent drugs. A study of false positives. *Nord Stud Alcohol Drugs*. 2006; s1:97–111. doi:<https://doi.org/10.1177/145507250602301S13>.
- Patil P, Peng RD, Leek JT. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspect Psychol Sci*. 2016; 11:539–544. DOI: 10.1177/1745691616646366 [PubMed: 27474140]
- Robinson-Cimpian JP. Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. *Educ Res*. 2014; 43:171–185. DOI: 10.3102/0013189X14534297
- Roesch NJ, Jamieson DW. Twenty years of bogus pipeline research: A critical review and meta-analysis. *Psychol Bull*. 1993; 114:363–375. DOI: 10.1037/0033-2909.114.2.363
- Satorra A, Bentler P. A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*. 2000; 66:507–514. DOI: 10.1007/bf02296192
- Savalei V. Understanding robust corrections in structural equation modeling. *Struct Equ Model A Multidiscip J*. 2014; 21:149–160. DOI: 10.1080/10705511.2013.824793
- Schmitt N. Uses and abuses of coefficient alpha. *Psychol Assess*. 1996; 8:350–353. DOI: 10.1037//1040-3590.8.4.350
- Vosylis R, Žukauskien R, Malinauskien O. Comparison of internet-based versus paper-and-pencil administered assessment of positive development indicators in adolescents' sample. *Psichologija/ Psychology*. 2012; 45:7–20.
- Wardell JD, Rogers ML, Simms LJ, Jackson KM, Read JP. Point and click, carefully: Investigating inconsistent response styles in middle school and college students involved in web-based longitudinal substance use research. *Assessment*. 2014; 21:427–42. DOI: 10.1177/1073191113505681 [PubMed: 24092819]
- Woods CM. Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *J Psychopathol Behav Assess*. 2006; 28:186–191. DOI: 10.1007/s10862-005-9004-7
- Wright KB. Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *J Comput Commun*. 2005; 10:0.doi: 10.1111/j.1083-6101.2005.tb00259.x

Highlights

- Minimal participant interaction may increase frequency of random self-report responses
- No research (< 1%) in the addictions reports screening for random responses.
- Random responses can inflate correlations between skewed variables.
- Studies collecting self-report data should screen for random responses.

Figure 1.

This figure illustrates three types of response probabilities where the true “trait” level of a construct is zero. The first response pattern illustrates a uniform random response, where a participant chooses from all response options at random. The second illustrates a random long-string response, where a participant has randomly chosen a single response option to repeat across all items (in this case, option 4). The third illustrates a valid response, where the participants’ trait level is 0, but the response probabilities for other response options are non-zero (due to measurement error or other sources of random error).

Figure 2.
The five different distributions estimated.

Figure 3.
Effects of different proportions of long-string and uniform responders on correlation coefficients across a range of distributions.

Figure 4a.

Figure 4b.

Figure 4.

The effects of replacing 2.5% of valid cases with uniform data. Circles represent “true” data, and x’s represent invalid data (i.e. uniform responses). 4(a) illustrates the correlation between mother warmth and alcohol use in valid data only. 4(b) illustrates the same correlation replacing 2.5% of cases with uniform cases. The even dispersion of the uniform cases, relative to the skewed valid responses, alters the correlation even with the injection of a small number of cases.

Figure 5.

Type I error inflation across magnitude of parameter bias and sample sizes. This figure illustrates a simulation of Type I error rate inflation (i.e. false positives) when the true correlation is 0 for parameter bias ranges from .05 to .40 in sample sizes ranging from 25 to 985. Simulation details and code are provided in Appendix 1³.

³Supplementary material can be found by accessing the online version of this paper at <http://dx.doi.org> and by entering doi:...

Table 1

Main effects of uniform and long-string proportions on statistical estimates

Statistic	<i>r</i>	
	Uniform	Long-string
Heavy Positive Skew	0.007	0.133
Moderate Positive Skew	−0.042	0.127
No Skew	−0.071	0.119
Moderate Negative Skew	−0.041	0.127
Heavy Negative Skew	0.011	0.138

Note: Table displays regression coefficients, scaled to represent the rate of change in the outcome per 10% increase in proportion within each type of distribution when none of the other type of responder is present. Uniform respondents reflect “random” responders who choose at random; long-string responders reflect responders who randomly choose 1 response to repeat across all items. All effects are significant ($p < 0.001$).

Table 2

The effects of replacing valid data with invalid responses in the Add Health Data

	True data Close to Mom	Long String Response Proportions					Uniform Response Proportions				
		2.5%	5%	10%	20%	30%	2.5%	5%	10%	20%	30%
Past year alcohol use	0.12	0.24	0.33	0.45	0.60	0.69	.18	0.22	0.28	0.34	0.35
Past year five or more	0.09	0.23	0.34	0.48	0.63	0.71	.17	0.23	0.30	0.37	0.38
Past year drunk	0.10	0.26	0.36	0.50	0.65	0.72	.19	0.25	0.32	0.38	0.40
<i>Mom really cares about me</i>											
Past year alcohol use	0.04	0.25	0.37	0.52	0.67	0.75	.16	0.23	0.34	0.37	0.39
Past year five or more	0.03	0.27	0.41	0.57	0.72	0.78	.18	0.27	0.36	0.43	0.44
Past year drunk	0.03	0.29	0.44	0.60	0.74	0.80	.20	0.29	0.38	0.44	0.45

* Note: coefficients reflect the estimated correlations between items in the Add Health data. We first display the correlations for the valid responses, then we display the effects of replacing a varying % of valid responses with simulated random responses.