Published in final edited form as:

Exp Clin Psychopharmacol. 2023 October ; 31(5): 908–919. doi:10.1037/pha0000645.

Are the Attention Checks Embedded in Delay Discounting Tasks a Valid Marker for Data Quality?

Shahar Almog[1], **Andrea Vásquez Ferreiro**[1], **Meredith S. Berry**[1,2], **Jillian M. Rung**[1]

[1]Department of Health Education and Behavior, University of Florida, Gainesville, FL, USA

[2]Department of Psychology, University of Florida, Gainesville, FL, USA

Abstract

To ensure good quality delay discounting (DD) data in research recruiting via crowdsourcing platforms, including attention checks within DD tasks have become common. These attention checks are typically identical in format to the task questions but have one sensical answer (e.g., “Would you prefer $0 now or $100 in a month?”). However, the validity of these attention checks as a marker for DD or overall survey data quality has not been directly examined. To address this gap, using data from two studies (total *N*=700), the validity of these DD attention checks was tested by assessing performance on other non-DD attention checks and data quality measures both specific to DD and overall survey data (e.g., providing nonsystematic DD data, responding inconsistently in questionnaires). We also tested whether failing the attention checks was associated with degree of discounting or other participant characteristics to screen for potential bias. While failing the DD attention checks was associated with a greater likelihood of nonsystematic DD data, their discriminability was inadequate, and failure was sometimes associated with individual differences (suggesting that data exclusion might introduce bias). Failing the DD attention checks was also not associated with failing other attention checks or data quality indicators. Overall, the DD attention checks do not appear to be an adequate indicator of data quality on their own, for either the DD task or surveys overall. Strategies to enhance the validity of DD attention checks and data cleaning procedures are suggested, which should be evaluated in future research.

Keywords

delay discounting; data quality; attention; crowdsourcing; Amazon Mechanical Turk; data screening

Online research recruiting participants on crowdsourcing platforms, in which researchers post tasks and interested individuals complete them on their own accord for compensation, has become popular across disciplines. Research relying on crowdsourcing for execution can be fast and inexpensive, reaching general or more specific populations under study.

Correspondence concerning this article should be addressed to Shahar Almog or Jillian Rung, Department of Health Education and Behavior, University of Florida, Gainesville, FL 32611. almogs@ufl.edu, rung.jillian@gmail.com.

Analysis codes are publicly available at OSF and can be accessed at https://osf.io/ax7dp/

This study was not preregistered.

It also has the benefit of increasing accessibility of research participation by way of reducing participant travel and scheduling burdens. However, conducting research utilizing crowdsourcing has challenges associated with data quality, such as participant misrepresentation and inattentive responding (Buhrmester et al., 2018; Newman et al., 2021).

In allowing participants to determine the location and timing of their participation, researchers have less control over participants' environments, which may compromise participant attention. A recent meta-analysis suggested an approximate prevalence rate of 11% of crowdsourced respondents providing inattentive responses (Jones et al., 2022). Because undetected inattentive responses might impact research results (Nichols & Edlund, 2020), researchers have developed best-practice suggestions to guard against this threat (Aguinis et al., 2021; Newman et al., 2021; Strickland et al., 2022). One example of these best practices is embedding specific items in the survey to assess data quality. For example, one type of item is a consistency check, which involves repeating a question in different parts of the survey and evaluating whether the responses were consistent across repetitions (e.g., age, sex). Another type of measure is an instructional attention check, which involves presenting an instruction that asks for a specific response. Only those who fully read the question are likely to answer these questions correctly. Failure of these types of items can be used to flag the participant's data for exclusion from analysis due to evidence suggesting poor quality data overall. Failure of these types of quality checks may also serve as a justification to decline participant compensation or bonuses, in accord with the study protocol and institutional ethical policies.

One measure that is frequently administered in online research and highly relevant to addictions is delay discounting (DD). Delay discounting is the decline in the value of a reward due to a delay to receive that reward and serves as a measure of "impulsive" decision making (Ainslie, 1975). In many DD tasks for humans, participants are asked to choose between two rewards with different delays: one is smaller and sooner, and the other is larger but delayed (e.g., "Would you prefer $50 now or $100 in a year?"). The question is repeated with different sooner-smaller amounts and delays, and participants' choices are ultimately used to map the decay of the larger delayed reward over time and estimate a degree of discounting (i.e., extent to which one devalues outcomes due to delay). Steep (or high) discounting coincides with a relatively high preference for the smaller immediate reward across delays. It is often referred to as relatively "impulsive" decision making, and it is associated with a range of addictive behaviors, including substance use disorders (Bickel et al., 2014, 2019; MacKillop et al., 2011). Delay discounting is widely studied because not only is it a significant and unique predictor of use disorders (e.g., Strickland et al., 2019), it can also serve as a behavioral index of recovery (e.g., Athamneh et al., 2019) or withdrawal processes (Xu et al., 2022), and is a manipulable construct that may be useful in treating/preventing addictions (Rung & Madden, 2018; Rung, Peck, Hinnenkamp, Preston, & Madden, 2019).

Because DD tasks often include many repetitive questions with a similar structure, it is important to ensure the participant is attending to each set of presented choices when making a response. With the expansion of delay discounting tasks to online research, it became common to embed two or more attention checks in DD tasks (e.g., Athamneh et al., 2020; Craft et al., 2022; Mellis et al., 2017; Rung & Madden, 2019; Sze et al., 2017). The most commonly used attention checks have the same format and structure as the actual questions posed in the task, however, one of the presented options is intentionally designed such that it would be highly unlikely to be preferred. An example of these checks is: “Would you prefer $0 now or $100 in a month?”. When a participant fails to choose the sensible answer, the participant is considered to not be fully attentive to the survey and his/her data may be flagged for exclusion from analysis (e.g., Athamneh et al., 2020; Craft et al., 2021, 2022; Pope et al., 2019; Stein et al., 2018).

Although researchers have been using the aforementioned types of questions in DD tasks, they have not, to the best of our knowledge, been specifically validated for capturing inattention. In using these types of attention checks in our own studies, this lack of information on their validity became a concern due to post-survey comments from participants. Some participants who failed an attention check acknowledged their error. For example, one participant said, “I accidentally clicked the choice '$0 in …' in one of the questions, but I can't go back. I hope my HIT doesn't get rejected by that.” This particular participant and others in our research who have failed these types of attention checks were aware of failing to choose the sensible option and remembered to comment on their error at the end of the survey, sometimes up to 20 minutes later. This behavior seems incongruent with the careless responding that these attention checks are thought to capture. Together with the lack of explicit evaluation of the validity of these attention checks, these participant responses stress the need to ensure the validity of these attention checks, especially if failing them serves as a justification to decline a bonus payment or exclude data from analysis in either a local or global way (e.g., excluding from discounting-related analyses or broadly from any analyses).

Given the lack of published methodological research on these types of attention checks for DD tasks, the goal of the present study was to evaluate the validity of these checks using data from two different studies (total of 700 participants) recruiting participants from Amazon Mechanical Turk (MTurk). In these secondary data analyses, we first investigated the relationship between performance on the DD attention checks and the quality of the DD data based on standard criteria (i.e., those from Johnson and Bickel, 2008). Second, we examined the relationship between performance on the DD attention checks and the quality of the survey data overall, as reflected in alternative measures of quality (e.g., scale reliabilities, measures of consistency). Lastly, we investigated whether failing the attention checks was associated with degree of discounting and other individual characteristics to determine if these DD task attention checks may be biased.

Study 1

Study 1 Methods

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the current study. This secondary analysis study was not preregistered.

Participants—Study 1 recruited MTurk Workers who were 21 years old or older, with a history of at least 100 approved tasks, an approval rate of at least 99%, and whose accounts reflected residency in the United States. Recruitment and data collection occurred between November 2019 and January 2021. Participants first completed a separate screening survey to verify adequate English language proficiency and other eligibility criteria relevant to the primary study (see Rung et al., 2022 for details on the screener). To be eligible to participate Workers had to report occasional alcohol use, no current regular use of nicotine, and no current uncontrolled mental health conditions (e.g., unmanaged major depressive disorder). Those who met criteria were then granted the ability to view and participate in the primary study on their own accord on MTurk (i.e., no direct invitation sent). Consent was obtained for both screening and the main study prior to completion of any study procedures. A total of 360 participants who were screened were eligible for and consented to complete the primary study. In order to be included in this secondary analysis, participants had to have completed a monetary DD task with attention checks and complete the study procedures in full. These criteria resulted in 90 participants being excluded for the aforementioned reasons, yielding a final sample of 270 participants.

Compensation for the screener and the main survey was based on anticipated survey duration at a rate of approximately $7.00 per hour. Compensation for the screener was always $0.35, and compensation for the primary study procedures varied based on the version individuals participated in (e.g., piloting vs. study proper). Compensation for the latter typically ranged from $4.10 to $5.25. All procedures were approved by the University of Florida Institutional Review Board (protocols IRB201902083 and IRB202102744). Secondary data analyses of both Study 1 and Study 2 were approved under IRB202200191.

Procedures—Regardless of the main survey version (e.g., piloting vs. study proper), all participants in the present analytic sample completed a monetary DD task with two attention checks embedded in it. Following the monetary DD task, participants also completed additional questionnaires that were theoretically relevant to the primary study hypotheses; these questionnaires are described in further detail below, which are included for evaluating data quality in the present research. After completion of these measures, participants provided information on alcohol and current and past nicotine use, demographic information (e.g., age, sex, subjective socioeconomic status), and were given the opportunity to leave comments about the survey. Finally, participants were provided a randomly-generated survey completion code to be submitted on MTurk for compensation verification purposes. Only measures pertinent to the primary aims of this report are included herein.

Measures

Monetary Delay Discounting Task.: The task was a titrating immediate amount procedure that arranged hypothetical monetary outcomes (Du et al., 2002). There were seven larger-later reward delays (i.e., 1 week, 2 weeks, 1 month, 6 months, 1 year, 5 years, 25 years), which were presented across blocks of trials. Each block consisted of six trials. In each trial, individuals were asked to choose between an immediate reward or a delayed one. On the first trial in a block, the immediate reward was always $50, and the larger later reward was always $100. On subsequent trials, the immediate amount changed based on the participant's choice. If the participant chose the immediate reward, it decreased on the next trial, and if the participant chose the delayed $100, the immediate reward increased. The adjustments to the immediate reward decreased in half from one trial to the next (i.e., $25 after the first trial and then $12.50, $6.25, etc.). After six trials, at the end of each block, the presumed seventh immediate amount (if it were to be presented), was taken as the indifference point. The indifference point reflects the subjective value of the larger reward given its delay. The indifference points for each delay were used to calculate degree of discounting (the area under the curve, AUC), and a measure of DD data quality. The AUC is the proportion of the graphical area that falls below the curve formed by adjacent indifference points when plotted (Myerson et al., 2001). The AUC ranges from 0 to 1, where lower values indicate greater discounting. The derived measure of DD data quality is outlined in the Data Analysis section.

Two attention checks were embedded in the DD task. The first check appeared in the second block of the one-month delay (“Would you rather have $50 now or $0 in 2 weeks?”), and the second check appeared in the sixth block of the five-year delay (“Would you rather have $0 now or $100 in 5 years?”). Choosing the $0 option for either question was considered failing the respective attention check. Because only three participants failed both attention checks (of a total of 46 who failed any), failure of these items was categorized dichotomously for analysis: failing at least one of the DD attention checks or passing both.

Internal Consistency.: We used the Consideration of Consequences questionnaire (CFC, Strathman et al., 1994) and the Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ-20, Aluja & Blanch, 2011; Torrubia et al., 2001) to obtain measures of internal consistency among participant subsamples (detailed in Data Analysis). In the full sample, Cronbach’s alpha was 0.96 for the CFC, and 0.84 and 0.79 for the reward and punishment subscales of the SPSRQ-20, respectively. The total scale scores were not analyzed or further discussed herein, but details on the questionnaires are provided in the Supplemental Material.

Other Quality Checks.: Response (in)consistency was quantified as whether or not individuals responded the same to items that were presented both in the screener and in the main survey, which should more often than not yield the same response over short durations. The items used for quantifying consistency were age, biological sex, number of siblings, past smoking status, current smoking status, and frequency and quantity of drinking alcohol. Several of these questions were added at later dates to the screener (biological sex, prior nicotine use) or screener and main survey (number of biological siblings) for the purposes of evaluating consistency. Of the present analytic sample, 130 participants completed the later version that had these additional questions (48% of the sample). Additionally, because inconsistencies were relatively rare, response consistency was represented dichotomously as inconsistently responding to any one of these seven items (inconsistent) or not (consistent). This coding scheme resulted in 25 datasets being coded as having inconsistent responses (9.3%). Inconsistency of responses to questions that involve relatively stable characteristics are indicative of low quality data and in particular, misrepresentation (MacInnis et al., 2020).

Data analysis—We used *R* Statistical Software (v4.1.2) and RStudio to conduct all analyses. First, using the rstatix package (Kassambara, 2021), descriptive statistics were used to characterize the sample overall, and by groups of those who failed at least one of the DD attention checks and those who passed both. The test type chosen to evaluate differences in demographic characteristics across the groups was based on whether the (continuous) variables met normality assumptions, which was judged by Shapiro-Wilks tests.

To determine the quality of the discounting data provided by each participant, the data were screened according to the Johnson and Bickel (2008) criteria. The first criterion captures instances of nonmonotonicity in the discounting data. Using the criteria with our task parameters, the data were flagged as nonsystematic if there was any increase of $20 or more in an indifference point relative to the immediately preceding one. The second criterion describes instances of non-discounting: if the last indifference point was not lower than the first indifference point by at least $10. Data meeting one or both criteria were coded as nonsystematic. The prevalence of nonsystematic DD data in the analytic sample was relatively low ($n = 26$, 9.6%). Eighteen of these datasets failed the first criterion (non-monotonicity), 12 of these failed the second criterion (non-discounting), and four failed both criteria.

The validity of the DD attention checks was tested in several ways. The first method determined whether failing the DD attention checks was associated with providing nonsystematic DD data. For this, we conducted a logistic regression with nonsystematic DD data as the outcome (systematic vs. nonsystematic), and DD attention check performance as a predictor (passed both or failed at least one). Covariates of duration to complete the DD task, the order of DD task completion (whether the monetary discounting task was first or not), and group membership (randomly assigned in the primary study) were included. Further details on covariates and their rationale for inclusion are provided in Appendix A). To better characterize the results of this regression analysis, using the ROCR package (Sing et al., 2005), we created a classification table using predictions from the logistic regression to compute the area under the Receiver Operating Characteristic curve (i.e., AUC-ROC). The ROC is the probability curve of the true and false positive rate, and it serves as a measure of a model’s capability to distinguish between the two groups—in this instance, systematic versus nonsystematic DD data. In general, an AUC-ROC of 0.50 indicates no discrimination, and values between 0.5 and 0.7 are considered poor. A minimum of 0.7 is considered acceptable, and a value of 0.8 and above is considered excellent (Hosmer et al., 2013).

The second method of evaluating the validity of these DD attention checks was via comparisons of reliability (internal consistency) coefficients from questionnaires across those who passed versus failed any of the checks (see Barends & de Vries, 2019 for a similar application). Reliabilities for the CFC and the subscales of the SPSRQ were calculated using the Psych package (Revelle, 2022). Differences in the scale reliabilities across the groups were then compared using the COCRON web interface (http://comparingcronbachalphas.org/) (Diedenhofen & Musch, 2016), which employs methods outlined in Feldt et al. (1987).

The third method of evaluating validity of the DD attention checks was to evaluate the relation between failure of these checks and performance on data quality measures from other portions of the survey. For this analysis, we used a logistic regression with consistency of responding as the outcome (any inconsistent responses vs. fully consistent), with failing at least one of the DD attention checks as the predictor (vs. passing both). Since the responses on the consistency items should not have been affected by the experimental manipulation, the only covariates included in this model were variables that captured speed of responding: the duration to complete the DD task, and the duration to complete the whole survey.

Finally, to assess whether those who failed at least one of the DD attention checks versus those who passed both differed in degree of discounting, we used beta regression (using the betareg package; Cribari-Neto & Zeileis, 2010). Beta regression is a type of generalized linear model that assumes a proportion outcome that is bounded between 0 and 1 (Ferrari & Cribari-Neto, 2004), which matches the properties of AUC (see Peck et al., 2020; Rung et al., 2018 for similar applications). In this model, failing or passing the attention checks predicted DD AUC, controlling for the duration to complete the DD task, the DD task Order, and Group, with the addition of whether the data were systematic.

For beta regression models, typical diagnostic procedures were conducted. These included screening for observations producing large standardized residuals (>2), observations having high influence on the fitted model (Cook's Distances exceeding the typical $4/N$ cutoff), and evaluation of Q-Q plots. Observations that exceeded the aforementioned thresholds were flagged for exclusion, and the model was re-run without them. Removal of these values did not change the nominal results of the model (i.e., significance of primary predictors of interest) and as such only the model results with all observations included are reported herein.

To ensure including individuals who failed both DD attention checks in the same group as those who only failed one did not impact the results, all analyses were conducted with and without the three individuals who failed both DD attention checks. The results of these analyses were nominally the same in terms of statistical significance and are not further discussed. In addition, all regression models were also run as unadjusted models without covariates to evaluate the robustness of results. The conclusions of the adjusted and unadjusted models were the same and only the results from the adjusted models are reported. The analysis code (for Studies 1 and 2) is publicly available via the Open Science Framework at https://osf.io/ax7dp/ (Almog et al., 2022).

Study 1 Results

Of the 270 participants, 46 (17.0%) failed at least one of the DD attention checks. Specifically, 42 participants failed the first-presented attention check, one failed the second-presented check, and three failed both. The demographics of the whole sample ($N = 270$) and of the two subgroups of those who failed at least one attention check ("Failed") versus those who passed both ("Passed") are presented in Table 1. There was only one significant demographic difference across those who passed versus failed at least one of the DD attention checks: there was a higher proportion of individuals who identified as Hispanic or Latinx that failed the DD attention checks (17.4% vs. 9.8% respectively, $p = .014$).

The first method of evaluating the validity of the DD attention checks focused on DD data quality defined as nonsystematic DD data. While the overall prevalence of nonsystematic DD data was low, it was higher among those who failed at least one DD attention check (n =9, 19.6%) than those who passed both (n = 17, 7.6%). Consequently, including failure of the DD attention checks as a predictor significantly improved the logistic regression model predicting nonsystematic DD data (i.e., compared to the covariate-only model; $\chi^2[1] = 5.78$, $p = .016$). Failing at least one of the DD attention checks significantly predicted a greater likelihood of nonsystematic data ($OR = 3.27$, $p = .012$). Table 2 presents coefficient estimates, standard errors, and odds ratios with 95% confidence intervals for each predictor. The model had an AUC-ROC of 0.68, which is below the 0.70 minimum recommendation of an acceptable fit (i.e., adequate discrimination). In other words, the model predicted nonsystematic discounting data 17.6% better than chance.

The second method of evaluating the validity of the DD attention checks via comparison of scale reliabilities indicated no differences in internal consistency as a function of DD attention check failure (ps > .46). For the CFC, those who failed at least one versus passed both of the DD attention checks each had a Cronbach's alpha of 0.96. For the reward subscale of the SPSRQ, Cronbach's alphas were 0.80 and 0.79; and for the punishment subscale Cronbach's alphas were 0.82 and 0.85 for those who failed any versus passed both DD attention checks, respectively.

For the third method of evaluating validity, failure of any DD attention checks was also not predictive of inconsistent responses on survey items asking about stable characteristics. This finding occurred despite a nominally higher percentage of individuals who failed the DD attention checks also failing at least one consistency check (n = 7, 15.2%) relative to those who passed both DD attention checks (n = 18, 8.0%). There was a lack of significant model improvement when adding failure of DD attention checks to the model controlling for necessary covariates ($\chi^2[1] = 1.83$, $p = .176$).

Lastly, the beta regression indicated that the DD attention checks might be biased towards certain degrees of discounting (presented in Table 3). Failing at least one of the DD attention checks was a significant predictor of DD AUC ($\chi^2[1] = 5.30$, $p = .021$). Those who failed at least one attention check showed less discounting, evidenced by higher AUCs (model-estimated *mean* AUC = .48) compared to those who passed both attention checks (model-estimated *mean* AUC = .39).

Study 1 Discussion

The purpose of Study 1 was to evaluate the validity of attention checks commonly embedded in DD tasks. We assessed performance on the attention checks in relation to the quality of discounting data itself, and to the performance on other indicators of data quality in the survey. These initial results suggest that failing DD task attention checks is a correlate of poor-quality DD data, as defined by the nonsystematic data algorithms of Johnson and Bickel (2008). However, the ability of these DD attention checks to correctly classify nonsystematic data was suboptimal, and failure of at least one of the DD attention checks was not associated with other measures of poor data quality (i.e., scale reliabilities, consistency checks). The DD attention checks also exhibited bias: those who failed any of these attention checks had lower degrees of DD (higher AUC) and were more likely to identify as Hispanic or Latinx. These findings raise concerns about sample representativeness and inclusivity if basing exclusions on these checks.

Overall, these results suggest that attention checks embedded in DD tasks may not be a valid indicator of data quality. Further, they may disproportionately exclude certain groups of individuals. To evaluate the robustness of these results, we replicated these analyses with a second, larger data set that involved the same DD task and additional data quality measures.

Study 2

The data for Study 2 were from a cross-sectional study on associations between exposure to nature, music habits, delay discounting, substance use, and subjective wellbeing. In addition to the DD attention checks, Study 2 included additional types of attention checks, which allows for extending the evaluation of the validity of the DD attention checks to other types of commonly used data quality metrics. Study 2 utilized the same DD task and analytic approach as used in Study 1. Where there were procedural differences, they are noted below.

Study 2 Methods

Participants—During March to October 2021, Study 2 recruited 450 MTurk Workers 18 years old or older, with a history of at least 500 approved tasks on the platform (in early stages of recruitment this was 100), an approval rate of at least 95%, and accounts indicating residency in the United States. After consenting to the screening procedures, participants completed a short unpaid screener that included English language proficiency checks and items intended to check for consistent responding in the main survey. In contrast to Study 1, Study 2 recruited a general sample that did not involve additional criteria. In addition, those who passed the screening were immediately presented with a second consent form for the main survey and proceeded to participate if they provided consent (i.e., all procedures conducted on the same day/in same survey). Those who did not pass the screening were branched out using survey logic.

The sample for the current analysis included all data available from 430 participants. This number excludes nineteen participants who completed the primary study procedures after making multiple attempts to pass the screener by modifying their responses across attempts, and one participant who responded twice (due to survey error) without changing the responses thus excluding the second observation. These participants were identified based on their Worker IDs appearing multiple times in the recorded data. The median duration of survey completion was 25 minutes, and the compensation was $3. The amount was based on an hourly rate of $7.20. All procedures were approved by the University of Florida Institutional Review Board under IRB202100096.

Procedures—For Study 2, participants first completed a monetary DD task, followed by questions and scales on health behaviors, subjective wellbeing, environmental factors related to health, and demographics. At the end of the survey participants were provided with a completion code, which they submitted on MTurk as verification of participation.

Measures—The DD task and the DD attention checks were identical to those in Study 1, though the questionnaires included were different and did not include the CFC or SPSRQ. Below are the unique measures used in Study 2.

Internal Consistency.: The BBC Subjective Wellbeing Scale (BBC-SWB, Kinderman et al., 2011; Pontin et al., 2013) and the Positive Affect Negative Affect Scale (PANAS-GEN, Watson et al., 1988) were used to calculate internal consistency measures among subsamples. In the full sample, Cronbach's alpha was 0.96 on the BBC-SWB and 0.94 for the positive and negative affect subscales of the PANAS-GEN, separately. As in Study 1, the total scores were not analyzed or further discussed herein. For details on these questionnaires see Supplemental Material.

Other Quality Checks.: In addition to the two DD attention checks, four general quality checks were embedded in the survey. These four quality checks included two consistency checks and two instructional attention checks. For this study, the consistency checks involved questions that should yield the same information as opposed to identical responses (i.e., year of birth and age, favorite musical genre and genres most frequently listened to). Inconsistent responding was considered indicative of poor data quality. For the instructional attention checks, the first asked participants to remember an instruction from an earlier page: to remember to type the word blue when they will be asked to type the word green later in the survey. The second attention check instructed participants to choose a specific response option rather than answering the initially presented question: "What is your main type of exercise? We want to make sure you are paying attention to our survey. This is an attention check question. Please do not choose your main type of exercise but rather choose Other". In earlier versions of the survey not all quality checks were presented, resulting in 34 participants that were not presented with the exercise attention check, and five that were not presented with the genre consistency item. Similar to Study 1, regardless of the number of attention items completed all participants were included in analyses. Because these additional quality checks relied on different behaviors, passing these was represented using two dichotomous variables. The first reflected inconsistency based on failing at least one of the two consistency items ($n = 30$, 7.0%), and the second reflected inattention based on failing at least one of the instructional checks ($n = 96$, 22.3%).

Data Analysis—The data analysis approach was identical to that for Study 1 (e.g., criteria for nonsystematic DD data, logistic regression analyses predicting nonsystematic DD data). For Study 2, the prevalence of nonsystematic DD data was relatively low: 55 DD datasets (12.8%) were nonsystematic, with 22 qualifying as nonsystematic for failing the first criterion (non-monotonicity), 21 for failing the second (non-discounting), and 12 for failing both. In addition to the analysis testing whether failure of the DD attention checks was predictive of failure of other data quality measures (i.e., inconsistent responses), an identical model using performance on the instructional attention checks as the outcome was also conducted.

Regarding covariates, only the duration to complete the DD task and the duration to complete the whole survey were included in the relevant regression models. In Study 2 there were no manipulations and therefore no procedural covariates to include. Similar to

Study 1, all analyses were conducted with and without individuals who failed both DD attention checks ($n = 2$), and both adjusted and unadjusted regressions were conducted. Omitting those who failed both DD attention checks did not impact results, but there was one differential outcome across the adjusted and unadjusted models. In the latter instance, the results of both analyses are reported. Otherwise, only the results with all participants and those from the adjusted models are reported.

Study 2 Results

Of the 430 participants, 48 (11.2%) failed at least one of the DD attention checks. Of these, 37 failed the first attention check, 9 failed the second, and 2 failed both. The demographics of the full sample and the two subgroups based on DD attention check performance (failed at least one for “Failed,” otherwise “Passed”) are presented in Table 4, with the p value for the appropriate statistical test. The groups were not significantly different on any of the demographic variables.

The first method of testing the validity of the DD attention checks focused on predicting nonsystematic DD data. In the Study 2 sample, there was a higher percentage of nonsystematic DD data among those who failed any of the DD attention checks ($n =$ 15, 31.3%) than among those who passed both ($n = 40$, 10.5%). Including failure of the DD attention checks as a predictor of nonsystematic DD data in the logistic regression significantly improved the basic covariate-only model ($\chi^2[1] = 12.59$, $p < .001$); and it was a significant predictor of a greater likelihood of providing nonsystematic DD data ($OR = 3.83$, $p < .001$). The coefficient estimates, standard errors, and odds ratios with 95% confidence intervals for the model are presented in Table 2. Despite the predictor improving the model and being significantly associated with nonsystematic DD data, the model only achieved an AUC-ROC of 0.60 (i.e., discriminated approximately 10% better than chance).

For the second method of evaluating DD attention check validity, there were no significant differences in questionnaire internal consistency between those who failed at least one of the DD attention checks and those who passed both (all ps > .112). For the BBC-SWB scale Cronbach’s alphas were 0.97 and 0.96 for those who failed any and those who passed both DD attention checks, respectively. For the Positive Affect subscale of PANAS the Cronbach’s alphas were 0.93 and 0.94, and for the Negative Affect subscale 0.96 and 0.94, for those who failed any versus passed both the DD attention checks, respectively.

For the third method of evaluating DD attention check validity, similar percentages of participants provided inconsistent responses among those who failed any of the DD attention checks ($n = 4$, 8.3%), and those who passed both ($n = 26$, 6.8%). The logistic regression showed that failure of any of the DD checks was not associated with inconsistent responses for stable characteristics ($\chi^2[1] = .13$, $p = .719$). For the instructional attention checks, a nominally higher percentage of participants who failed at least one of the DD attention checks also failed at least one of the instructional checks ($n = 15$, 31.2%) compared to those who passed both DD attention checks ($n = 81$, 21.2%), but the logistic regression indicated failure of the DD attention checks did not significantly predict failure of the instructional attention checks ($\chi^2[1] = 2.26$, $p = .133$).

Lastly, the DD attention checks were inconsistently associated with degree of discounting. In contrast to Study 1, inclusion of whether individuals failed any of the DD attention checks as a predictor did not significantly improve the beta regression predicting DD AUC beyond survey/task duration-related covariates ($\chi^2[1] = .81$, $p = .369$). However, the unadjusted model showed agreement with Study 1 in that failing any of the DD attention checks was predictive of lower discounting ($\chi^2[1] = 4.45$, $p = .03$). Those who failed any attention checks had an estimated mean AUC of .36, and those who passed both had an estimated mean of .28.

Study 2 Discussion

The results of Study 2 largely paralleled those of Study 1. While failing the DD attention checks was associated with an increased likelihood of providing nonsystematic DD data, their predictive utility was poor. Moreover, those who failed any of the DD attention checks did not show poorer performance on other indicators of data quality in the survey. The primary difference in findings relative to Study 1 were those pertaining to bias. There was no robust evidence for differences in degree of discounting between individuals who failed any of the DD attention checks and those who did not (although the unadjusted model aligned with Study 1 findings of suggested bias). Similarly, there were no demographic differences between those who failed DD attention checks and those who did not. Thus, the results of Study 2 suggest that while these typically used DD attention checks are not reliably biased, they are not a valid indicator of poor data quality.

General Discussion

The purpose of the current study was to examine the validity of commonly used DD attention checks and whether they may be biased using two large datasets containing an identical measure of DD and multiple, varied measures of overall data quality. Across both datasets, failing the DD attention checks was associated with an increased likelihood of nonsystematic DD data (typically considered a marker of quality; Johnson & Bickel, 2008), but the discriminability of these attention checks was poor. The consistency of this finding suggests that the validity of these commonly used attention checks in DD tasks is questionable. These findings inform techniques used to identify data quality and may be particularly relevant for substance use research, given the continued impactful and wide use of DD in this context.

At a broader level, failing the DD attention checks was also not associated with other various markers of poor data quality in other parts of the survey. Failure of these checks was not associated with response consistency for more stable characteristics, failure to follow task instructions (instructional attention checks), or incongruous responses on validated, standardized questionnaires (lower scale reliabilities). These findings challenge the idea that this style of attention check in DD tasks can serve as either a valid local (in-task) or global (across-survey) indicator of data quality. As such, we caution against using failure of these types of attention checks in DD tasks, at least in isolation, as a justification for excluding data on a task- or survey-level basis, or for purposes of withholding compensation bonuses, etc.

Overall, the DD attention checks showed few instances of bias; when they did, the findings were not robust across studies and analyses. In the first data set, failure of the DD attention checks was associated with a greater likelihood of identifying as Hispanic or Latinx but this finding failed to replicate in Study 2. In Study 1, failing the attention checks was associated with lower discounting, but this finding only replicated in Study 2 when speed of responding was not controlled for. The inconsistent results across the two studies, especially with the relatively low number of participants identifying as Hispanic/Latinx, call for additional research and caution in interpretation. Nevertheless, researchers should be diligent and ensure these and other types of attention checks used are not systematically flagging certain groups of individuals, because exclusions based on them may bias the sample and results (Berinsky et al. 2014).

Our findings suggest that failing the DD attention checks is not necessarily an outcome of inattentiveness or careless responding. This raises the question of why participants who provide good quality data otherwise fail them. Given the predictability of the algorithm in the adjusting task, it may be the case that these types of attention checks are not appropriate gauges of attention. After the first several trials or block, individuals are familiarized with the task structure and may adopt response strategies. As one participant commented at the end of the survey:

> During the part where I had to select what I preferred for money between now and a certain time, I accidentally selected the $0 option. I realized my mistake right away and I apologize. I had thought about my minimum amount before going through that particular phase for that amount of time and that is why I selected it so fast.

Of note, in the two studies we analyzed data from, the response options were also always presented on the same side of the screen. Adoption of particular strategies such as “I’m willing to wait 2 weeks no matter what” or “I’ll pick the immediate amount until it goes below $23” are reasonable and with the side of the immediate/delayed options being static, do not require the participant to carefully review all options in a given block. Most participants only failed one of the DD attention checks, and those who did tended to fail the first one presented (91% and 77% of those failing one check failed the first in Studies 1 and 2, respectively). That few individuals failed the second check (or both) supports the idea that the task structure with static response placement did not initially encourage attention to both options. Randomizing the side that the options are presented on would allow for adoption of such response strategies while necessitating attention to what is presented on each side of the screen; consequently, failure of these attention checks may more unambiguously reflect careless responding, which is of primary concern when excluding data. Future research should randomize the order of the task choices and examine the performance on, and validity of the attention checks under such conditions.

Several results across Studies1 and 2 support the above interpretation. In Study 1, those who failed attention checks had lower discounting, and nearly all of these participants failed the first check (where $0 is shown instead of the delayed $100). In Study 2, where fewer participants failed the first check, the difference in AUC was not as robust. Together, preference for the delayed larger outcome (i.e., shallower discounting) may be related to failing the first attention check, which supports the recommendation to not exclude these participants without other indication.

Although further research validating and refining these typically used DD attention checks is worthwhile, researchers should also consider the various ways of interpreting performance on them for subsequently judging the quality of data. For instance, Nichols and Edlund (2020) suggested that inattention can be approached as a continuum measured with several checks on the basis of a threshold rather than being based on performance of a single item with a pass/fail decision. When quality is judged based on a single item, there is a 50% chance of passing it even with a random response. Additionally, and especially in longer surveys, the level of one’s attention might change over the course of survey completion. Following this approach, failing a single DD attention check would not be used as a determinant of data quality. When including additional quality checks in the survey, the DD attention checks can serve as part of a larger constellation of inattentive or careless responding, which provides greater confidence for classifying data as poor quality and subsequently excluding it.

While the present study makes several important contributions, they must be considered in light of limitations. First and foremost, the studies from which data were used to evaluate our hypotheses were not originally designed to test the validity of the DD attention checks. Additional research specifically designed to address this question and extend our findings should be conducted, which will help inform researchers how to best assess DD in online studies while ensuring good quality data. Second, in both studies, not all participants were presented with all non-DD related data quality measures, which resulted in some participants being flagged as good quality responders in a more lenient manner than others (e.g., some participants were evaluated for consistency based on 2 items vs. 4). Third, Study 1 included participants from all phases of data collection for the parent study, across which study procedures differed (number of DD tasks, certain experimental groups present/ absent). However, to reduce the influence of these factors in analyses, study parameters were controlled for to the extent possible (e.g., group assignment, DD task order). Finally, our studies had lower percentages of nonsystematic DD data when compared to less restrictive inclusion/screening criteria (Yeh et al., 2022, Craft et al., 2022, Naude et al., 2021), which may have reduced our statistical power to detect associations between performance on the DD attention checks and the data quality measures herein. Additional research evaluating their utility under different recruitment contexts/screening procedures is needed.

Conclusions

Based on secondary analyses using two datasets, the commonly used attention checks in DD tasks do not appear to be a valid indicator of poor quality DD data on their own or for data obtained more broadly. Additional research is needed to determine if these types of checks may be biased, as we found some support that they may differentially flag participants’ data based on individual characteristics. As such, caution is warranted in solely using failure of these types of attention checks for excluding data—DD or otherwise—from analysis. Because more and more studies are recruiting participants on crowdsourcing platforms, additional research on the validity of the DD attention checks specifically, and on other quality checks in general is warranted to ensure data quality remains high. While our findings and recommendations are directly related to online research, they may also be relevant to any laboratory or clinical settings where these measures are used as DD tasks are informative and highly used tools in the domain of substance use research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Disclosures and Acknowledgements

MSB gratefully acknowledges that her time was supported in part by the National Institute on Drug Abuse (NIDA) grant K01DA052673. JMR's time was supported in part by the Center for Translational Science Training to Reduce the Impact of Alcohol on HIV Infection under award T32 AA025877 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA); and award K99 AA029732 (NIAAA). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDA or NIAAA.

All authors contributed in a significant way to the manuscript and that all authors have read and approved the final manuscript.

The authors declare no real or potential conflict(s) of interest, including financial, personal, or other relationships with other organizations or pharmaceutical/biomedical companies that may inappropriately impact or influence the research and interpretation of the findings.

The findings of this manuscript have not been previously published. However, portions of this work were presented at the 2022 annual meeting of the Society for the Quantitative Analyses of Behavior (SQAB).

References

Aguinis H, Villamor I, & Ramani RS (2021). MTurk research: Review and recommendations. Journal of Management, 47(4), 823–837. 10.1177/0149206320969787

Ainslie G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. Psychological Bulletin, 82(4), 463–496. 10.1037/h0076860 [PubMed: 1099599]

Almog S, Vásquez Ferreiro A, Berry MS, & Rung JM (2022). Are the attention checks embedded in delay discounting tasks a valid marker for data quality? https://osf.io/ax7dp/

Aluja A, & Blanch A (2011). Neuropsychological behavioral inhibition system (BIS) and behavioral approach system (BAS) assessment: A shortened Sensitivity to Punishment and Sensitivity to Reward questionnaire version (SPSRQ–20). Journal of Personality Assessment, 93(6), 628–636. 10.1080/00223891.2011.608760 [PubMed: 21999386]

Athamneh LN, DeHart WB, Pope D, Mellis AM, Snider SE, Kaplan BA, & Bickel WK (2019). The phenotype of recovery III: Delay discounting predicts abstinence self-efficacy among individuals in recovery from substance use disorders. Psychology of Addictive Behaviors, 33(3), 310–317. 10.1037/adb0000460 [PubMed: 30896193]

Athamneh LN, Stein MD, Lin EH, Stein JS, Mellis AM, Gatchalian KM, Epstein LH, & Bickel WK (2020). Setting a goal could help you control: Comparing the effect of health goal versus general episodic future thinking on health behaviors among cigarette smokers and obese individuals. Experimental and Clinical Psychopharmacology, 29(1), 59–72. 10.1037/pha0000351 [PubMed: 32191071]

Barends AJ, & de Vries RE (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. Personality and Individual Differences, 143, 84–89. 10.1016/j.paid.2019.02.015

Berinsky AJ, Margolis MF, & Sances MW (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. American Journal of Political Science, 58(3), 739–753. 10.1111/ajps.12081

Bickel WK, Athamneh LN, Basso JC, Mellis AM, DeHart WB, Craft WH, & Pope D (2019). Excessive discounting of delayed reinforcers as a trans-disease process: Update on the state of the science. Current Opinion in Psychology, 30, 59–64. 10.1016/j.copsyc.2019.01.005 [PubMed: 30852411]

Bickel WK, Koffarnus MN, Moody L, & Wilson AG (2014). The behavioral- and neuro-economic process of temporal discounting: A candidate behavioral marker of addiction. Neuropharmacology, 76, 518–527. 10.1016/j.neuropharm.2013.06.013 [PubMed: 23806805]

Buhrmester MD, Talaifar S, & Gosling SD (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. Perspectives on Psychological Science, 13(2), 149–154. 10.1177/1745691617706516 [PubMed: 29928846]

Craft WH, Tegge AN, & Bickel WK (2022). Narrative theory IV: Within-subject effects of active and control scarcity narratives on delay discounting in alcohol use disorder. Experimental and Clinical Psychopharmacology, 30(5), 500–506. 10.1037/pha0000478 [PubMed: 34166034]

Craft WH, Tegge AN, Freitas-Lemos R, Tomlinson DC, & Bickel WK (2022). Are poor quality data just random responses?: A crowdsourced study of delay discounting in alcohol use disorder. Experimental and Clinical Psychopharmacology, 30(4), 409–414. 10.1037/pha0000549 [PubMed: 35175071]

Cribari-Neto F, & Zeileis A (2010). Beta regression in R. Journal of Statistical Software, 34(2), 1–24. 10.18637/jss.v034.i02.

Diedenhofen B, & Musch J (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. International Journal of Internet Science, 11(1), 51–60.

Feldt LS, Woodruff DJ, & Salih FA (1987). Statistical inference for coefficient alpha. Applied Psychological Measurement, 11(1), 93–103. 10.1177/014662168701100107

Ferrari S, & Cribari-Neto F (2004). Beta regression for modelling rates and proportions. Journal of Applied Statistics, 31(7), 799–815. 10.1080/0266476042000214501

Hosmer DW, Lemeshow S, & Sturdivant RX (2013). Applied logistic regression (Third edition). Wiley.

Johnson MW, & Bickel WK (2008). An algorithm for identifying nonsystematic delay-discounting data. Experimental and Clinical Psychopharmacology, 16(3), 264–274. 10.1037/1064-1297.16.3.264 [PubMed: 18540786]

Jones A, Earnest J, Adam M, Clarke R, Yates J, & Pennington CR (2022). Careless responding in crowdsourced alcohol research: A systematic review and meta-analysis of practices and prevalence. Experimental and Clinical Psychopharmacology, 30(4), 381–399. 10.1037/pha0000546 [PubMed: 35130007]

Kassambara A (2021). Pipe-Friendly Framework for Basic Statistical Tests (Version 2021) [Computer software].

Kinderman P, Schwannauer M, Pontin E, & Tai S (2011). The development and validation of a general measure of well-being: The BBC well-being scale. Quality of Life Research, 20(7), 1035–1042. 10.1007/s11136-010-9841-z [PubMed: 21243528]

MacInnis CC, Boss HCD, & Bourdage JS (2020). More evidence of participant misrepresentation on Mturk and investigating who misrepresents. Personality and Individual Differences, 152, 109603. 10.1016/j.paid.2019.109603

MacKillop J, Amlung MT, Few LR, Ray LA, Sweet LH, & Munafò MR (2011). Delayed reward discounting and addictive behavior: A meta-analysis. Psychopharmacology, 216(3), 305–321. 10.1007/s00213-011-2229-0 [PubMed: 21373791]

Mellis AM, Woodford AE, Stein JS, & Bickel WK (2017). A second type of magnitude effect: Reinforcer magnitude differentiates delay discounting between substance users and controls. Journal of the Experimental Analysis of Behavior, 107(1), 151–160. 10.1002/jeab.235 [PubMed: 28101922]

Myerson J, Green L, & Warusawitharana M (2001). Area under the curve as a measure of discounting. Journal of the Experimental Analysis of Behavior, 76(2), 235–243. 10.1901/jeab.2001.76-235 [PubMed: 11599641]

Naudé GP, Dolan SB, Strickland JC, Berry MS, Cox DJ, & Johnson MW (2021). The influence of episodic future thinking and graphic warning labels on delay discounting and cigarette demand.

International Journal of Environmental Research and Public Health, 18(23), 12637. [PubMed: 34886370]

Newman A, Bavik YL, Mount M, & Shao B (2021). Data collection via online platforms: Challenges and recommendations for future research. Applied Psychology, 70(3), 1380–1402. 10.1111/apps.12302

Nichols A, & Edlund J (2020). Why don't we care more about carelessness? Understanding the causes and consequences of careless participants. International Journal of Social Research Methodology, 23, 1–14. 10.1080/13645579.2020.1719618

Peck S, Rung JM, Hinnenkamp JE, & Madden GJ (2020). Reducing impulsive choice: VI. Delay-exposure training reduces aversion to delay-signaling stimuli. Psychology of Addictive Behaviors, 34(1), 147–155. 10.1037/adb0000495 [PubMed: 31343195]

Pontin E, Schwannauer M, Tai S, & Kinderman P (2013). A UK validation of a general measure of subjective well-being: The modified BBC subjective well-being scale (BBC-SWB). Health and Quality of Life Outcomes, 11(1), 150. 10.1186/1477-7525-11-150 [PubMed: 24004726]

Pope DA, Poe L, Stein JS, Snider SE, Bianco AG, & Bickel WK (2019). Past and future preference reversals are predicted by delay discounting in smokers and non-smokers. Experimental and Clinical Psychopharmacology, 27(1), 19–28. 10.1037/pha0000224 [PubMed: 30382730]

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revelle W (2022). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.2.3, https://CRAN.R-project.org/package=psych

R Studio Team (2019). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. http://www.rstudio.com/

Rung JM, Almog S, Ferreiro AV, & Berry MS (2022). Using crowdsourcing for alcohol and nicotine use research: Prevalence, data quality, and attrition on Amazon Mechanical Turk. Substance Use & Misuse, 57(6), 857–866. 10.1080/10826084.2022.2046096 [PubMed: 35258409]

Rung JM, Johnson PS, & Madden GJ (2018). Differential relations between delay discounting and distress tolerance as a function of opportunity cost and alcohol use. Experimental and Clinical Psychopharmacology, 26(3), 278–289. 10.1037/pha0000198 [PubMed: 29863384]

Rung JM, & Madden GJ (2018). Experimental reductions of delay discounting and impulsive choice: A systematic review and meta-analysis. Journal of Experimental Psychology: General, 147(9), 1349–1381. https://psycnet.apa.org/doi/10.1037/xge0000462 [PubMed: 30148386]

Rung JM, & Madden GJ (2019). Demand characteristics in episodic future thinking II: The role of cues and cue content in changing delay discounting. Experimental and Clinical Psychopharmacology, 27(5), 482. 10.1037/pha0000260 [PubMed: 30762382]

Rung JM, Peck S, Hinnenkamp JE, Preston E, & Madden GJ (2019). Changing delay discounting and impulsive choice: Implications for addictions, prevention, and human health. Perspectives on Behavior Science, 42(3), 397–417. Doi: 10.1007/s40614-019-00200-7 [PubMed: 31650104]

Sing T, Sander O, Beerenwinkel N, & Lengauer T (2005). ROCR: Visualizing classifier performance in R. Bioinformatics, 21(20), 3940–3941. 10.1093/bioinformatics/bti623 [PubMed: 16096348]

Stein JS, Heckman BW, Pope DA, Perry ES, Fong GT, Cummings KM, & Bickel WK (2018). Delay discounting and e-cigarette use: An investigation in current, former, and never cigarette smokers. Drug and Alcohol Dependence, 191, 165–173. 10.1016/j.drugalcdep.2018.06.034 [PubMed: 30121475]

Strathman A, Gleicher F, Boninger DS, & Edwards CS (1994). The consideration of future consequences: Weighing immediate and distant outcomes of behavior. Journal of Personality and Social Psychology, 66(4), 742–752. 10.1037/0022-3514.66.4.742

Strickland JC, Amlung M, & Reed DD (2022). Crowdsourcing methods in addiction science: Emerging research and best practices. Experimental and Clinical Psychopharmacology, 30(4), 379–380. 10.1037/pha0000582 [PubMed: 35862134]

Strickland JC, Lile JA, & Stoops WW (2019). Evaluating non-medical prescription opioid demand using commodity purchase tasks: test-retest reliability and incremental validity. Psychopharmacology, 236(9), 2641–2652. 10.1007/s00213-019-05234-y [PubMed: 30927021]

Sze YY, Stein JS, Bickel WK, Paluch RA, & Epstein LH (2017). Bleak present, bright future: Online episodic future thinking, scarcity, delay discounting, and food demand. Clinical Psychological Science, 5(4), 683–697. 10.1177/2167702617696511 [PubMed: 28966885]

Torrubia R, Ávila C, Moltó J, & Caseras X (2001). The Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ) as a measure of Gray's anxiety and impulsivity dimensions. Personality and Individual Differences, 31(6), 837–862. 10.1016/S0191-8869(00)00183-5

Watson D, Clark LA, & Tellegen A (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54(6), 1063–1070. 10.1037/0022-3514.54.6.1063 [PubMed: 3397865]

Xu Y, Towe SL, Causey ST, Dennis PA, & Meade CS (2022). Effects of substance use on monetary delay discounting among people who use stimulants with and without HIV: An ecological momentary assessment study. Experimental and Clinical Psychopharmacology, 30(1), 39–50. 10.1037/pha0000423 [PubMed: 32757596]

Yeh YH, Freitas-Lemos R, Craft WH, & Bickel WK (2022). The relationship between nonsystematic delay discounting and low-quality survey responses in a sample of smokers: ROC curve analysis. Experimental and Clinical Psychopharmacology. Advance online publication. 10.1037/pha0000584

Public Health Significance

The present study showed that a commonly used method to determine data quality from a behavioral economic task of impulsive decision-making (delay discounting) is not adequate to discriminate poor from good quality data on its own. These results suggest that researchers should determine data quality using multiple methods and highlight the need to continue refining data screening practices for delay discounting tasks which are highly used in and relevant to addiction science.

Table 1.

Demographic characteristics for the full Study 1 sample and as a function of performance on the delay discounting attention checks (*n* and %; or median and first and third quartiles).

	DD Attention Check Performance			
Variable	**Full Sample** ***N* = 270**	**Failed Any** ***n* = 46**	**Passed Both** ***n* = 224**	***p***
Age	33 (28 – 40)	35 (30 – 40)	33 (28 – 39)	0.482
Sex				0.410
Male	136 (50.4%)	22 (47.8 %)	114 (50.9 %)	
Female	132 (48.9%)	23 (50.0 %)	109 (48.7 %)	
N/A	2 (0.7%)	1 (2.2 %)	1 (0.4 %)	
Race				0.088
White	204 (75.6%)	33 (71.7 %)	171 (76.3 %)	
Black/African American	21 (7.8%)	4 (8.7 %)	17 (7.6 %)	
Asian	19 (7.0%)	1 (2.2%)	18 (8.0%)	
Other	12 (4.4%)	2 (4.4%)	10 (4.5%)	
N/A	14 (5.2%)	6 (13.0%)	8 (3.6%)	
Ethnicity				0.014
Hispanic/LatinX	21 (7.8%)	8 (17.4 %)	13 (5.8 %)	
Not Hispanic/LatinX	249 (92.2%)	38 (82.6 %)	211 (94.2 %)	
Yearly income ($10K)	42 (25 – 65)	51 (28 – 65)	41 (25 – 65)	0.494
Years of education	16 (14 – 16)	16 (14 – 16)	16 (14 – 16)	0.524

Note. Age is rounded to the nearest year.

Table 2.

Coefficient estimates, test statistics, and effect sizes for logistic regressions evaluating the association between failure of the discounting task attention checks and nonsystematic delay discounting data for Studies 1 and 2.

Predictor	Coefficient	Std. Error	*z*	*p*	*OR* [95% CI]
Study 1					
Intercept	−2.65	0.57	−4.61	<.001	0.70 [0.02, 0.21]
Failing Any DD Attention Checks	1.19	0.47	2.51	.012	3.27 [1.26, 8.19]
DD Duration	0.09	0.08	1.08	.280	1.09 [0.91, 1.27]
DD Task Order [Second]	−0.19	0.52	−0.36	.720	0.83 [0.28, 2.25]
DD Task Order [Third]	−0.47	0.53	−0.89	.376	0.62 [0.20, 1.71]
Group [Experimental Group 1]	−0.34	0.58	−0.59	.559	0.71 [0.21, 2.13]
Group [Experimental Group 2]	0.12	0.48	0.24	.808	1.12 [0.43, 2.93]
Study 2					
Intercept	−2.07	0.31	−6.71	<.001	0.13 [0.07, 0.24]
Failing Any DD Attention Checks	1.34	0.36	3.76	<.001	3.82 [1.87, 7.62]
DD Duration	−0.02	0.06	−0.30	0.767	0.98 [0.85, 1.09]

Note. DD = Delay Discounting. The intercept coefficient for Study 1 reflects those who passed both attention checks, completed the monetary discounting task first, and were assigned to the control group in the primary study from which data were drawn. The intercept coefficient for Study 2 reflects those who passed both attention checks.

Table 3.

Coefficient estimates and test statistics for the beta regression evaluating the association between failure of the DD attention checks and degree of discounting in Study 1.

Predictor	**Coefficient**	**Std. Error**	***z***	***p***
Intercept	−0.08	0.26	−0.32	0.750
Failing Any DD Attention Checks	0.39	0.17	2.32	0.020
Systematic DD Data [Yes]	−0.68	0.21	−3.25	0.001
DD Duration	−0.02	0.03	−0.52	0.603
DD Task Order [Second]	0.40	0.16	2.57	0.010
DD Task Order [Third]	0.08	0.15	0.57	0.567
Group [Experimental Group 1]	−0.23	0.16	−1.45	0.147
Group [Experimental Group 2]	−0.23	0.15	−1.58	0.113

Note. DD = Delay Discounting. The intercept coefficient reflects those who passed both attention checks, completed the monetary DD task first, and were assigned to the control group in the primary study from which data were drawn.

Table 4.

Demographic characteristics for the full Study 2 sample and as a function of performance on the delay discounting attention checks (*n* and %; or median and first and third quartiles).

	DD Attention Check Performance			
Variable	**Full Sample** ***N* = 430**	**Failed Any** ***n* = 48**	**Passed Both** ***n* = 382**	***p***
Age	38 (31 – 48)	36 (31 - 41)	38 (31 – 49)	0.122
Sex				0.065
Male	214 (49.8%)	30 (62.5 %)	184 (48.2 %)	
Female	213 (49.5%)	17 (35.4 %)	196 (51.3 %)	
N/A	3 (0.7%)	1 (2.1 %)	2 (0.5 %)	
Race				0.350
White	335 (77.9%)	37 (77.1 %)	298 (78 %)	
Black/African American	32 (7.4%)	6 (12.5 %)	26 (6.8 %)	
Asian	36 (8.4%)	4 (8.3%)	32 (8.4%)	
Other	27 (6.3%)	1 (2.1%)	26 (6.8%)	
N/A	0 (0 %)	0 (0 %)	0 (0 %)	
Ethnicity				0.436
Hispanic/LatinX	41 (9.5%)	6 (12.5 %)	35 (9.2 %)	
Not Hispanic/LatinX	389 (90.5%)	42 (87.5 %)	347 (90.8 %)	
N/A	0 (0 %)	0 (0 %)	0 (0 %)	
Yearly Income				0.678
$29,999 or less	130 (30.2%)	10 (20.8%)	120 (31.4%)	
$30,000 - $59,999	150 (34.9%)	19 (39.6%)	131 (34.3%)	
$60,000-$89.999	68 (15.8%)	8 (16.7%)	60 (15.7%)	
$90,000-$119,999	40 (9.3%)	6 (12.5%)	34 (8.9%)	
$120,000 or more	31 (7.2%)	4 (8.3%)	27 (7.1%)	
N/A	11 (2.6%)	1 (2.1%)	10 (2.6%)	
Education (Highest Degree)				0.144
Less than high school	4 (0.9%)	0 (0%)	4 (1.0%)	
High school	49 (11.4%)	2 (4.2%)	47 (12.3%)	
Some college (no degree)	77 (17.9%)	4 (8.3%)	73 (19.1%)	
Associate	51 (11.9%)	7 (14.6%)	44 (11.5%)	
Bachelor's	182 (42.3%)	24 (50.0%)	158 (41.4%)	
Master's	58 (13.5%)	9 (18.8%)	49 (12.8%)	
Doctoral	8 (1.9%)	2 (4.2%)	6 (1.6%)	
Professional (JD, MD)	1 (0.2%)	0 (0%)	1 (0.3%)	
N/A	0 (0%)	0 (0%)	0 (0%)	

Note. Age is rounded to the nearest year. Descriptive statistics for income are shown in 30K categories for brevity, but were assessed in $15K-increment categories.