Can Carelessness Be Captured? Assessing Careless Responding in Attitudes Toward Novel Stimuli

Article *in* Social Cognition · October 2019

DOI: 10.1521/soco.2019.37.5.468

6 authors, including:

Brian A. O'Shea
Harvard University
8 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Liz Redford
University of Florida
18 PUBLICATIONS **181** CITATIONS
SEE PROFILE

Gabrielle Pogge
University of Florida
8 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Richard A. Klein
Université Grenoble Alpes
21 PUBLICATIONS **810** CITATIONS
SEE PROFILE

Abstract

Detecting careless responding has the potential to improve the quality of data obtained from research participants. In three samples (*Ns* = 570, 602, 210), we used multiple indices of careless responding to predict the strength of implicit and explicit attitudes formed toward novel social groups as well as error rates on an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998). We tested each measure of careless responding on its own, simultaneously with other predictors, and with Bayesian analyses. In three samples, there were strong and consistent effects such that more careful participants made fewer overall errors on the IAT; however, careless responding did not consistently predict implicit and explicit attitudes formed toward novel social groups. These results suggest that caution should be exercised when removing participants based on indices of careless responding.

Word Count: 133
Keywords: Participant carelessness, Attitude formation, Attention, Implicit Attitudes

Participant responses that are made in a careless, inattentive, or random fashion can threaten the quality of data in research (Bowling, Huang, Bragg, Khazon, Liu, & Blackmore, 2016; Clifford & Jerit, 2014). But what kind of responses constitute high quality data? What methods can researchers rely on to identify and control for careless or otherwise inattentive responding? Determining which features of participant quality matter (e.g., carefulness, attention) is important for researchers aiming to collect high-quality data. The current research is designed to test the effectiveness of several data quality indices to capture careless responding in the context of attitude formation toward novel stimuli.

Attitude Formation

Because attitudes are formed based on attention to relevant information (e.g., negative attitudes toward a target form after exposure to negative information about the target), the current research focused on participant attention and careless responding in the context of attitude formation. We explored the predictive utility of several indices of careless responding in predicting the strength of implicit and explicit attitudes formed toward novel social groups as well as their utility in predicting error rates on an Implicit Association Test (IAT; Greenwald, McGhee, & Schwartz, 1998).

Attitudes are psychological tendencies expressed by evaluating a given entity with some degree of favor or disfavor (Eagly & Chaiken, 1993) or associations between attitude objects (e.g., a social group) and summary evaluations (e.g., negativity; Fazio, 2007). Because attitudes can play a significant role in predicting behavior (Fazio, 1990), researchers have paid considerable attention to understanding the processes by which attitudes form.

One method for examining attitude formation is inducing attitudes toward novel individuals or groups (Ranganath & Nosek, 2008; Ratliff & Nosek, 2011; Rydell, McConnell,

Mackie, & Strain, 2006; Rydell, McConnell, Strain, Claypool, & Hugenburg, 2007). In one type of attitude induction procedure (Hamilton & Gifford, 1976; Ranganath & Nosek, 2008; Ratliff & Nosek, 2011), participants read a series of positive and negative traits or behaviors that describe members of two hypothetical groups. Each name-behavior or name-trait pair appears on the screen for several seconds before the next automatically appears. This attitude induction procedure reveals that even small amounts of information produces measurable attitudes toward a target -- both explicitly (i.e., attitudes measured by direct measures such as self-report) and implicitly via indirect measures (such as the IAT; Greenwald, McGhee, & Schwartz, 1998; Gawronski, Hofmann & Wilbur, 2006).

Researchers frequently use this attitude induction paradigm (e.g., Chen & Ratliff, 2015; Hawkins & Ratliff, 2014; Ranganath & Nosek, 2008; Ratliff & Nosek, 2010; Ratliff & Nosek, 2011; Ratliff, Swinkels, Klerx, & Nosek, 2012; Rydell et al., 2006; Rydell et al., 2007). However, because the success of the attitude induction relies on participants' receipt of the positive and negative information presented, inattentiveness during the induction procedure may lead participants to fail to form attitudes. Thus, the current work focuses on understanding whether participant carelessness influences the formation of attitudes toward novel social groups following an attitude induction paradigm.

What is Careless Responding?

Researchers have increasingly drawn attention to carelessness as a source of inaccuracy in self-report data (Huang, Curran, Keeney, Poposki, & DeShon, 2012; Maniaci & Rogge, 2014; Meade & Craig, 2012). But it is unclear whether and when "careless" responses alter results or data quality for non-self-report outcomes. For example, Berinsky et al. (2016) found that they were able to induce participants to improve their performance on manipulation checks but that

the same incentives did not improve the quality of the responses participants provided. Conversely, Oppenheimer, Meyvis, & Davidenko (2009) found that removing participants who did not follow instructions (instructional manipulation check; IMC) increased statistical power and reliability in a dataset. Including participants who fail an IMC reduces the likelihood of finding statistically significant differences between Amazon's Mechanical Turk (MTurk) participants and college students on other dimensions (i.e., rates of passing other attention checks; Goodman et al., 2013). In the context of attitude formation, this could mean a lack of differences between groups exposed to attitude inductions and control groups.

One reason that identifying careless responding is difficult is because researchers can define "carelessness" in several ways. One strategy is to use "screeners" that differentiate between participants who respond carelessly versus carefully. Research shows that screeners—such as IMCs (Oppenheimer et al., 2009)—can reveal careless responding and reduce noise. Studies using screeners can balance the goals of internal and external validity by presenting results conditional on different levels of attention (Berinsky, Margolis, & Sances, 2014). Yet, other researchers recommend using multiple items to measure attention (Maniaci & Rogge, 2014).

Other methods to identify careless responding include using multiple response-qualifying questions that include bogus or very low-incidence behaviors (e.g., "I get paid bi-weekly by leprechauns."; Meade & Craig, 2012). Researchers have also included consistency checks by asking essentially the same question but in two separate ways (e.g., positive versus negative wording) in separate parts of the questionnaire (e.g., Schmitt & Stults, 1985). Assessing self-reported attention (Meade & Craig, 2012) and identifying elevated levels of item nonresponse (Baker & Downes-Le Guin, 2007) offer other measures to detect careless responding.

Careless responding has consequences, including satisficing (e.g., spending less time reading vignettes; Krosnick, 1991; Krosnick, 1999; Miura & Kobayashi, 2016), stereotypical patterns of responding (e.g., using nationality-based stereotypes when making estimates of immigrants' personality traits; Miura & Kobayashi, 2016), and inconsistent responses (Gao, House, & Bi, 2016).

Careless Responding and Attitudes

Research has identified that a participant's willingness or ability to attend to information is an important contextual factor in attitude formation and change. For example, the Elaboration Likelihood Model (ELM; Petty & Cacioppo, 1986) posits distinctive styles of information processing depending on attentional resources. A person who is either unwilling or unable to attend to information may process information very differently than a person who is more motivated to do so (Petty & Cacioppo, 1986). Participants who are more motivated to attend to information may process it more deeply and form more durable attitudes that are better predictors of behavior when compared to participants who are less motivated (Haugtvedt & Petty, 1989; Petty, Haugtvedt & Smith, 1995).

The same factors that contribute to careless responding are also implicated in attitude formation and change. This poses a special problem for attitude research. The same participants that are inclined to respond carelessly may also be forming differentially strong attitudes. Further, if participants show weaker or less reliable attitudes it might be because of careless responses, not that their actual attitudes differ. If the same processes that contribute to attitude formation should also contribute to careless responding, then the same indices that are used to identify careless responding should also have implications for identifying people who form stronger and weaker attitudes.

The Current Research

The current research tested how various metrics predict implicit and explicit attitude formation toward novel stimuli. We chose to test the metrics in an attitude formation paradigm for several reasons. First, attitude formation is a reliable phenomenon with known standards for parameters (e.g., attitude strength, error rates, etc.; Ranganath & Nosek, 2008). Second, standard indicators of careless responding (e.g., IMCs) may not have significant implications for the strength of attitudes formed or measures of automatic processes. People often form impressions automatically, with little or no effort or with only very little information (e.g., Winter & Uleman, 1984).

The current research assessed careless responding via six indicators: an IMC, self-reported attention, self-reported data retention, error rates on the IAT, response time during the experiment, and a bogus item. In three samples we assessed the ability of each indicator to predict known standards for parameters (e.g., attitude strength, error rates, etc.) within an attitude formation paradigm. Participants were induced to form positive or negative attitudes toward novel social groups. We assessed the relative strength of each of the indicators of careless responding in predicting responses to this attitude induction procedure.

Study 1

Method

Participants

All participants volunteered at the Project Implicit website (https://implicit.harvard.edu; Nosek, 2005).

Sample A. 590 participants completed all materials and were included in the final sample (63.7% women, 76.1% White, M_{age} = 41.7 years, SD = 14.8). We chose this sample size based

on an *a priori* decision to collect data from 500 participants. Because studies at Project Implicit do not get taken down from the site immediately upon request, the sample size is slightly larger than that. This sample size provides 91% power to detect an effect size of $d = 0.30$ for the attitude induction procedure (i.e., two-sample t-test demonstrating an implicit or explicit preference for one group over another).

Sample B. 602 participants completed all materials and were included in the final sample (64.9% women, 75.6% White, $M_{age} = 35.9$ years, $SD = 14.8$). We chose this sample size based on an *a priori* decision to collect data from 600 participants. This sample size gives us 96% power to detect an effect size of $d = 0.30$ for the attitude induction procedure (i.e., two-sample t-test demonstrating an implicit or explicit preference for one group over another).

We collected Sample B for replication purposes and to compare the results of Sample A. For both samples, statistical power for the individual participant quality markers predicting attitudes differs based on how many people "failed" each check. Further discussion of this issue is provided in the results section.

Materials

Attitude induction manipulation. The primary manipulation was an attitude induction procedure designed to make participants more favorable to one fictitious social group over a comparison fictitious social group. To induce these attitudes, participants read a series of positive and negative behaviors and traits describing members of two fictitious social groups – Niffians and Laapians (adapted from Gregg, Seibt & Banaji, 2006; Ratliff & Nosek, 2010). All Niffians had names that end with "nif" (e.g., Vabbenif, Ibonnif) and all Laapians had names that end in "lap" (e.g., Reemolap, Bosaalap). Participants saw each name-behavior pair presented on the

screen for four seconds before the next automatically appeared. See online supplement (http://bit.ly/2gPA1fE) for a full list of names and traits used in the attitude induction procedure.

In the two *Niffian-Positive* conditions, participants read either: (a) 16 positive & 4 negative statements about Niffians and 16 negative & 4 positive statements about Laapians, or (b) 8 positive & 2 negative statements about Niffians and 8 negative & 2 positive statements about Laapians. In the two *Laapian-Positive* conditions, participants read either: (a) 16 positive & 4 negative statements about Laapians and 16 negative & 4 positive statements about Niffians, or (b) 8 positive & 2 negative statements about Laapians and 8 negative & 2 positive statements about Niffians. Because the induction length manipulation (whether they read more or less information about the groups) did not produce large effects in this study, and produced no effects in the replication study, we report the results here collapsed across the two length manipulation conditions. Due to a programming error, the *Niffian-Positive* conditions were not included in Sample B.

Attitude formation outcomes.

Explicit attitudes toward the target groups. We measured attitudes toward the two fictitious groups (Niffians and Laapians) were measured on six evaluative dimensions: unpleasant/pleasant, unfriendly/friendly, unlikeable/ likeable, unpopular/popular, bad/good, and unkind/kind. Responses were on 7-point scales (-3 = *Very unpleasant*; +3 = *Very pleasant*). We averaged participants' responses across the six evaluative dimensions to create a composite score. Then we computed a difference score such that higher scores indicate a preference for one group relative to the other based on which induction condition participants were assigned to.

Implicit attitudes toward the target groups. We measured implicit attitudes toward the two fictitious groups using an IAT (Greenwald et al., 1998). The IAT measures the degree to

which participants associate two target concepts (e.g., “Laapians” and “Niffians”) and two attributes (e.g., “good words” and “bad words”). The stimuli used to represent the “Niffians” and “Laapians” categories were the names of group members from the attitude induction. Stimuli representing “good words” were *nice, heaven, happy, pleasure*; stimuli representing “bad words” were *nasty, hell, horrible, rotten*. The IAT comprised seven trial blocks as recommended by Nosek, Greenwald, and Banaji (2005). A positive *D* score indicates a stronger association between Laapians + good word/Niffians + bad words relative to the reverse (i.e., a stronger implicit attitude). Typically, data from participants who have too-high error rates (40% on any given block or 30% overall) are excluded. In Sample A 39 (7.0%) participants met these criteria and in Sample B 60 (10.0%) participants met these criteria; however, because we wanted to look at IAT error rates as a potential indicator—and outcome—of participant quality, we left participants with too-high error rates in the dataset and will discuss this issue further in the results section.

Participant quality indicators.

Instructional manipulation check. Participants completed an instructional manipulation check (IMC; Oppenheimer et al., 2009). In this task, participants were presented with a large block of text with a textbox at the bottom. The question, in bolded text, immediately above the textbox was “What city were you born in?”; however, the larger block of instructions earlier on the page instructed participants to ignore the question in bold and instead to write “Dog” in the text box. We accepted only “Dog” and “dog” as correct responses; we considered anything else a failure of the IMC. 409 participants (70%) passed the instructional manipulation check in Sample A and 375 participants (62%) passed the instructional manipulation check in Sample B.

Self-reported data retention. Participants were asked to respond with a "yes" or "no" to the following question: "In your honest opinion, do you think we should use your responses in our research?" In sample A, 448 participants (76%) indicated that their responses should be included in the analysis, and in Sample B 444 participants (75%) indicated that their responses should be included in the analysis.

Self-reported attention. Participants responded to a single attention item "To what extent were you paying attention during the study?" on a 5-point scale with the following response options: *Not at all*, *A little*, *A moderate amount*, *Very much*, *Extremely* (Sample A: $M = 3.85$, $SD = 0.80$; Sample B: $M = 3.75$, $SD = 0.86$). Higher scores reflect greater self-reported attention.

"Bogus" scale item. Participants completed the 22-item Williams and Eberhardt's (2008) Race Conceptions Scale (Sample A: $\alpha = .83$; Sample B: $\alpha = .80$), which is designed to measure the extent to which people see race as being biologically determined (i.e., essentialist thinking; sample item: "Siblings born to the same parents will always be of the same race as each other."). We embedded a "bogus item" from Meade & Craig (2012) in the scale: "I am paid bi-weekly by leprechauns." Participants responded using a seven-point scale ranging from *Strongly Disagree* to *Strongly Agree* (Sample A: $M = 1.48$, $SD = 1.34$; Sample B: $M = 1.45$, $SD = 1.17$). We categorized as failing the attention check all participants who chose a response other than "Strongly disagree" or "Disagree". 492 participants (82%) passed this attention check in Sample A and 494 (82%) passed this check in Sample B.

Response time. We measured the amount of time each participant spent on each question following the attitude induction (in milliseconds) and averaged those measures into a composite. We then log-transformed the composite measure (Sample A: $M = 3.66$, $SD = 0.18$; Sample B: $M = 3.94$, $SD = 0.19$).

Procedure

Participants volunteered through the Project Implicit website. Upon arriving at the site, they completed demographics as part of the registration process (if a new participant) or logged into their account (if a returning participant). They were randomly assigned to this study from a pool of approximately eight studies. Participants read that the purpose of this study was to measure their attitudes and behaviors, and were then randomly assigned to one of two induction conditions (Niffians-Positive versus Laapians-Positive). Following the attitude induction procedure, participants completed explicit attitude measures, an IAT measuring implicit preference for Niffians vs Lappians, and participant quality measures in randomized order. After beginning this study, they could not be assigned to it again.

Results

Attitude Formation Results

Implicit attitudes were induced. A one-sample t-test revealed that implicit attitudes were induced. Higher scores indicate a stronger preference for the positive group relative to the negative group (i.e., collapsing across group-name counterbalancing conditions). Implicit attitudes were induced in the expected direction (Sample A: $M = 0.11$, $SD = 0.47$, $t(584) = 5.872$, $p < 0.001$, Cohen's $d = 0.24$, 95% CI = [0.07, 0.15]; Sample B: $M = 0.20$, $SD = 0.46$, $t(587) = 10.59$, $p < 0.001$, Cohen's $d = 0.44$, 95% CI = [0.16, 0.24]). That is, participants had more positive implicit attitudes toward a group described as performing predominantly positive behaviors than toward a group described as performing predominantly negative behaviors.

Explicit attitudes were induced. To simplify interpretation of these results and provide additional power for the analyses of participant quality, higher scores indicate a stronger preference for the positive group relative to the negative group (i.e., collapsing across group-

name counterbalancing conditions). A one-sample t-test revealed that explicit attitudes were induced in the expected direction (Sample A: $M = 1.03$, $SD = 1.51$, $t(569) = 16.316$, $p < 0.001$, Cohen's $d = 0.68$, 95% CI = [0.91, 1.16]; Sample B: $M = 1.23$, $SD = 1.66$, $t(578) = 17.84$, $p < 0.001$, Cohen's $d = 0.74$, 95% CI = [1.10, 1.37]). That is, participants had more positive explicit attitudes toward a group described as performing predominantly positive behaviors than toward a group described as performing predominantly negative behaviors[1].

Relations among study variables. Implicit and explicit attitudes toward the two groups correlated significantly in Sample A ($r = .24$, $p < .01$, 95% CI = [.16, .32]) and in Sample B ($r = .21$, $p < .01$, 95% CI = [.13, .29]). See Table 1 for a full breakdown of the relations between study variables in each sample and the individual relations between each predictor of participant quality and implicit and explicit attitudes.

Overall Analysis Plan

The primary test of each hypothesis was a regression in which each of the three dependent measures (implicit attitudes, explicit attitudes, overall IAT error rates) were separately predicted by one of the six indices of participant quality. A second regression also predicted each dependent measure from the simultaneous combination of each measure of participant quality (whether participants made too many errors on the IAT, whether they passed IMC, their self-reported attention, whether they felt their data should be retained, overall response time to questions, and whether they passed the bogus item check). Throughout, we report Bayes factors as our primary analysis. Null hypothesis significance testing (NHST) results are available in the online supplement. The primary reason we adopted this strategy was because NHST can only

[1]Prior to collapsing, participants reported more positive attitudes toward Laapians relative to Niffians in the Laapian-Positive condition ($M = 1.22$, $SD = 1.57$) compared to the Niffian-Positive Condition ($M = 0.86$, $SD = 1.43$), $t(568) = 2.824$, $p = < .01$, Cohen's $d = 0.237$, 95% $CI_{diff} = [0.11, 0.60]$.

estimate if variables are statistically different from one another. Bayes factors, on the other hand, can estimate the likelihood that variables are statistically similar and hence this method can test how much evidence there is in favor of the null hypothesis relative to the alternative hypothesis (Jarosz, & Wiley, 2014). We use Bayes Factor 10 (BF_{10}) throughout. See Table 2 for a description of how to interpret Bayes Factor 10. For a visual depiction of the results using NHST, please see Table 3 for a description of the results when predictors are entered individually and Table 4 for results when the predictors are entered simultaneously.

Effects of the Individual Data Quality Predictors

Instructional manipulation check predictor (binary pass/fail).

Predicting implicit attitudes. Passing versus failing the IMC did not predict implicit attitudes. The estimated Bayes factor suggested ($BF_{10} = .103$) that the data were substantially to strongly in favor of the null hypothesis.

Predicting explicit attitudes. Passing versus failing the IMC did not predict explicit attitudes. The estimated Bayes factor suggested ($BF_{10} = .095$) that the data were strongly in favor of the null hypothesis.

Predicting overall IAT error rates. Passing versus failing the IMC significantly predicted overall IAT error rates such that failing the IMC was associated with making more errors. The estimated Bayes factor suggested ($BF_{10} > 100$) that the data were very strongly in favor of the alternative hypothesis.

Self-reported data retention predictor (binary pass/fail).

Predicting implicit attitudes. Whether participants felt their data should be retained did not predict implicit attitudes. The estimated Bayes factor suggested ($BF_{10} = .309$) that the data were moderately in favor of the null hypothesis.

Predicting explicit attitudes. Whether participants felt their data should be retained did not predict explicit attitudes. The estimated Bayes factor suggested (BF_{10}= .297) that the data were moderately in favor of the null hypothesis.

Predicting overall IAT error rates. Whether participants felt their data should be retained significantly predicted overall IAT error rates such that participants who reported we should not use their data made more errors. The estimated Bayes factor suggested (BF_{10} > 100) that the data were decisively in favor of the alternative hypothesis.

Self-reported attention predictor (continuous).

Predicting implicit attitudes. Self-reported attention did not predict implicit attitudes. The estimated Bayes factor suggested (BF_{10} = .101) that the data were strongly in favor of the null hypothesis.

Predicting explicit attitudes. Self-reported attention significantly predicted explicit attitudes such that participants who reported paying more attention formed stronger explicit attitudes. The estimated Bayes factor suggested (BF_{10} = 41.74) that the data were very strongly in favor of the alternative hypothesis.

Predicting overall IAT error rates. Self-reported attention significantly predicted overall IAT error rates such that participants who reported paying more attention made fewer errors on the IAT. The estimated Bayes factor suggested (BF_{10} > 100) that the data were decisively in favor of the alternative hypothesis.

Bogus scale item predictor (binary pass/fail).

Predicting implicit attitudes. Passing the bogus item check did not predict implicit attitudes. The estimated Bayes factor suggested (BF_{10} = 0.132) that the data were strongly in favor of the null hypothesis.

Predicting explicit attitudes. Passing the bogus item check did not predict explicit attitudes. The estimated Bayes factor suggested (BF_{10} =0.96) that the data were substantially in favor of the null hypothesis.

Predicting overall IAT error rates*.* Passing the bogus item check significantly predicted overall IAT error rates, such that participants who passed the item check made fewer errors on the IAT. The estimated Bayes factor suggested (BF_{10} > 100) that the data were decisively in favor of the alternative hypothesis.

Too many IAT errors predictor (binary pass/fail).

Predicting implicit attitudes*.* Whether participants made too many errors on the IAT did not predict implicit attitudes based. The estimated Bayes factor suggested (BF_{10} = .133) that the data were substantially in favor of the null hypothesis.

Predicting explicit attitudes*.* Whether participants made too many errors on the IAT did not predict explicit attitudes. The estimated Bayes factor suggested (BF_{10} = 0.412) that the data is anecdotally in favor of the null hypothesis.

Predicting overall IAT error rates*.* We did not conduct a test using the categorical (pass/fail) measure of error rates to predict overall continuous error rates.

Response time predictor (continuous).

Predicting implicit attitudes*.* Response time did not predict implicit attitudes. The estimated Bayes factor suggested (BF_{10} = 0.14) that the data were strongly in favor of the null hypothesis.

Predicting explicit attitudes*.* Response time did not predict explicit attitudes. The estimated Bayes factor suggested (BF_{10} = 0.14) that the data were strongly in favor of the null hypothesis.

Predicting overall IAT error rates. Response time significantly predicted overall IAT error rates, such that participants who spent more time on the study made fewer errors on the IAT. The estimated Bayes factor suggested ($BF_{10} > 100$) that the data were decisively in favor of the alternative hypothesis.

Simultaneous Regression Results for Each DV

In addition to testing each predictor separately, we were also interested in how each of the indices of participant quality would function when we entered them simultaneously. For these tests we used simultaneous regression models in which we separately predicted implicit attitudes, explicit attitudes, and IAT error rates from the combination of each of the six indices (see Tables 5-6 for full regression results and the online supplement for the full text results).

Failing Multiple Participant Quality Indicators

We were also interested in differences in our outcomes based on how many indicators participants passed or failed. Compared to participants who failed zero, one, or two of the quality indicators, relatively few participants failed three or more of the data quality indicators. Because of the lower sample size for these categories we collapsed those participants who failed three or more of the data quality indicators into a single category. For the attention and response time (log-transformed) continuous variables, we classified scores one standard deviations below—or faster than/less attentive—the mean (respectively) as a fail. When predicting overall IAT error rates, we used only five indicators because we excluded as a predictor the binary pass/fail IAT error indicator.

Predicting implicit attitudes. One-way ANOVA revealed no effect of the number of quality indicators a participant failed on implicit attitude formation in Sample A, $F(3, 529) = 0.38$, $p > .250$, $\eta\rho^2 = .00$. The estimated Bayes factor suggested ($BF_{10} = 0.021$) the data were very

strongly in favor of the null hypothesis. However, examining the same analysis in Sample B, revealed a significant effect, $F(3, 526) = 4.17$, $p = .006$, $\eta\rho^2 = .02$. Post hoc comparisons using the t-test with Bonferroni's correction indicated that the mean implicit attitude score among participants who failed 3+ indicators ($M = 0.03$, $SD = 0.61$) was significantly lower ($p = .011$) than the mean implicit attitude score among participants who did not fail any of the indicators ($M = 0.27$, $SD = 0.44$). All other comparisons were non-significant ($p > .217$). Importantly, when using Bayes factor the data suggested ($BF_{10} = 2.742$) only anecdotal evidence that the more indicators a participant failed predicted weaker implicit attitude formation.

Predicting explicit attitudes. One-way ANOVA revealed no effect of the number of quality indicators a participant failed explicit attitude formation in Sample A, $F(3, 519) = 2.13$, $p > .250$, $\eta\rho^2 = .00$. The estimated Bayes factor suggested ($BF_{10} = 0.062$) the data were strongly in favor of the null hypothesis. Examining the same analysis in Sample B revealed a significant effect, $F(3, 528) = 2.83$, $p = .038$, $\eta\rho^2 = .02$, but when using Bayes, the data suggested ($BF_{10} = .399$) there was no evidence in favor of or against both the null and the alternative hypothesis.

Predicting overall IAT error rates. One-way ANOVA revealed a significant effect of the number of indicators a participant failed on overall IAT error rates in Sample A, $F(3, 529) = 76.83$, $p < .001$, $\eta\rho^2 = .30$. A Post hoc analysis using Bonferroni's correction indicated that the mean scores for participants who failed 0, 1, 2, or 3+ indicators were significantly different from each other ($p < .001$), except between those in the zero-fail condition and the single fail condition where $p = .011$. Overall, the findings indicated that the more tests a participant failed the higher their error rates were on the IAT. The estimated Bayes factor suggested ($BF_{10} > 100$) that the data were decisively in favor of the alternative hypothesis. We found comparable results in Sample B.

Statistical Power

Statistical power to detect an effect is partially a function of sample size (Cohen, 1992). To have a 95% chance to detect an effect of average size (f = .25) in a two-cell design, a researcher should plan on collecting approximately 210 participants (Faul, Lang, & Buchner, 2007). However, not all participants are necessarily going to be included in the final analysis. In addition to the above analyses, we tested the mean difference from zero for each of the dependent measures after removing participants who failed one or more indices of participant quality. Because self-reported attention is a continuous item, we excluded from analysis participants who self-reported spending greater than one standard deviation of attention below the mean of the sample. Across the two samples the evidence consistently suggests that excluding participants who failed one or more measures of participant quality did not significantly impact implicit and explicit attitudes toward the two groups (see Figure 1a and Figure 1b). However, evidence supported fluctuations in the mean percentage of errors made by participants on the IAT after excluding participants who failed one or more indices of participant quality (see Figure 1c).

Discussion

In two different samples, participants read information about two novel groups that was biased such that they should form more positive attitudes toward one of the groups relative to the other. Participants also completed a number of measures designed to identify careless responding. When used individually, none of the metrics consistently predicted the strength of attitudes formed toward the two groups. When implemented in a combined approach, the metrics also did not consistently predict the strength of attitudes formed toward the two groups. However, the metrics did consistently predict the number of overall errors participants made on

the IAT. Although these metrics did not perform well at identifying participants that formed differently strong attitudes, excluding participants who failed one or more of these metrics would have had a deleterious effect on statistical power to detect an effect if one was present.

Study 2

MTurk (See Crump, McDonnell, Gureckis, 2013 for a review) has become a popular outlet for data collection in the social sciences. Some research suggests that MTurk workers are less attentive than participants in other populations (Hauser & Schwarz, 2015). However, other research suggests that unsupervised MTurk workers perform as well or better on attention checks when compared to supervised undergraduate participants (Briones & Benham, 2017). Given the conflicting evidence regarding MTurk workers and attention, in Study 2 we aimed to replicate our results in a sample of MTurk workers.

Method

Participants

210 participants completed all materials posted to a human-intelligence task (HIT) on MTurk (53.1% women, 75.2% White, M_{age} = 34.2 years, SD = 10.7). Participants received $1.50 for completion of the task. Only participants who had completed at least 100 previous HITs with an 80% positive rating were eligible to complete the HIT. We chose this sample size based on a practical concern to collect as much data as possible while providing participants with the equivalent of the U.S. minimum wage. This sample size gives us 82% power to detect an effect size of d = 0.40 for the attitude induction procedure. We used a higher estimated effect size in this power calculation based on the results from Study 1.

Materials

Attitude induction manipulation. All participants read either: 16 positive & 4 negative statements about Laapians and 16 negative & 4 positive statements about Niffians, or 16 positive & 4 negative statements about Niffians and 16 negative & 4 positive statements about Laapians.

Attitude formation outcomes.

Explicit attitudes toward the target groups. We measured attitudes toward the two fictitious groups (Niffians and Laapians) on eight evaluative dimensions: unpleasant/pleasant, unfriendly/friendly, unlikeable/ likeable, unpopular/popular, bad/good, rude/polite, untrustworthy/trustworthy and unkind/kind. Responses were recorded and scored exactly as in Study 1.

Implicit attitudes toward the Target Groups. We measured implicit attitudes toward the target groups exactly as in Study 1. About a quarter of participants (49; 23%) were flagged as making too many errors on the IAT.

Participant quality indicators.

Instructional manipulation check. We implemented the IMC exactly as in Study 1. Most participants (205; 98.6%) passed the IMC.

Self-reported data retention. We measured self-reported data retention exactly as in Study 1. Most participants (206; 98.1%) indicated that their responses should be included in the analysis.

Self-reported attention. We measured self-reported attention exactly as in Study 1 (M = 4.78, SD = 0.57). Higher scores reflect greater self-reported attention.

Self-reported effort. We measured self-reported effort via a single item "I put forth ____ effort towards this study". Participants responded using a 5-point scale with the following

response options: *almost no*, *very little*, *some*, *quite a bit*, *a lot of* (M = 4.77, SD = 0.56). Higher scores reflect greater self-reported effort.

"Bogus" scale item. In addition to the bogus item used in Study 1, we embedded an additional "bogus item" from Meade & Craig (2012) in the scale: "I have been to every country in the world." We flagged participants as failing the attention check if they chose a response other than "Strongly disagree" or "Disagree." Most participants (198; 94.3%) passed this attention check. Additionally, most participants (199; 94.8%) passed the attention check regarding being paid bi-weekly by leprechauns.

Response time. We measured response time exactly as in Study 1 (M = 3.66, SD = 0.18).

Procedure

Participants registered on MTurk read that the purpose of this study was to measure their attitudes and behaviors, and then completed the attitude induction. Following the attitude induction procedure, participants completed the explicit attitude measures, an IAT measuring implicit preference for Niffians versus Lappians, and responded to participant quality measures (in randomized order). After beginning this study, they could not begin it again.

Results

Attitude Formation Results

Implicit attitudes were induced. A one-sample t-test revealed that implicit attitudes were induced in the expected direction (M = 0.21, SD = 0.48, $t(200)$ = 6.44, $p < 0.001$, Cohen's d = 0.45, 95% CI = [0.15, 0.29]). That is, participants had more positive implicit attitudes toward the group described as performing predominantly positive behaviors than toward the group described as performing predominantly negative behaviors. Higher scores indicate a stronger preference for Laapians relative to Niffians.

Explicit attitudes were induced. A one-sample t-test revealed that explicit attitudes were induced. To simplify the interpretation of these results and provide additional power for the analyses of participant quality, higher scores indicate a stronger preference for the positive group relative to the negative group. This results in an overall mean explicit attitude score of 1.82 (*SD* = 1.87) which differs significantly from zero, $t(209) = 14.06$, $p < .0001$, Cohen's $d = 0.97$, 95% CI = [1.56, 2.07], confirming that explicit attitudes were induced in the expected direction. That is, participants reported more positive explicit attitudes toward the group described as performing predominantly positive behavior than toward the group described as performing predominantly negative behaviors.

Relations among Study Variables

Implicit and explicit attitudes toward the two groups did not correlate significantly ($r = .14$, $p = .06$). See Table 1 for a full breakdown of the relations between study variables in each sample and the individual relations between each predictor of participant quality and implicit and explicit attitudes.

Overall Analysis Plan

We used the same analytic strategy as in Study 1. The primary test of each hypothesis was a regression in which each of the three dependent measures were separately predicted by one of the indices of participant quality. However, given the low numbers of participants who failed each binary indicator (i.e., the IMC, self-reported data retention, and bogus item measures), those measures were excluded from this analysis.

Effects of the Individual Data Quality Predictors

Self-reported attention predictor (continuous).

Predicting implicit attitudes. Self-reported attention did not predict implicit attitudes. The estimated Bayes factor suggested ($BF_{10} = 0.25$) the data were anecdotally in favor of the null hypothesis.

Predicting explicit attitudes. Self-reported attention significantly predicted explicit attitudes such that participants who reported paying more attention formed stronger explicit attitudes. The estimated Bayes factor suggested ($BF_{10} = 0.98$) that the data were anecdotally in favor of the null hypothesis.

Predicting overall IAT error rates. Self-reported attention significantly predicted overall IAT error rates such that participants who reported paying more attention made fewer errors on the IAT. The estimated Bayes factor suggested ($BF_{10} = 89.22$) the data were decisively in favor of the alternative hypothesis.

Self-reported effort predictor (continuous).

Predicting implicit attitudes. Self-reported effort did not predict implicit attitudes. The estimated Bayes factor suggested ($BF_{10} = .18$) that the data were substantively in favor of the null hypothesis.

Predicting explicit attitudes. Self-reported effort significantly predicted explicit attitudes such that participants who reported exerting more effort formed stronger explicit attitudes. The estimated Bayes factor suggested ($BF_{10} = 0.65$) the data were anecdotally in favor of the null hypothesis.

Predicting overall IAT error rates. Self-reported effort significantly predicted overall IAT error rates such that participants who reported exerting more effort made fewer errors on the

IAT. The estimated Bayes factor suggested (BF_{10} = 2374.20) the data were decisively in favor of the alternative hypothesis.

Response time predictor (continuous).

Predicting implicit attitudes. Response time did not predict implicit attitudes. The estimated Bayes factor suggested (BF_{10} = 1.02) the data were anecdotally in favor of the alternative hypothesis.

Predicting explicit attitudes. Response time significantly predicted explicit attitudes such that participants who spent more time on each self-report item formed stronger explicit attitudes. The estimated Bayes factor suggested (BF_{10} = 509.95) the data were decisively in favor of the alternative hypothesis.

Predicting overall IAT error rates. Response time significantly predicted overall IAT error rates such that participants who spent more time on each self-report item made fewer errors on the IAT. The estimated Bayes factor suggested (BF_{10} = 176.95) the data were decisively in favor of the alternative hypothesis.

Too many IAT errors predictor (binary pass/fail).

Predicting implicit attitudes. We did not conduct a test using the categorical (pass/fail) measure of error rates to predict implicit attitudes.

Predicting explicit attitudes. Whether participants made too many errors on the IAT (M = 1.10, SD = 2.16) or not (M = 2.04, SD = 1.73) significantly predicted explicit attitudes such that participants who made too many errors formed weaker explicit attitudes. The estimated Bayes factor suggested (BF_{10} = 13.43) that the data is strongly in favor of the alternative hypothesis.

Predicting overall IAT error rates. We did not conduct a test using the categorical (pass/fail) measure of error rates to predict overall continuous error rates.

Simultaneous Regression Results for Each DV

In addition to testing each predictor separately, we compared each of the indices of participant quality against each other in simultaneous regressions separately predicting each DV (implicit attitudes, explicit attitudes, IAT errors). See Tables 5-7 for full regression results.

Implicit attitudes. The combined measures of participant quality did not exert an overall effect on implicit attitudes, $F(3, 197) = 1.65$, $p = .18$.

Explicit attitudes. The combined measures of participant quality exerted a significant overall effect on explicit attitudes, $F(4, 205) = 5.84$, $p < .001$. Of the four measures, only response time predicted explicit attitudes such that spending more time on each item corresponded with stronger explicit attitudes, $b = 2.55$, $SE = .78$, $t(198) = 3.69$, $p < .01$. Self-reported attention, $b = 0.18$, $SE = .05$, $t(205) = 0.05$, $p = .54$, self-reported effort, $b = 0.17$, $SE = .29$, $t(205) = 0.61$, $p = .54$, and whether participants made too many errors on the IAT, $b = -0.48$, $SE = .32$, $t(205) = -1.50$, $p = .13$, did not predict explicit attitudes. The estimated Bayes factor of the combined model suggested that the data were very strongly in favor of the alternative hypothesis ($BF_{10} = 65.87$).

Overall IAT error rates. The combined measures of participant quality (excluding whether participants made too many errors on the IAT) exerted a significant overall effect on overall IAT error rates, $F(3, 199) = 12.80$, $p < .001$. Self-reported effort, $b = -0.10$, $SE = .04$, $t(199) = -2.67$, $p < .01$, and response time, $b = -0.33$, $SE = .09$, $t(199) = -3.78$, $p < .001$, predicted overall IAT error rates such that participants who reported expending more effort and spending more time responding also made fewer errors. The estimated Bayes factor of the combined model suggested ($BF_{10} > 100$) that the data were decisively in favor of the alternative hypothesis.

Discussion

Although we initially intended to replicate and expand on the same analysis plan as in Study 1, MTurk participants failed the measures of participant quality at such low rates that we decided to exclude several measures from the analysis. However, MTurk participants did make substantially more errors on the IAT compared with either sample collected from Project Implicit. This suggests MTurk participants (compared to Project Implicit participants) may be relatively less familiar with the measure. Although making too many errors on the IAT predicted explicit attitude formation when tested individually, this was not the case in the simultaneous regression. Of the measures we tested, response time was the most consistent predictor.

Quad Modeling Comparison

Quad modeling is a multinomial technique used on IAT data to determine response biases other than automatic processes that drive participant's responding (Conrey, Sherman, Gawronski, Hugenberg, & Groom, 2005; Sherman, 2006). It is useful for the current study because it can inform whether Project Implicit participants and MTurk participants use similar or different processes when completing the IAT[2]. Quad modeling separates out four distinct processes in IAT data using the frequency of correct and incorrect responses across both the compatible and incompatible blocks. The four processes include: (1) the association activation **(AC)** parameter which reflects the probability that an association is automatically activated by a stimulus, (2) the overcoming bias **(OB)** parameter which indicates an automatic association is activated but it is overcome in favor of more deliberate responding, (3) the discriminability/detection **(D)** parameter which shows the probability that the correct response

[2] We would like to thank a reviewer for suggesting we apply Quad modeling to the IAT data.

can be determined and lastly, and (4) the guessing **(G)** parameter which is useful for indicating when no association is activated and when no correct answer is accessible and therefore, a guess is used.

We expect these processes to be partly influenced by the degree to which participants were paying attention during attitude formation. Specifically, higher scores on the AC parameter would indicate greater attention and perhaps reflection during attitude formation. However, attitudes that are well-elaborated should result in higher AC scores. Therefore, we do not expect the relatively short attitude formation training to produce high AC scores. Likewise, we expect low OB scores because there will be less opportunity for participants to overcome a weakly held automatic bias (AC parameter). Additionally, participants may have no motivation to overcome any automatic biases they learn (Calanchini, Sherman, Klauer, & Lai, 2014) because it is relatively socially acceptable to hold and express prejudicial attitudes towards fictional groups. High scores on the D parameter would indicate participants who are more conscientious and careful when completing the IAT. For the G parameter, a score of .50 would indicate no guessing while other scores would indicate participants were not being careful and attentive both during attitude formation and when completing the IAT.

Analysis & Results

We used the Multitree (Moshagen, 2010) software package to perform Quad modeling on data from the Project Implicit and MTurk participants who completed the long attitude formation induction training. Rather than reaction times, Quad modeling uses correct and incorrect response frequencies from IAT data to estimate the four parameters.

The overall error rate for the Project Implicit and MTurk sample was 9.55% and 19.61% respectively. The chi-square for the model fit of the Project Implicit sample was 215.95, $df = 3$, p

< .001 and for the MTurk sample it was 51.06, $df = 3$, $p < .001$. Importantly, chi-square tests are dependent on samples size and therefore, high power, as in the current study, can artificially deflate model fit (Cohen, 1988). Based on the effect size of model fit between the actual data and the model's predicted data, the MTurk data was acceptable and in line with previous research when controlling for sample size ($w = .05$; e.g., Calanchini et al., 2014). However, the Project Implicit data did not fit the model as well, $w = .09$, even when controlling for sample size (see Table 8 for the four parameter scores for the Project Implicit and MTurk samples).

As predicted, for the AC parameters, the probability of automatic associations being activated partially occurred for both the Project Implicit sample (Lap-Positive = χ^2 (1) = 0.42, p = .51, w = .00; Niff-Negative = χ^2 (1) = 5.313, p = .021, w = .01) and the MTurk sample (Lap-Positive = χ^2 (1) = 3.757, p = .053, w = .01; Niff-Negative = χ^2 (1) = 21.761, $p < .001$, w = .03). For both groups, the Niff-Negative AC parameter showed a larger, yet still tiny effect size. The MTurk sample formed significantly stronger automatic associations compared to the Project Implicit sample for both the Lap-Positive, χ^2 (1) = 1.374, p = .024, w = .03, and Niff-Negative associations, χ^2 (1) = 5.586, p = .018, w = .06. No biases were overcome (OB) and this parameter did not differ between the two groups, χ^2 (1) = 0.000, p = .999, w = .00. For the D parameters, high scores were observed for both the Project Implicit, χ^2 (1) = 16475.082, $p < .001$, w = .82, and MTurk samples, χ^2 (1) = 9909.112, $p < .001$, w = .66. Crucially, the Project Implicit sample was much better at determining correct responses than the MTurk sample, χ^2 (1) = 443.142, p < .001, w = .50. Lastly, the Project Implicit sample made significantly more "guess" responses than the MTurk sample χ^2 (1) = 4.327, p = .038, w = .05. More specifically, participants in the MTurk sample did not demonstrate any guessing response bias, χ^2 (1) = 0.001, p = 973, w = .00. However, participants in the Project Implicit sample were significantly more likely to make a

guess response using the "good" key press, χ^2 (1) = 6.678, p =.010, w = .02. See Figure 2 for a visual representation of the AC, OB, D, and G processes for both the PI and MTurk samples.

Discussion

As expected, both the Project Implicit and MTurk sample showed weak response biases on the AC parameters, with the MTurkers showing stronger biases on this parameter compared to the Project Implicit sample. Importantly, an extremely small effect size was shown. The lack of any effect in the OB parameter is likely due to the weak AC effects. The D parameter scores showed a large effect size difference between the two groups, with the Project Implicit sample showing higher scores. Although the PI sample did show a positivity response bias on the G parameter, the effect size was negligible. Overall, these results indicate that there is no real-world difference, based on effect sizes, between the two groups on the AC, OB, and G parameters. In contrast, for the D parameter, the Project Implicit sample was far more conscientious and accurate when making responses on the IAT compared to the MTurkers. It appears that when the paid MTurk participants are accustomed to explicit questions that aim to determine if there are being attentive, they perform remarkably well. However, completing an unaccustomed task seems to greatly diminish their attentiveness and conscientiousness. Researchers should be cognizant of giving paid MTurkers novel tasks such as reaction time tasks because doing so may result in more careless responding.

General Discussion

Across three samples, we tested the effects of several different measures of participant quality on the formation of attitudes toward two novel groups. When examining explicit (self-reported) attitudes, only self-reported attention consistently predicted the strength of the attitudes participants formed and the effect remained consistent even after controlling for the effect of the

other measures of participant quality. Across three samples, we did not see consistent evidence that any of the measures of participant quality predicted the strength of implicit attitudes formed toward the two groups as measured by the IAT. Further, we observed strong and robust effects of each of the measures of participant quality on the overall error rates participants made on the IAT such that participants who either passed each of the dichotomous measures or reported paying more attention during the study also made fewer errors on the IAT.

None of the indices of participant quality consistently predicted attitude formation on the explicit and implicit measures. Even when examining the data quality indices collectively, we failed to find consistent evidence that failing even three or more of the selected measures of participant quality consistently impacted attitude formation. This failure to consistently predict attitude formation is striking: in our paradigm, participants who pay more attention to the study should form stronger attitudes. The one measure that did predict attitude formation was a continuous self-reported attention measure and not the dichotomous pass-fail items (i.e., the instructional manipulation check). Even after dichotomizing self-reported attention by excluding anyone one standard deviation below the mean, there was not an appreciable change in the overall strength of the attitudes formed by participants. This finding suggests that if researchers are interested in measuring attention, self-reported attention may provide a more useful approach than the other popular strategies we included in this analysis. Researchers who want to include self-reported attention as a measure of participant quality may benefit from including it as a continuous rather than dichotomous covariate.

Despite the lack of ability to differentiate between people who formed stronger and weaker attitudes, each of the measures of participant quality performed well at differentiating participants who made more or fewer errors on the IAT. This relation was most noticeable in

Study 2 where MTurk workers made substantially more overall errors on the IAT (a function of how quickly and how accurately a person performs the task). These analyses suggest that across both samples participants who made fewer errors on the IAT (either going too slow or too fast on a trial or miscategorizing a stimulus) were also more likely to pass other measures of participant quality. For example, participants who read the full directions of the IMC were also more likely to categorize stimuli correctly even though that increase in accuracy did not result in an appreciably different IAT score between participants who passed or failed the IMC.

Our findings suggest that indices of participant quality are measuring something, but that something might not necessarily be what researchers intend. It appears these measures were better indicators of participant conscientiousness than of attention. Participants who passed these measures were generally better at following instructions and identifying trick questions. Further, the consistent correlations observed between study variables suggest that there is a shared underlying construct measured by these items (see Table 1 for all relations between study variables).

Greater attention did not correspond stronger attitudes toward either of the two novel groups, a surprising finding for which there are two possible explanations. The first is that these measures may not capture the type of attention--depth of processing--that leads people to form attitudes. A second possible explanation is that only limited attention to stimuli is required to form attitudes. If this is the case, then any attention above a certain low threshold may not result in differentially strong attitudes toward the two groups. This robustness against low attention could be considered a strength of this paradigm.

Despite the difference in incentives between the Project Implicit (volunteer) and MTurk (paid) samples, supplementary analysis using quad modeling (Conrey, Sherman, Gawronski,

Hugenberg, & Groom, 2005; Sherman, 2006) revealed the sole difference between the two samples was the D parameter (discriminability) which indicates attentional control. Overall, Project Implicit volunteers, who were all uncompensated, were more attentive during the IAT and correspondingly made fewer overall errors than the MTurk sample. Although we did not have a direct measure of how this difference might have impacted the formation of novel attitudes, this pattern does at least suggest that the IAT is able to detect differences in discriminability. However, we can use mean IAT scores to infer that the differences in discriminability had a negligible impact on the implicit attitude formation. This pattern suggests that although we observed differences in self-rated attention between participants at Project Implicit and MTurk, that difference did not correspond to stronger attitudes in the expected direction. These results suggest that attention and motivation, as measured by the D parameter, were unrelated to the actual strength and valence of the attitudes formed.

After identifying low-quality responses, researchers may struggle to know what to do with them and may ultimately choose to remove them from the analysis. However, in our research most participants failed between one and two of the checks. Further, even participants who failed three or more checks still had comparable attitudes to participants who did not fail a single check. Accordingly, excluding these participants would result in an appreciable loss of statistical power. Based on these analyses, researchers should exercise caution when deciding to exclude participants from their samples.

Although there were not notable differences in the consistency of effects within studies, there are notable differences between the MTurk and Project Implicit samples. Specifically, compared to the Project Implicit samples MTurk participants made more errors on the IAT but were better at passing the attention check measures. Several possibilities could explain this

difference. First, the difference may be due to relative familiarity with the measures. MTurk participants, compared to the Project Implicit participants, were less familiar with the IAT measure and more familiar with the other attention check measures. Another possibility is that Project Implicit participants were paying less attention during the study and correspondingly performed worse on the attention check measures. The second possibility may be less likely considering Project Implicit participants scored higher on the D parameter of the quad model indicating they may have been more attentive on average compared to the MTurk sample. Based on this finding, the Project Implicit sample does not appear to be lower quality than the MTurk sample despite the volunteer nature of the study. Rather, the differences in rates of passing the checks seem to suggest that these measures are contextually effective based on the measures the participants in question are familiar with.

There are, however, limitations of this design. The studies listed here used a simple attitude induction paradigm with novel groups about whom the participants had no prior knowledge. Information about the groups was intentionally valenced to form a preference for one group relative to the other. Our results may not hold within other attitude formation paradigms where the information is more mixed or there is prior knowledge of the groups.

Given that participants read very simple information about the novel groups, it is possible participants were unlikely to elaborate on the information. Models of attitude change such as the Elaboration Likelihood Model (Petty & Cacioppo, 1986) would suggest that different processes might come into play with more complicated information. Although our data cannot speak to the relation between measures of participant quality in more complicated attitude formation paradigms, researchers should be mindful of potential confounds. For example, using these measures to exclude participants may produce samples that are biased toward “high elaboration”

participants and findings may not generalize back to the population. In these cases, researchers should exercise caution and carefully consider the consequences of using these types of participant quality metrics as exclusion criteria.

Second, we examined participant quality only in relation to attitude formation. While we have confidence in our results, different researchers may be interested in using these measures of participant quality outside of the context of attitude formation and we cannot speak to their efficacy in those cases.

Conclusion

Across three attitude induction studies (*N*s = 570, 602, and 210) , four of six measures of participant quality either inconsistently differentiated or failed to differentiate between participants on the basis of attitude formation. Self-reported attention was the most robust predictor of attitude formation. However, all of the tested measures were effective at identifying participants who made more errors during an IAT. When looking at the cumulative effect of failing multiple measures, participants who failed three or more measures still did not consistently differ from participants who passed every measure. Thus, it may be the case that these measures of participant quality are better measures of participant conscientiousness. Researchers should exercise caution when considering excluding participants from an analysis based on whether they passed or failed a single measure of participant quality.

References

Baker, R., & Downes-Le Guin, T. (2007, September). Separating the wheat from the chaff: ensuring data quality in internet samples. In *Proceedings of the Fifth International Conference of the Association for Survey Computing: The Challenges of a Changing World* (pp. 157-166).

Bentler, P. M., Jackson, D. N., & Messick, S. (1971). Identification of content and style: A two dimensional interpretation of acquiescence. *Psychological Bulletin, 76*, 186-204.

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

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers?. *Journal of Experimental Social Psychology, 66*, 20-28.

Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology, 111* (2) 218-229.

Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples. *Behavior Research Methods*, *49*(1), 320-334.

Calanchini, J., Sherman, J. W., Klauer, K. C., & Lai, C. K. (2014). Attitudinal and non-attitudinal components of IAT performance. *Personality and Social Psychology Bulletin*, *40*(10). https://doi.org/10.1177/0146167214540723

Chen, J., & Ratliff, K. A. (2015). Implicit attitude generalization from Black to Black-White biracial group members. *Social Psychological and Personality Science, 6*, 544-550.

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates.

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, *1*(3), 98-101.

Conrey, F. R., Sherman, J. W., Gawronski, B., Hugenberg, K., & Groom, C. J. (2005). Separating multiple processes in implicit social cognition: The quad model of implicit task performance. *Journal of Personality and Social Psychology*, *89*(4), 469–487. https://doi.org/10.1037/0022-3514.89.4.469

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The MODE model as an integrative framework. *Advances in Experimental Social Psychology*, *23*, 75-109.

Fazio, R. H. (2007). Attitudes as object–evaluation associations of varying strength. *Social cognition*, *25*(5), 603-637. https://doi.org/10.1521/soco.2007.25.5.603

Gao, Z., House, L., & Bi, X. (2016). Impact of satisficing behavior in online surveys on consumer preference and welfare estimates. *Food Policy*, *64*, 26-36.

Gawronski, B., Hofmann, W., & Wilbur, C. J. (2006). Are "implicit" attitudes unconscious?. *Consciousness and cognition*, *15*(3), 485-499. https://doi.org/10.1016/j.concog.2005.11.007

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213-224.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464-1480.

Gregg, A. P., Seibt, B., & Banaji, M. R. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90*, 1–20.

Hamilton, D. L., & Gifford, R. K. (1976). Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *Journal of Experimental Social Psychology*, 12(4), 392-407.

Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *Sage Open*, *5*(2), https://doi.org/10.1177/2158244015584617.

Hawkins, C. B., & Ratliff, K. A. (2015). Trying but failing: Implicit attitude transfer is not eliminated by objectivity manipulations. *Basic and Applied Social Psychology*, 37, 31-43.

Haugtvedt, C. P., and Petty, R. E. 1989. "Need for Cognition and Attitude Persistence," *Advances in Consumer Research*, 16, 33-36.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology, 27*, 99-114.

Huang, J. L., Liu, M., & Bowling, N. A. (2015). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology, 100*, 828-845.

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2-9.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krosnick, J. A. (1999). Survey research. *Annual review of psychology*, *50*, 537.

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

Meade, A.W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*, 437-455.

Miura, A., & Kobayashi, T. (2016). Survey Satisficing Inflates Stereotypical Responses in Online Experiment: The Case of Immigration Study. *Frontiers in Psychology*, *7: 1563.*

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*(1), 42-54. https://doi.org/10.3758/BRM.42.1.42

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General, 134(4),* 565.

Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. *Personality and Social Psychology Bulletin, 31(2),* 166-180.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872.

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in experimental social psychology*, *19*, 123-205. doi: https://doi.org/10.1016/S0065-2601(08)60214-2.

Petty, R. E., Haugtvedt, C. P., & Smith, S. M. (1995). Elaboration as a determinant of attitude strength: Creating attitudes that are persistent, resistant, and predictive of behavior. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences*. (pp. 93-130). Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc.

Ranganath, K. A., & Nosek, B. A. (2008). Implicit Attitude generalization occurs immediately; explicit attitude generalization takes time. *Psychological Science, 19,* 249–254.

Ratliff, K. A., & Nosek, B. A. (2010). Creating distinct implicit and explicit attitudes with an illusory correlation paradigm. *Journal of Experimental Social Psychology*, *46*(5), 721-728.

Ratliff, K. A., & Nosek, B. A. (2011). Negativity and outgroup biases in attitude formation and transfer. *Personality and Social Psychology Bulletin, 37*, 1692–1703.

Ratliff, K. A., Swinkels, B. A. P., Klerx, K., & Nosek, B. A. (2012). Does one bad apple(juice) spoil the bunch? Implicit attitudes toward one product transfer to other products by the same brand. *Psychology & Marketing, 29,* 531-540.

Rydell, R. J., McConnell, A. R., Mackie, D. M., & Strain, L. M. (2006). Of two minds: Forming and changing valence inconsistent implicit and explicit attitudes. *Psychological Science, 17, 954–*958.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37,* 867–878.

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9*, 367-373.

Sherman, J. W. (2006). AUTHORS'RESPONSES: Clearing Up Some Misconceptions About the Quad Model. *Psychological Inquiry*, *17*(3), 269-276. https://doi.org/10.1207/s15327965pli1703_7

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291-298.

Williams, M. J., & Eberhardt, J. L. (2008). Biological conceptions of race and the motivation to cross racial boundaries. *Journal of Personality and Social Psychology*, *94*, 1033-1047.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, *47*(2), 237.

Table 1
Correlations among Measured Variables in Study Samples A and B and Study 2

Measure	1	2	3	4	5	6	7	8
Study 1 Sample A								
1. Implicit Attitude								
2. Explicit Attitude	**.236							
3. Total IAT Error Rate	.051	-.069						
4. Self-reported Attention	.018	**.148	**-.359					
5. Too High IAT Error Rate	-.036	.073	**-.768	**.261				
6. "Bogus" Scale Item	-.035	.005	**-.217	*.102	**.250			
7. Data Retention	-.066	.065	**-.280	**.328	**.153	*.092		
8. IMC	-.019	.003	**-.266	**.200	**.205	**.137	**.142	
9. Reaction Time	.037	.037	-**.259	**.142	**.174	.001	.048	**.215
Study 1 Sample B								
1. Implicit Attitude								
2. Explicit Attitude	**.212							
3. Total IAT Error Rate	**-.124	**-.112						
4. Self-reported Attention	*.089	**.120	**-.267					
5. Too-High IAT Error Rate	**.109	*.086	**-.733	**.235				
6. "Bogus" Scale Item	.004	.048	**-.157	*.097	**.131			
7. Data Retention	*.099	.076	**-.265	**.333	**.240	*.102		
8. IMC	*.106	.056	**-.245	**.167	**.184	.077	**.204	
9. Reaction Time	*.083	.049	**-.308	**.181	**.240	.000	.030	**.211
Study 2								
1. Implicit Attitude								
2. Explicit Attitude	0.135							
3. Total IAT Error Rate	*-.161	**-.208						
4. Self-reported Attention	-0.072	*.138	**-.255					
5. Self-reported Effort	-0.044	0.122	**-.310	**.624				
6. Too-High IAT Error Rate	*.173	**.211	**-.833	**.223	**.198			
7. Reaction Time	*.142	**.281	**-.267	0.12	0.066	**.345		

* $p < .05$, ** $p < .01$

Table 2
Adapted from Wetzels, Matzke, Lee, Rouder, Iverson, & Wagenmakers, 2011

Bayes Factor 10 (BF_{10})	*Interpretation*
> 100	*Decisive evidence for H_A*
30 - 100	*Very strong evidence for H_A*
10 - 30	*Strong evidence for H_A*
3 - 10	*Substantial evidence for H_A*
1 - 3	*Anecdotal evidence for H_A*
1	*No evidence*
0.33 - 1	*Anecdotal evidence for H_0*
0.10 - 0.33	*Substantial evidence for H_0*
0.03- 0.10	*Strong evidence for H_0*
0.01 - 0.03	*Very strong evidence for H_0*
< 0.01	*Decisive evidence for H_0*

Table 3

Visual representation between the outcome measures and measures of participant quality in each sample. Boxes marked with an "X" indicate a statistically significant relationship ($p < .05$) when each predictor was entered separately. N/A indicates a measure that was not included in that analysis.

Outcome measure	*Instructional Manipulation Check*	*Self-Reported Data Retention*	*Self-Reported Attention*	*Self-Reported Effort*	*Bogus Scale Item*	*Too Many IAT Errors*	*Response Time*
				Sample A			
Implicit Attitude	—	—	—	N/A	—	—	—
Explicit Attitude	—	—	X	N/A	—	—	—
IAT error rate	X	X	X	N/A	X	N/A	X
				Sample B			
Implicit Attitude	X	X	—	N/A	—	X	—
Explicit Attitude	—	—	X	N/A	—	X	—
IAT error rate	X	X	X	N/A	X	N/A	X
				Study 2			
Implicit Attitude	N/A	N/A	—	—	N/A	X	—
Explicit Attitude	N/A	N/A	—	—	N/A	—	X
IAT error rate	N/A	N/A	—	X	N/A	N/A	X

Table 4

Visual representation between the outcome measures and measures of participant quality in each sample. Boxes marked with an "X" indicate a statistically significant relationship ($p < .05$) when all predictors were entered simultaneously.

Outcome measure	*Instructional Manipulation Check*	*Self-Reported Data Retention*	*Self-Reported Attention*	*Self-Reported Effort*	*Bogus Scale Item*	*Too Many IAT Errors*	*Response Time*
				Sample A			
Implicit Attitude	—	—	—	N/A	—	—	—
Explicit Attitude	—	—	X	N/A	—	—	—
IAT error rate	X	X	X	N/A	X	N/A	X
				Sample B			
Implicit Attitude	—	—	—	N/A	—	—	—
Explicit Attitude	—	—	—	N/A	—	—	—
IAT error rate	X	X	X	N/A	X	N/A	X
				Study 2			
Implicit Attitude	N/A	N/A	—	—	N/A	X	—
Explicit Attitude	N/A	N/A	—	—	N/A	—	X
IAT error rate	N/A	N/A	—	X	N/A	N/A	X

Table 5

Explicit Attitudes as a Function of Participant Quality Entered Simultaneously in Study 1 (Samples A and B) and Study 2

Predictor	*B*	T_{522}	$p \leq$	r_p [95% CI]
Study 1 Sample A				
Intercept	-0.579	—	—	—
Instructional Manipulation Check	-0.104	-0.69	.49	-.030 [-.115, .055]
Self-Reported Data Retention	0.109	0.63	.53	.027 [-.058, .112]
Self-Reported Attention	0.236	2.68	.01	.117 [.032, .200]
"Bogus" Scale Item	-0.085	-0.42	.68	-.018 [-.103, .067]
Too many IAT errors	0.129	0.42	.68	.018 [-.067, .103]
Response Time	0.171	0.44	.66	.019 [-.066, .104]
Study 1 Sample B				
Intercept	-0.563	—	—	—
Instructional Manipulation Check	0.030	0.18	.85	.008 [-.077, .093]
Self-Reported Data Retention	0.168	0.88	.38	.038 [-.047, .122]
Self-Reported Attention	0.201	2.20	.03	.095 [.010, .178]
"Bogus" Scale Item	0.110	0.49	.62	.022 [-.063, -.106]
Too many IAT errors	0.298	1.12	.27	.049 [-.036, .133]
Response Time	0.211	0.45	.66	.020 [-.065, 0.104]
Study 2				
Intercept	-7.992	—	—	—
Self-Reported Attention	0.175	0.62	.54	.041 [-.095, .175]
Self-Reported Effort	0.174	0.61	.54	.040 [-.096, .174]
Too many IAT errors	-0.479	-1.50	.13	.-.099 [-.231, .037]
Response Time	2.390	3.31	.01	.219 [.086, .344]

Table 6

Implicit Attitudes as a Function of Participant Quality Entered Simultaneously.

Predictor	*B*	T_{532}	$p \leq$	r_p [95% CI]
Sample A				
Intercept	-0.239	—	—	—
Instructional Manipulation Check	-0.016	-0.34	.74	-.015 [-.090, .070]
Self-Reported Data Retention	-0.077	-1.43	.15	-.062 [-.143, .016]
Self-Reported Attention	0.024	0.86	.39	.037 [-.041, .119]
"Bogus" Scale Item	-0.055	-0.86	.39	-.037 [-.118, .042]
Too many IAT errors	0.030	0.32	.75	.014 [-.065, .095]
Response time	0.075	0.62	.54	.027 [-.065, .095]
Sample B				
Intercept	-0.643	—	—	—
Instructional Manipulation Check	0.056	1.30	.19	.056 [-.029, .140]
Self-Reported Data Retention	0.087	1.68	.09	.072 [-.013, .156]
Self-Reported Attention	0.023	0.94	.35	.040 [-.045, .124]
"Bogus" Scale Item	-0.039	-0.65	.52	-.028 [-.112, .057]
Too many IAT errors	0.134	1.87	.06	.080 [-.005, .164]
Response time	0.176	1.40	.16	.060 [-.025, .144]
Study 2				
Intercept	-0.092	—	—	—
Self-Reported Attention	-0.062	-.70	.49	-.048 [-.185, .091]
Self-Reported Effort	-0.037	-.47	.64	-.032 [-.169, .107]
Too many IAT errors	-0.191	-2.18	.03	-.152 [-.284, -.014]
Response Time	0.277	1.38	.17	-.096 [-.231, .043]

Table 7

Overall IAT Error Rates as a Function of Participant Quality Entered Simultaneously.

Predictor	*B*	T_{532}	$p \leq$	r_p [95% CI]
Sample A				
Intercept	0.482	—	—	—
Instructional Manipulation Check	-0.031	-3.26	.001	-.126 [-.208, -.042]
Self-Reported Data Retention	-0.033	-3.06	.002	-.118 [-.200, -.034]
Self-Reported Attention	-0.033	-6.05	.000	-.234 [-.312, -.153]
"Bogus" Scale Item	-0.058	-4.61	.000	-.178 [-.259, -.095]
Response time	-0.060	-2.43	.015	-.094 [-.177, -.010]
Sample B				
Intercept	0.659	—	—	—
Instructional Manipulation Check	-0.040	-3.51	.01	-.143 [-.225, -.059]
Self-Reported Data Retention	-0.036	-2.64	.01	-.108 [-.191, -.024]
Self-Reported Attention	-0.018	-2.73	.01	-.111 [-.194, -.027]
"Bogus" Scale Item	-0.047	-2.94	.01	-.120 [-.203, -.036]
Response time	-0.116	-3.45	.01	-.141 [-.223, -.057]
Study 2				
Intercept	2.062	—	—	—
Self-Reported Attention	-.045	-1.21	.23	-.079 [-.214, .059]
Self-Reported Effort	-.096	-2.67	.01	-.173 [-.303, -.036]
Response Time	-.328	-3.78	.01	-.245 [-.370, -.111]

Table 8
The Quad modeling parameter scores for both the Project Implicit sample and MTurk sample

Sample	AC (Lap-Positive)	AC (Niff-Negative)	OB	D	G
PI Sample	0.003 [-.001, .007]	0.011 [0.006, 0.016]	0.000 [0.000, 0.000]	0.821 [0.818, 0.824]	0.529 [0.508, 0.549]
MTurk Sample	0.013 [-.001, .026]	0.031 [0.017, 0.044]	0.000 [-0.764, 0.764]	0.673 [0.660, 0.687]	0.500 [0.485, 0.516]