"'Could you please pay attention?" Comparing in-person and MTurk Responses on a Computer Code Review Task

Anthony M Gibson
Consortium of Universities
Dayton, OH
Anthony.mgibson89@gmail.com

Gene M. Alarcon
Air Force Research Laboratory
Wright Patterson AFB, OH
Gene.alarcon.1@us.af.mil

Michael A. Lee
GDIT
Dayton, OH
Michael.lee@gdit.com

Izz aldin Hamdan
GDIT
Dayton, OH
Izzy.hamdan@gdit.com

Abstract

The current study examined the differences in data quality across two environments (i.e., in a laboratory and online via Amazon's Mechanical Turk) on a computer code review task. Researchers and practitioners often collect data online for the sake of convenience, as well as for obtaining a more generalizable sample of participants. The lack of social contact between the researchers and participants, however, may result in less effort dedicated to the experimental task resulting in poor quality data. The results of the current study showed that data quality—at least when measuring the individual difference variables—was drastically worsened when the experimental task was presented online. In contrast, we observed little differences in the experimental task perceptions across the two samples. Rather, participants spent significantly less time examining the computer code when completing the experiment online. The current study has implications for the effects of using online platforms (like MTurk) to collect experimental data.

1. Introduction

The extent to which results and the corresponding conclusions from experimental studies are valid depends on researchers and practitioners collecting high-quality data. Thus, the conclusions drawn from self-report data can be negatively affected when inattentive participants provide invalid data. Although prior research has focused on inattentive participants completing online questionnaires [1, 2], inattentive respondents could also lead to low-quality data in experimental tasks. For example, respondents are expected to fully read experimental instructions and text-based manipulations. Failure to read the directions properly can influence the results of the experiment and lead to unexpected results. Thus, the possible negative effects of inattentiveness during experimental tasks deserves more research focus. In the current study, we examined the data quality of computer programmers completing a computer code review task either inside the laboratory or online (i.e., on MTurk). We first describe inattentive, or careless, responding in general and then describe the negative effects of inattention on experimental tasks specifically.

Careless responding (CR) occurs when participants answer questions without the motivation to provide a valid response based on question content [1]. Note that CR is separate from other response biases (e.g., faking or social desirability). Rather than having a response pattern that exaggerates participants' positive qualities, CR occurs when participants answer the experimental questions in a manner that is unrelated to the item content. For example, a respondent motivated to complete the experiment quickly could select the first response option for every question in the experiment.

Unsurprisingly, the presence of CR can negatively affect the validity of research data for both survey and experimental data [1, 2, 3]. For example, CR can artificially attenuate the correlation between two theoretically related variables [1], reduce the observed internal consistency of variables, [4] obfuscate the results of factors analyses [5], and artificially inflate the correlation between two unrelated variables [1]. The latter occurs when the mean of the careful respondents is located on either ends of the response scale, as careless participants' data typically has an average near the midpoint of the scale. When the careless respondents have a mean

URI: https://hdl.handle.net/10125/71121
978-0-9981331-4-0

that is in the center of the scale—but the careful respondents have a mean on one of the end points of the scale—this creates a confound that can artificially inflate the observed correlation between otherwise unrelated variables. Given the negative, and sometimes unpredictable, effects of CR on data quality, it is important for researchers to scan for CR in their own data sets, as well as implement manipulations in their experimental and correlational designs to prevent CR from occurring. In the following subsections, we highlight methods to detect careless responding when using self-report data, as well as examine deterrent methods to prevent CR in both surveys and experimental studies.

1.1. Detecting Careless Responding

Methods for detecting CR have been around for decades [6, 7]. In general, CR indices are separated into two general groups: (a) indices that can be considered a-priori and inserted into the experiment itself (e.g., bogus or infrequency items) and (b) indices that are considered post-hoc and analyzed after the experiment is finished [8, 9]. Common a-priori CR indices are infrequency items [7], response time indices [10], and semantic consistency indices [4]. Infrequency, or bogus, items are items inserted into self-report scales that all careful participants should provide the same response (e.g., I like getting speeding tickets; [4]). Page time indices record how long participants spend on each experimental page and have a cutoff for those who complete the page excessively fast. A typical cutoff value for self-report items has been two seconds per question, which converged significantly with other CR indices [1, 10]. Unlike the infrequency index, page time indices can surreptitiously record participants' response time and thus measure careless responding without attracting attention from otherwise careful respondents.

Researchers and practitioners can also use post-hoc methods to detect poor data quality [8]. Conventional post-hoc methods include long string indices to detect patterned responses (e.g., selecting the same response across multiple questions; [1, 2]), consistency indices to detect responses that contradict each other [1, 2], and indices that detect unlikely distributions of responses to the test items, such as Mahalanobis D scores [9, 10]. A benefit of the post-hoc indices is that researchers and practitioners can use these methods after the data has been collected [2, 9].

Although there are numerous methods to detect low-quality responses, researchers have found them to relate to each other in predictable ways. For example, it has been shown that CR indices load onto two separate factors [2]. One factor includes the long string index, which captures respondents who repeatedly select the same item response repeatedly across different experimental questions. The indices that load onto the second factor include the infrequency items, the response time indices, and the reliability, or consistency, indices [2]. Importantly, for the second factor, the different indices all have error variance associated with the scores. Thus, it is important to observe participants' scores across various indices to determine whether or not people have provided low quality data [8, 9].

Identifying problematic participants and removing careless data can enhance data quality. Huang et al. [1] found that removing suspected careless responders improved the Cronbach's alpha estimates of traditional self-report scales. Also, removing cases showing signs of CR has improved statistical power [4]. Thus, it is important that careless responders be identified before traditional hypothesis testing is conducted. In the subsection below, we highlight potential benefits of preventative techniques to reduce careless responding, rather than omitting suspected cases after data collection.

1.2. Preventing Careless Responding

Contemporary research has shown that removing careless participants may have unintended negative consequences, which has shifted research attention to preventing careless responding. First, data collection takes much time and effort and removing participants wastes valuable resources. Specifically, Meade and Craig [2] found that approximately 10 to 12 percent of undergraduate students respond carelessly on typical low-stakes online questionnaires. As such, the amount of wasted resources can increase substantially, as researchers attempt to achieve the desired statistical power. For example, researchers and practitioners may need to collect an extra 10-20% of the required sample knowing they will need to omit careless cases.

Perhaps even more concerning, contemporary researchers have found that removing CR cases can inadvertently remove a particular subset of the population, potentially limiting the external validity of research findings. Specifically, Bowling et al. [10] found that CR was correlated negatively with other reports of conscientiousness and agreeableness, as well as negatively with extraversion and emotional stability. If researchers delete cases that have been detected as having a high likelihood of CR, they may systematically remove a particular subset of the population (e.g., removing participants who score low in conscientiousness). Thus, in order to ensure the sample accurately matches the intended population of interest, it is important to consider ways to possibly prevent CR from occurring before launching the scientific study.

Compared to the methods to detect CR, researchers have had more difficulty finding effective ways to prevent CR. One method that has shown promise across numerous studies is a warning message, which highlights negative consequences for participants providing low-quality data [2]. For example, Gibson and Bowling [11] showed that a warning message highlighting a punishment for CR consisting of a revocation of research credits for undergraduate students prevented CR. Other preventative methods that have been considered include: a) adding an avatar to the survey page to monitor participants' responding behaviors that only reduced CR when paired with a warning manipulation [12], b) having participants personally sign their names to a pledge promising to respond carefully [2], and c) removing anonymity from questionnaires without a pledge to respond carefully [2]. Although some methods have shown promise in preventing CR, it is important to understand theoretically why some techniques may be more effective than others.

Meade and Craig [2] provided rationale for why certain manipulations may effectively prevent CR. Specifically, the authors described various factors that likely promote careless responding on online surveys. First, an increase in anonymity likely reduces participants' perceived accountability to complete the study carefully. Second, a lack of social contact between the participants and researchers may also reduce participants' motivation to put their full effort into the task [2]. Specifically, when participants interact with the researchers, they should feel more apt to put forth effort in the study compared to when participants complete studies online. Third, environmental distractions are likely more common when participants complete studies online. Indeed, Gibson and Bowling [11] found that participants had much higher incidence of CR when completing an online study versus completing the study in a laboratory. These environmental distractions may also lead to CR on experimental tasks. With the ubiquity of the internet, participants can now complete studies in various environments, many of which are assumed to have more distractions than the laboratory.

Prior research has found that a lack of participant effort can also lead to invalid findings outside the scope of self-report individual difference items [13]. Oppenheimer et al. [13] showed that participants skipping instructions on research studies can reduce the effectiveness of classic manipulations in the social psychology literature. Furthermore, the inclusion of participants who skip reading experimental instructions added random error into the measurement of study variables and reduced statistical power. Given that our current study used text-based instructions for our manipulations, we expect that some participants may miss key details about the manipulation, due to inattentiveness. In the section below, we describe the current experiment in more detail.

1.2. The Current Study

The current study examined the presence of careless respondents across two data sets performing an identical experimental task. In the first data set, participants were computer programmers reviewing computer code on Amazon's Mechanical Turk (MTurk). A second data set consisted of computer programmers examining the same piece of computer code but inside the laboratory with the experimenter present. Given the description above, we expect that participants completing the survey online should provide more careless responses on both the individual difference items and the experimental task compared to participants completing the study in the laboratory.

- Hypothesis 1: Online participants will exhibit more careless responding than in-person participants.

2. Method

2.1. Participants

A between subjects' design was used to compare laboratory (*N* = 58) and online (*N* = 158) participants. (Note the sample for in-person participants was to be higher but Covid-19 prevented completion of data collection). Participants had to be at least 18 years of age, have a minimum three years of programming experience, and know the Java programming language. The laboratory sample had a mean age of *M* = 29 years (*SD* = 10 years), was 77% male, 67% white, and had an average of 6.6 years of programming experience. The online sample had a mean age of *M* = 30 years (*SD* = 23 years), was 80% male, 73% white, and had an average of 10.2 years of programming experience. As the samples had similar ages and experience, we did not test for any differences in regards to demographics.

2.2. Experimental task

For the in-person group, participants were welcomed into the laboratory and informed that they would be assessing several pieces of code. Participants were instructed to answer a demographic questionnaire and background survey. Upon completion, participants moved on to the main task, where they assessed the trustworthiness of six pieces of code. In-person participants were given a $50 American Express gift card as compensation for their participation. We chose this renumeration amount as it reflects the average programmer's hourly rate [14].

The online study was administered through MTurk. The same instructions and processes presented to the in-person group were also administered to the online participants in the MTurk format. Online participants were renumerated $10 USD for their time and effort. As $50 is not a normal amount of renumeration on the platform, we did not want to unduly influence MTurk workers [see 14].

2.3. Manipulations

All participants viewed six pieces of code, each with a different function. We will not describe the details of the code, as they are beyond the scope of the current study. However, it is important to note that code six (the final code) was flawed. Thus, we used this code as a manipulation check to ensure participants were carefully reviewing the code.

2.4. Measures

2.4.1. Dispositional Trust. Trust was measured using the 10-item measure from the International Personality Item Pool [15]. A sample item was, "Believe that others have good intentions." Respondents answered these items on a 5-point rating scale (1 = *Very Inaccurate* to 5 = *Very Accurate*).

2.4.2 Need for Cognition. Need for cognition was assessed using Cacioppo, Petty, and Kao's [16] 19-item scale. A sample item was, "I would prefer complex to simple problems." Respondents answered these items using a 5-point rating scale (1 = *Strongly Disagree* to 5 = *Strongly Agree*).

2.4.3 Propensity to Trust in Technology. Propensity to Trust in Technology was measured using Jessup et al.'s [17] 6-item scale. A sample item was, "Generally, I trust technology." Respondents answered these items on a 5-point rating scale (1 = *Strongly Disagree* to 5 = *Strongly Agree*).

2.4.4 Suspicion Propensity Inventory. Suspicion Propensity was assessed using Calhoun et al.'s [18] 8-item scale. A sample item was, "I am naturally suspicious". Respondents answered these items on a 5-point rating scale (1 = *Strongly Disagree* to 5 = *Strongly Agree*).

2.4.5 Code ratings. Participants were presented with five single-item measures to assess different aspects of each piece of code: Reputation was measured by the item, "How reputable is the code?" Maintainability was measured by the question, "How maintainable is this code?" Transparency was measured by asking, "How transparent is this code?" We measured perceived performance with the item, "How well do you think this code will perform?" Finally, perceived trustworthiness was measured with the item, "How trustworthy is the code?". Respondents answered these items on a 7-point rating scale (1 = *Not at All* to 7 = *Very*). For the current study, we only considered the reputation item and the trustworthiness item. The condition for this specific piece of code was Reputable, so we converted the item into a dichotomous manipulation check item. Specifically, we coded any endorsement of this item as careful, whereas any participants who marked the code as unreputable was coded as careless.

2.4.6 Code description. For each piece of code, participants were prompted to describe what the code does: "To the best of your knowledge, please describe what this code does in the text box below." As a proxy for the amount of effort participants exerted, we compared the word and character counts between the online and in-person samples.

2.5. Careless Responding Indices

2.5.1. Long string. We measured long string by computing the number of identical, consecutive responses across all the self-report scales [2]. Because the scales measured different psychological constructs and some items were reverse-scored, participants selecting the same response across a large number of consecutive items were assumed to be careless. The index was calculated for the initial individual difference items, as well as the self-report data for the computer code perception items.

2.5.2 Page time. We assessed page time both in the individual difference items and the code perception pages. For the individual difference items, we recoded the page time submissions (in seconds) into binary variables. We used a cutoff of two seconds per item, in which participants completing the individual difference items faster than two seconds per item (per page) was identified as careless [see 10]. Because the self-report code perceptions questions were on the same page as the computer code itself, we set the cutoff for this page at 120 seconds. That is, participants reading the code and answering the code perceptions questions in under two minutes were flagged as careless.

2.4.3 Even-odd consistency. Even-odd consistency estimates were measured only for the individual difference self-report items. First, we split the scales into halves and then computed the mean score for each participant across all the scale halves [8]. Next, we created two separate vectors of mean scores across the half scales. Finally, we computed the within-person correlation of the two vectors that comprised the mean scores for each scale [see 8 for a full description]. Note we multiplied participants' scores by negative one, so positive scores were indicative of a higher likelihood of careless responding. For example, a person with a score of positive one on this index would have provided perfectly inconsistent responses. Participants with even-odd consistency scores greater than zero were identified as careless.

2.4.3 Mahalanobis D index. Similar to the even-odd consistency index, the Mahalanobis D index was measured only for the individual difference items. In short, Mahalanobis D is a multivariate outlier statistic, which can be used to determine whether participants' pattern of responses deviate from the pattern of the rest of the sample's responses [8]. In order to ensure the pattern of responses cluster together, we ran Mahalanobis D scores for each of the individual difference variables. Next, we transformed the estimates into *z*-scores and computed an average score across the individual difference scales [10]. Then, we observed any participants with excessively large average *z*-scores.

3. Results

We show the means and standard deviations for all CR indices in Table 1. Participants first answered the self-report individual difference items. The scales were included on their individual survey page, with a total of four survey pages. Then, participants saw the computer code snippets and answered the code perception items. To test our research question, we performed a series of t-tests on the variables of interest.

3.1. Long string index

We first measured the number of cases across in-person and online conditions that had high long string values. Given that some survey pages contained only a few items, we observed the long string value across the four individual difference variable pages. In total, participants answered 36 items measuring the individual difference variables.

Table 1
Mean Scores of CR Indices Across Online and In-Person Environments

Index	Environment	
	Online	In-Person
Long string	4.18 (3.19)	4.42 (1.99)
Page time	0.97 (1.34)	0.08 (0.57)
Mahalanobis D	0.02 (0.99)	0.00 (1.00)
Even-odd	-.71 (.52)	-.64 (.51)

Note. Online *N* = 158. In-person *N* = 50. Mahalanobis D scores were transformed to *z*-scores. All indices were recoded so larger values show increased probability of CR. These results correspond to the individual difference items only. Respondents within the online sample with large long string values increased the observed standard deviation.

In the online sample, we observed five participants (out of a total of 158; approximately 3%) who had long string values greater than 10. We chose ten as a cutoff value to represent greater than 25% of the total number of items. In contrast, only one participant, or two percent, who completed the study in the laboratory had a long string value over 10. For the code perceptions self-report data, we measured the number of consecutive, identical responses across all six pieces of code. Note that the six pieces of code were different, so we expected that individual participants would rate the computer codes differently. In total, participants rated the six pieces of code with five items each, resulting in a total of 30 items total. We chose a more conservative cutoff value of fifteen, as participants answered the same code perception items across the different pieces of code. In the online sample, we observed that three participants, or two percent, had a long string value greater than fifteen. There was no statistically significant difference between the online and in person samples, $t(133.3) = -.062$, $p = .53$. Thus, within the MTurk sample, three participants answered at least half the code perception items with the identical response. In contrast, zero participants had long string values over 15 for the code review items. However, although long string responses were more problematic in the online sample compared to the in-person environment, the results were not statistically significant. These findings do not support the expectations of Hypothesis 1.

3.2. Completion time index

Next, we compared the completion time scores across the in-person and online samples. We first observed the completion time index across the individual difference items. Because there is error variance associated with all CR indices, we flagged participants who completed two or more survey pages conspicuously fast. In the online sample, 41 participants (i.e., 26%) had more than one survey page that was completed faster than two seconds per item. In contrast, only one participant (or two percent) was identified as completing the individual difference items conspicuously fast when the sample completed the study in the laboratory. As shown in Table 1, respondents online had on average one page flagged by the page time index, whereas the mean for the in-person study was close to zero. It should be noted that both samples completed the study on similar web pages.

Next, we examined the page time for the survey page that contained both the computer code and the code perception items. We used a cutoff value of 120 seconds, as participants had to read approximately 500 lines of computer code, answer five self-report Likert-type items, and answer one open-ended question. For the MTurk sample, 57 participants (i.e., 36%) completed the code review task and reported their perceptions of the code in under 120 seconds. In contrast, only seven participants (14%) completing the study in the laboratory completed the code review and answered the self-report items in under 120 seconds. Lastly, we conducted an independent samples t-test. Results indicated the online sample was significantly higher than the in-person sample for the completion time index $t(190.27) = 6.71$, $p < .001$. In total, the findings support the expectation of Hypothesis 1 that participants would be more careless than those completing the study in the laboratory based on the completion time index.

3.3. Even-odd consistency index

Next, we compared the even-odd consistency index across the two samples. Note that we considered even-odd consistency scores for the individual difference items only. For the MTurk sample, 14 participants (i.e., nine percent) of the sample were flagged for having even-odd scores above zero (see Figure 1). In contrast, four participants (i.e., eight percent) had even-odd consistency values over zero in the laboratory sample. Results of the t-test indicated no differences between the samples $t(83.86) = -0.82$, $p = .41$. Thus, although the count for participants who were flagged by the even-odd consistency index was higher for the online sample, the percent flagged was relatively equal across samples. Thus, we failed to observe any differences in even-odd consistency scores across samples.

3.4. Mahalanobis D index

When computing the Mahalanobis D estimates, we considered the individual difference items only. We used a cutoff of positive three, as this represents three standard deviations above the mean when the scores are standardized. For the Mahalanobis D index estimates, two participants (one percent) had *z*-scores greater than three in the online sample, whereas zero participants had Mahalanobis D *z*-scores greater than three in the in-person sample. The distributions across the two groups for the standardized Mahalanobis D scores are shown in Figure 2. The independent samples t-test indicated no differences between the samples $t(82.52) = 0.14$, $p = .88$. In total, there was little evidence that aberrant response patterns were an issue for either sample. Thus, we observed limited evidence of greater Mahalanobis D scores for the online sample.

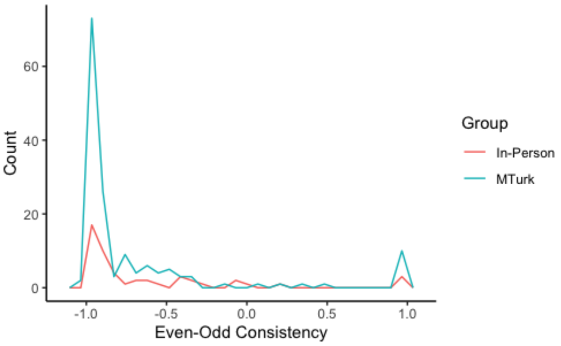

***Figure 1*. Even-odd consistency distributions across groups. Positive scores indicate CR.**

3.5. Code perceptions

Finally, we examined the differences in code perceptions across the two groups (i.e. online or in-person participants). First, we examined the number, and corresponding proportions, of participants who correctly identified that the code was extracted from a reputable source. At the top of the code, there was a line that explicitly stated the code was from a reputable source. Thus, all participants read the code should report that the code was written by a reputable source. We recoded the reputation item into either one (i.e., participants incorrectly stated that the code was from an unreputable source) or zero (i.e., they endorsed the question that the code was from a reputable source). Results of the independent samples t-test indicated the online sample had more instances of careless responding, $t(191.09) = 9.46$, $p < .001$. Thus, the online sample reported was more likely to report that the code was from a unreputable source.

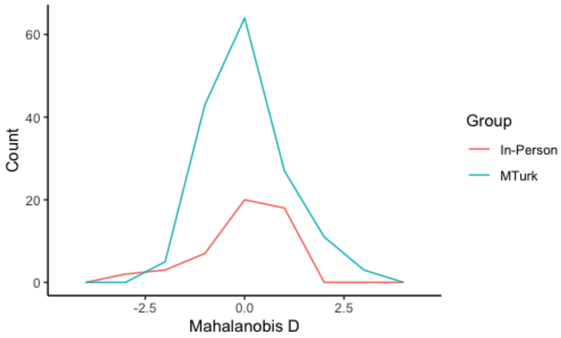

Figure 2. **Mahalanobis D standardized score distributions across groups. Positive scores indicate CR.**

Next, we examined the number of words that participants wrote when describing the function of the code (see Figure 3). We expected that participants exerting sufficient effort in the study would write more words when describing the code than participants putting forth little effort. As the distribution was skewed for the wordcount variable, we examined the median number of words written across the two groups. In the MTurk sample, the median word count was 12 per participant. The median word count for the in-person sample was 17.50. An independent samples t-test indicated no differences between the two-groups, $t(93.53) = 0.13$, $p = .89$.

Table 2
Descriptive Statistics of CR Indices Across Online and In-Person Environments

Measure	Environment	
	Online	In-Person
Time spent	164.88	257.10
Word count	12.00	17.50
Reputable	58%	56%

Note. Online $N = 158$. In-person $N = 50$. Time spent on code and word count variables were positively skewed, so we report the median values. For Reputation, we included all participants who incorrectly stated that the code was Reputable.

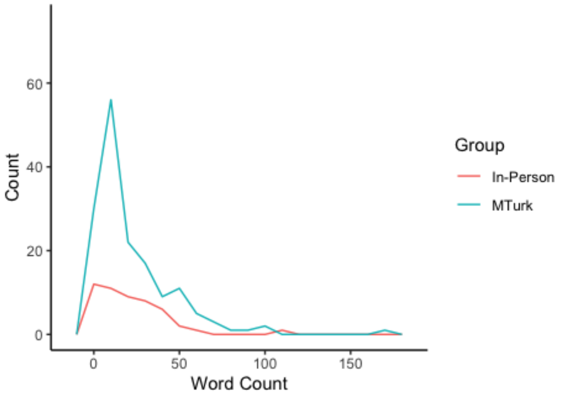

***Figure 3*. Distributions of the word counts of descriptions of the computer code across groups.**

Finally, this particular piece of code was designed to be unorganized and difficult to understand. Thus, we expected attentive participants to rate this final code as untrustworthy and compared the level of perceived code trustworthiness across samples. In the MTurk sample, the average trustworthiness score for the computer code was 4.61 out of a maximum of seven, with a standard deviation of 1.50; the mean and standard deviation for the in-person sample was 4.70 and 1.56, respectively. Thus, in contrast to our hypothesis, mean trustworthiness scores were higher than we expected across both groups.

4. Discussion

In general, we found some support for a higher incidence of CR in online data collection platforms compared to in-person environments. Specifically, we found that a larger percent of respondents answered with long string patterns (i.e., both on the individual difference items and the experimental questions) when they completed the experiment online, compared to participants who completed the survey in the laboratory. However, these results were not statistically different. Additionally, participants completing the study online were more likely to complete both the individual difference questions and the experimental task egregiously fast compared to participants who completed the study in the laboratory. Specifically, those completing the experimental code completion task online had a median response time that was nearly 100 seconds faster than those completing the study in the laboratory. Finally, we observed similar even-odd consistency scores across samples for the individual difference items, which may indicate that although participants were much faster in their responses to the individual difference items, they responded somewhat consistently in their responses. This finding was unexpected, and we expand on this further in the section below. In total, these findings provide some evidence that participants completing the study online put forth less effort than programmers completing the experiment in the laboratory.

In terms of potential negative effects of inattention on experimental manipulations, we found evidence that participants completing the experiment online responded differently to the computer code portion than participants completing the experiment in-person. First, the percent of participants who correctly identified that the computer code was extracted from a trustworthy source was similar across samples. Interestingly, approximately 40% of both samples failed to endorse the item that the code came from a trustworthy source, which is surprising in that this information was written clearly at the top of the computer code. One explanation for these findings could be that the code was purposefully written to be unorganized and difficult to understand. Thus, even though we explicitly stated that the code was taken from a reputable source, participants may have been skeptical of this information based on the low quality of the code. We should note, however, that the online sample was significantly faster on their page times and more terse in their responses to the code. It may be that gaining meaningful written responses from online samples requires more direction or motivation to get respondents to write.

Second, both online and in-person participants rated the trustworthiness of the code similarly, with the mean trustworthiness perceptions being above the midpoint across both samples. These findings were also surprising, as this computer code was manipulated to have poor organization and little transparency. One possible explanation for these findings is that the majority of participants in both samples failed to read the computer code thoroughly enough to accurately rate the trustworthiness of the code. Another possible explanation is that participants considered other aspects of the code than what was manipulated in the current study (i.e., source reputation, transparency, and organization). Future research should include an open-ended question asking for rationale for participants' ratings on the code trustworthiness items.

4.1. Theoretical and Practical Implications

The current findings have implications for measuring CR when using online platforms to collect data. First, we found evidence of increased CR (i.e., more long string patterns and more egregiously fast response times) when participants completed the study online, though the differences were not statistically different from in-person. Note that we observed more long string patterns and faster response times in both the individual difference variables portion and the experimental task portion of the study. Although previous studies have found CR to be more problematic within online surveys [11], the current findings extend previous research to computer programmers completing HITs on MTurk. Similarly, participants in the current study had shorter response times for the experimental tasks, along with the self-report items. Thus, inattentiveness may be more general than responses on self-report Likert-type items. Indeed, the same rationale for CR during online surveys correspond to other experimental tasks [2]. For example, Meade and Craig [2] described reasons for increased CR with online surveys including increased anonymity, reduced social contact with the researcher(s), and increased vulnerability to environmental distractions. Given that these occurrences could also reduce accountability and effort on experimental tasks, data quality appears to suffer on online experiments as well. Thus, we recommend that researchers implement methods to detect and/or prevent CR when collecting survey and experimental data using MTurk.

The findings also have implications for researchers using MTurk to collect experimental data. In the current study, we refrained from using stringent criteria to recruit MTurk participants, as we were collecting data on a specific population (i.e., computer programmers). Specifically, we failed to specify a minimum number of Human Intelligence Tasks (HITs) that participants had to complete or a minimum acceptance rate. Although numerous studies have specified completion of a minimum of 100 HITS and an approval rate of 95% or higher from their workers [19], we were interested in the rates of CR within computer programmers on MTurk. Thus, the results we observed may differ from experiments that have stringent worker requirements. Second, researchers using MTurk workers without implementing methods to prevent CR should be prepared to collect more data than originally determined by a power analysis, as many participants may complete the study faster than is reasonably possibly when answering effortfully. Researchers using MTurk, and other similar online platforms, to collect human-subjects research should take proactive steps to account for the possibility of reduced attentiveness of participants on these sites.

Finally, in contrast to collecting more participants than a power analysis suggests, researchers may want to implement techniques to limit the possibility of MTurk workers rushing through the study. For example, Gibson and Bowling [11] found that both a warning describing negative consequences of engaging in CR, as well as including a potential reward to respond carefully, both reduced CR rates. Thus, researchers may want to include a possible reward (e.g., a gift card raffle) for those participants who put forth sufficient effort on the experiment. Experimenters could also introduce safeguards into the experimental stimuli to prevent participants from rushing through the experiments. For example, researchers and practitioners could state the minimum amount of time a task should reasonably require to complete and refrain from displaying the submit button on the survey page until that time has elapsed. Future research is needed, however, to test whether implementing constraints into the experiment improves data quality.

4.2. Limitations and Future Research

This study has several limitations. The first limitation is the aforementioned discrepancy between our HIT completion and approval requirements for MTurk workers compared to other studies in the field. Because the qualification standards were comparatively minimal for this study, there are limitations to how well the results for our MTurk sample may generalize to other studies with stricter worker qualifications. The results may also fail to generalize to other populations outside of computer programmers. Thus, future research should attempt to replicate these findings with other worker requirements and different experimental tasks and populations.

Additionally, our laboratory sample was notably smaller than our MTurk worker sample, with 93 fewer participants completing the study in-person than on MTurk. The discrepancy in sample size can be largely attributed to the relative ease of recruitment and study completion on MTurk, allowing us to obtain data from substantially more participants in a shorter span of time than the process of recruiting and running participants *in vivo*. Regardless, because the in-person sample included fewer participants and those who were recruited were largely from Midwestern university samples, the generalizability of the laboratory findings may be comparatively limited relative to the MTurk sample. Thus, future research should attempt to collect computer programmers in person across multiple geographic areas.

Additionally, the current study utilized a sample of computer programmers. There may be differences between programmers and the general public in terms of personality constructs which may influence attention to details are careless responding. Future research should explore the current hypotheses in other samples utilizing different stimuli.

Finally, this study only featured four individual difference measures, which may have limited the validity of the CR indices in this study. Like other psychometric properties of psychological scales, the accuracy of the even-odd consistency indices and the Mahalanobis D index increases with the increased number of observations [8]. For example, the even-odd consistency should be able to detect CR more accurately when using multiple scales with sound psychometric properties. Stated simply, the greater the number of high-quality observations that are collected, the greater capability of detecting people who have inconsistent or aberrant responses. Because this study only contained four individual difference scales, this may have reduced the detection accuracy of these indices, particularly in relation to other studies considering CR, which typically incorporate a larger number of measures.

This study creates opportunities for future research in the detection and prevention of CR for online and *in vivo* studies. First, future research could examine whether participants put forth less effort online versus in-person on other types of tasks or other types of study instructions. For example, researchers could replace large blocks of text instructions with videos of the experimenter stating the instructions verbally. As researchers have shown that many participants don't read large text blocks [13], participants may be more likely to process text-based manipulations that are provided in a video. Furthermore, as this study focused on programmers for both samples, other studies could consider other specialized samples to determine if similar patterns of CR apply as well. Given the variety of worker filtering options available on MTurk, researchers could finely control their online sample recruitment while adjusting their in-person recruitment accordingly. Finally, this area of investigation could be extended to other popular online data collection platforms. Crowdsourcing platforms such as Prolific Academic and CrowdFlower have been compared to MTurk for data quality in a previous study [20]. However, that study did not compare the platforms for CR. Extending this line of research by comparing CR across crowdsourcing platforms would provide a more comprehensive understanding of where researchers can gather the highest quality data possible when developing an online study.

In total, we found that participants completing the experiment online submitted faster page times and had high long string values compared to when participants completed the experiment in the laboratory. Thus, researchers would need to remove more cases for inattentiveness if completing the experiment online. Unexpectedly, participants rated the computer code similarly across study locations. Although it may have been specific aspects of the code itself that influenced these ratings (e.g., low transparency in the code), future research should investigate whether these findings replicate to other types of computer code (e.g., highly transparent code).

5. Acknowledgements

This material is based upon work supported by U.S. Air Force Research Laboratory contracts FA8650-16-D-6616, 0003. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the U.S. Air Force. Distribution Statement A. Approved for public release : 88ABW Cleared 7/10/20, 88ABW-2020-2193.

6. References

[1] J.L. Huang, M. Liu, and N.A. Bowling, "Insufficient Effort Responding: Examining an Insidious Confound in Survey Data", Journal of Applied Psychology, American Psychological Association, United States, 2015, pp. 828.

[2] A.W. Meade, and S.B. Craig, "Identifying Careless Responses in Survey Data", Psychological Methods, American Psychological Association, United States, 2012, pp. 437.

[3] C.C.S. Kam, and J.P. Meyer, "How Careless Responding and Acquiescence Response Bias Can Influence Construct Dimensionality: The Case of Job Satisfaction". Organizational Research Methods, SAGE Publications, United States, 2015, pp. 512-541.

[4] M.R. Maniaci, and R.D. Rogge, "Caring About Carelessness: Participant Inattention and its Effects on Research", Journal of Research in Personality, Elsevier, Netherlands, 2014, pp. 61-83.

[5] C.M. Woods, "Careless Responding to Reverse-Worded Items: Implications for Confirmatory Factor Analysis." Journal of Psychopathology and Behavioral Assessment, Springer, Berlin, 2006, pp. 189-194.

[7] D.T. Berry, M.W. Wetter, R.A. Baer, L. Larsen, C. Clark, and K. Monroe, "MMPI-2 Random Responding Indices: Validation Using a Self-Report Methodology", Psychological Assessment, American Psychological Association, United States, 1992, pp. 340.

[8] D.A. Beach, "Identifying the Random Responder", Journal of Psychology: Interdisciplinary and Applied, Taylor and Francis, United Kingdom, 1989, pp. 101-103.

[9] P.G. Curran, Methods for the Detection of Carelessly Invalid Responses in Survey Data", Journal of Experimental Social Psychology, Elsevier, Netherlands, 2016, pp. 4-19.

[10] J.A. DeSimone, P.D. Harms, and A.J. DeSimone. "Best practice recommendations for data screening." Journal of Organizational Behavior, Wiley, United States, 2015, pp. 171-181.

[11] N.A. Bowling, J.L. Huang, C.B. Bragg, S. Khazon, M. Liu, and C.E. Blackmore, "Who Cares and Who Is Careless? Insufficient Effort Responding as a Reflection of Respondent Personality", Journal of Personality and Social Psychology, American Psychological Association, United States, 2016, pp. 218–229.

[12] A.M. Gibson, and N.A. Bowling, "The Effects of Questionnaire Length and Behavioral Consequences on Careless Responding", European Journal of Psychological Assessment, Advance online publication. Hogrefe Publishing Corp, Germany, 2019 https://doi.org/10.1027/1015-5759/a000526

[13] M.K. Ward, and S.B. Pond III, "Using Virtual Presence and Survey Instructions to Minimize Careless Responding on Internet-Based Surveys'. Computers in Human Behavior, Elsevier, Netherlands, 2015, pp. 554-568.

[14] D.M. Oppenheimer, T. Meyvis, and N. Davidenko, "Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power", Journal of Experimental Social Psychology, Elsevier, Netherlands, 2009, pp. 867-872.

[15] G. M. Alarcon, R. F. Gamble, S. A. Jessup, C. Walter, T.J. Ryan, D. W. Wood, and C. S. Calhoun, "Application of the heuristic-systematic model to computer code trustworthiness: The influence of reputation and transparency," *Cogent Psychology*, 2017, 1389640.

[15] International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences (http://ipip.ori.org/). Internet Web Site.

[16] J.T. Cacioppo, R.E. Petty, & C.F. Kao, "The Efficient Assessment of Need for Cognition". Journal of Personality Assessment, Taylor and Francis, United Kingdom, 1984, pp. 306-307.

[17] S.A. Jessup, T.R. Schneider, G.M. Alarcon, T.J. Ryan, & A. Capiola, "The Measurement of the Propensity to Trust Technology", In International Conference on Human-Computer Interaction Springer, 2019, pp. 476-489.

[18] C. Calhoun, P. Bobko, M. Schuelke, S. Jessup, T. Ryan, C. Walter…C. Stokes. "Suspicion, Trust, and Automation", SRA International Inc. Publication No. AFRL-RH-WP-TR-2017-0002, 2017.

[19] R. Kennedy, S. Clifford, T. Burleigh, R. Jewell, and P. Waggoner, "The Shape of and Solutions to the MTurk Quality Crisis", (SSRN Scholarly Paper No. ID 3272468). Retrieved from Social Science Research Network website: https://papers. ssrn. com/abstract, 3272468, 2018.

[20] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research". Journal of Experimental Social Psychology, Elsevier, Netherlands, 2017, pp. 153-163.