Replication of ‘Data quality of platforms and panels for online behavioral research’ by Peer et al. (2021, Psychological Science)

Kimia Saadatian ([kimia@stanford.edu](mailto:kimia@stanford.edu))

October 17, 2021

Table of Contents

[Introduction 2](#_Toc85394814)

[Justification for choice of study 2](#_Toc85394815)

[Anticipated challenges 2](#_Toc85394816)

[Links 2](#_Toc85394817)

[Methods 2](#_Toc85394818)

[Power Analysis 2](#_Toc85394819)

[Planned Sample 2](#_Toc85394820)

[Materials 3](#_Toc85394821)

[Procedure 3](#_Toc85394822)

[Analysis Plan 4](#_Toc85394823)

[Differences from Original Study 4](#_Toc85394824)

[Methods Addendum (Post Data Collection) 4](#_Toc85394825)

[Results 4](#_Toc85394826)

[Data preparation 4](#_Toc85394827)

[Confirmatory analysis 4](#_Toc85394828)

[Exploratory analyses 4](#_Toc85394829)

[Discussion 5](#_Toc85394830)

[Summary of Replication Attempt 5](#_Toc85394831)

[Commentary 5](#_Toc85394832)

## Introduction

### Justification for choice of study

I will be replicating a recent study by Peer et al. (2021), which compared data quality (attention, comprehension, reliability and honesty) between multiple sites used for online participant recruitment: Amazon MTurk, CloudResearch (formerly TurkPrime) and Prolific.

### Anticipated challenges

As of right now, I do not anticipate running into any challenges in replicating the results from the original study.

### Links

Project repository: <https://github.com/psych251/saadatian2021.git>

Original paper: <https://github.com/psych251/saadatian2021/blob/3c2eeb3b8a352c51dea439f7878b2ab95f67acff/original-paper/Peer%20et%20al.%20(2021).pdf>

Original study’s OSF: <https://osf.io/342dp/>

## Methods

### Power Analysis

Main analysis I am replicating is "We found statistically significant differences between the sites on this measure (overall data quality score), F(2, 1458) = 129.4, p < .001, which showed higher scores for Prolific and CR (M = 5.87, 5.78, SD = 1.0, 1.1, respectively) compared to MTurk (M = 4.55, SD = 1.9)).

I will use an effect size converting spreadsheet to convert the f statistic to an r value, get square root of r, and plug that into g\*power to get a projected sample size of – participants for power = .95

### Planned Sample

I will recruit participants on Amazon MTurk, CloudResearch, and Prolific. Participants must be U.S. residents age 18 or above. Similar to the original study, I will exclude participants who do not complete all of the study.

“We recruited 500 participants from each platform (MTurk, CR, and Prolific), who reported residing in the United States, in March 2021. Participants were paid 1.5 USD on CR and MTurk and 1.1 GBP on Prolific plus a bonus of up to 0.5 USD/GBP. We applied data quality prescreening filters on all sites by restricting the study to participants with at least 95% approval rating and at least 100 previous submissions; on CR we also used the site setting to “block low data quality workers.” We excluded participants on Prolific who completed the previous study on Prolific and participants on CR who completed the previous study on CR or MTurk. However, because our study had to be posted twice on MTurk (once through our MTurk account and once through our CR account), 39 participants completed the study twice, and we removed their later submission (although they were still paid for their submissions). The final sample thus included 1461 participants who completed the study. Table 4 presents the samples. Additional demographics can be found in the Appendix."

### Materials

“1. Attention is measured using two attention-check questions. The first asks participants to answer”six" and “three” to two items regardless of their actual preference (other responses are coded as failures); the second is an item within a scale worded “I currently don’t pay attention to the questions I’m being asked in the survey” (response other than “strongly disagree” is coded as a failure) 2. Comprehension is examined through the oral summaries of instructions to two tasks. The first task asks participants to identify faces in a picture, but includes an instruction to only report zero; the second includes instructions for the “Matrix task” (Mazar et al., 2008). Two raters will code the oral responses independently and blind to the origin of the participant. We will consider an answer correct if both readers agree it is correct, and will apply a third reader to any answer with splits. 3. Reliability will be measured using Cronbach’s alpha measure for the Need for Cognition scale. 4. Honesty will be measured using an online version of the Matrix task (Mazar et al., 2008) that will include two unsolvable matrices. Reporting solving any of these two problems will be coded as a dishonest response. Additionally, we will examine whether participants lie about their eligibility for a future study by asking them to indicate if they want to be invited to a study that samples participants of their own gender but whose age will be described as 5-10 years above the age participants reported in the beginning of the study.

In addition we will examine drop-out rates, duration for completing the survey, overall response time and speed between sites, differences in NFC, demographics, and patterns of usage of the site (main purpose, frequency of usage, number of submissions and approval ratings, usage of other sites), and we will also ask participants to report did they complete a study similar to this study in the last months."

### Procedure

“Participants were invited to complete a survey on individual differences in personal attitudes, opinions, and behaviors. All participants began the survey by answering demographic questions, followed by the data quality measures described below. Participants finished the survey by answering questions related to their usage of the online platform including how often they use the site, for what purposes, how much they earn in an average week, their percent of approved submissions (responses that participants submit and are approved by the researcher), and how often (if at all) they use other sites.”

“We will apply data quality filters on all three sites in the following scheme: 1) MTurk - only workers who have completed 100 submissions or more and 95% of those submissions were approved. 2) CloudResearch - only workers who have completed 100 submissions or more and 95% of those submissions were approved + the default setting on CloudResearch of”block low data quality workers"

### Analysis Plan

I plan to follow the exact analyses that were conducted in the original study.

The main analysis I am replicating from study 2nwhere Peer et al. (2021) found higher data quality scores for data obtained from Prolific and CloudResearch (M = 5.87, 5.78, SD = 1.0, 1.1, respectively) compared to MTurk (M = 4.55, SD = 1.9; F(2, 1458) = 129.4, p < .001).

In order to compute the overall data quality score, the original authors “used chi-square tests to examine differences in the rates of attention, comprehension and dishonesty. They also computed average scores for (across the items per aspect) and compared them between sites using ANOVA and regression analyses. They tested for differences between reliability coefficients using Hakistan & Whalen (1976) method. Lastly, they computed an overall composite score of data quality (based on the average scores of attention, comprehension and dishonesty) and compare them between sites using ANOVA and regression.”

### Differences from Original Study

I do not anticipate any deviations from the original analysis plan at this stage.

### Methods Addendum (Post Data Collection)

You can comment this section out prior to final report with data collection.

#### Actual Sample

Sample size, demographics, data exclusions based on rules spelled out in analysis plan

#### Differences from pre-data collection methods plan

Any differences from what was described as the original plan, or “none”.

## Results

### Data preparation

Data preparation following the analysis plan.

### Confirmatory analysis

The analyses as specified in the analysis plan.

*Side-by-side graph with original graph is ideal here*

### Exploratory analyses

Any follow-up analyses desired (not required).

## Discussion

### Summary of Replication Attempt

Open the discussion section with a paragraph summarizing the primary result from the confirmatory analysis and the assessment of whether it replicated, partially replicated, or failed to replicate the original result.

### Commentary

Add open-ended commentary (if any) reflecting (a) insights from follow-up exploratory analysis, (b) assessment of the meaning of the replication (or not) - e.g., for a failure to replicate, are the differences between original and present study ones that definitely, plausibly, or are unlikely to have been moderators of the result, and (c) discussion of any objections or challenges raised by the current and original authors about the replication attempt. None of these need to be long.