Transactions on
Replication Research

Methodological Replication DOI: 10.17705/1atrr.00058 ISSN 2473-3458

A Replication of Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research – Sometimes Preferable to Student Groups

Troy L Adams
Ira A Fulton Schools of Engineering
Arizona State University
troy.l.adams@asu.edu

Yuanxia Li
Eller College of Management
The University of Arizona
yuanxiali@email.arizona.edu

Hao Liu
Eller College of Management
The University of Arizona
liuhao16@email.arizona.edu

Abstract:

This study is a replication of one of two studies found in "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research" (Peer, Brandimarte, Samat, & Acquisti, 2017). We conduct an empirical analysis and comparison between two online crowdsourcing platforms, Amazon Mechanical Turk (MTurk) and Prolific Academic (ProA), as well as to a traditional student group. The online crowdsourcing platform (e.g., MTurk and others) used for years as a launching point for many types of microwork, including academic research. Today, MTurk has several competitors, including one that was built to focus on research tasks, ProA. Across the four segments, we reinforce the original study by finding both MTurk and ProA to provide inexpensive, reliable, and significantly faster methods of conducting surveys over traditional methods. Our results indicate that ProA provides superior service. By centering on research, ProA results are similar to MTurk's. However, ProA's response and completion rates, diversity, attention, naivety, reproducibility, and dishonest behavior are better.

Keywords: Online research, Crowdsourcing, Data quality, Amazon Mechanical Turk

Prolific Academic, Replication

The manuscript was received 10/02/2018 and was with the authors 17 months for 4 revisions.

1 Introduction

Amazon's Mechanical Turk (MTurk) and other online crowdsourced platforms are popular for conducting research surveys, among other tasks. They are understandably very popular for researchers interested in a fast and reasonably reliable replacement for traditional forms of conducting surveys. Several studies have sought to compare and contrast online crowdsourced platforms and evaluate their value as research tools (Casler, Bickel, & Hackett, 2013; Mason & Suri, 2012; Peer et al., 2017). Some studies focus on the comparison between traditional students and MTurk's participants (Smith, Roster, Golden, & Albaum, 2016). However, we believe that understanding the characteristics of online crowdsourced platforms would be equally important, as they have gained increasing popularity in collecting research surveys.

In one such example completed in 2016, several researchers conducted a pair of studies entitled *Beyond the Turk: Alternative platforms for crowdsourcing behavioral research (BTT)*, which sought to evaluate the data quality of several of these platforms when conducting surveys, as well as with more traditional methods (i.e. a group of student participants) for comparison (Peer et al., 2017). Among the works comparing online crowdsourcing platforms (Bentley, Daskalova, & White, 2017; Lutz, 2015; Peer et al., 2017), Peer et al. (2017) tested the most complete set of constructs, which provide the essential information for a broader audience.

In the original set of studies depicted in BTT, one sought to compare a number of results from three different online crowdsourcing alternatives, MTurk, ProA, and Crowdflower as well as one traditional group of student participants, and the other sought only to compare results from MTurk and Prolific Academic (Peer et al., 2017). BTT's measurements included an examination of several factors between platforms and used questionnaires and experimental tasks adopted from prominent psychology studies to assess data quality, to include reliability of data (via Need for Cognition and Rosenberg Self-Esteem scales), participant attention (via attention check questions), non-naivety (via survey familiarity questions), reproducibility of known effects (via Asian Disease framing effect questions, among others), and dishonesty (via post-hoc statistical comparison) (Peer et al., 2017). Between each platform alternative, in the first study researchers sought to compare participant behavior in their response rates, attention, reliability, reproducibility, non-naivety, and dishonesty; but also overlap between alternatives, and participant demographics and usage patterns (Peer et al., 2017). In their second study, the BTT researchers only compared MTurk and ProA, using the same measurements as in their first study, but with deeper demographic analysis (Peer et al., 2017). Their findings indicate that although trade-offs exist, several options are just as good or better in some respects (e.g., speed and reproducibility) to more traditional methods of conducting surveys (Peer et al., 2017).

This work seeks to replicate the original Beyond the Turk study #1 (but not #2) by collecting and analyzing a similar set of survey responses. We replicated only study #1 for two reasons: First, all the platforms (MTurk and ProA) and constructs (e.g., attention, reproducibility, and others) in study #2 repeated study #1. By replicating only study #1, we will be able to compare our results with both studies in the original paper and focus at the heart of the comparison, which was between a few popular online crowdsourced platforms and their relative effectiveness against a traditional group composed of students. Second, the comparison in study #2 is only between two online platforms, and the primary purpose of it is to compare the two online platforms more convincingly with a more significant number of participants. As we are also interested in comparing online platforms and traditional participant pools, we consider study #1 to be appropriate. Further, to simplify the study design and maintain our budget, we focused only on the two online crowdsourcing platforms highlighted in study #2, MTurk and ProA, and a traditional study group made up of predominantly student participants as done in study #1. In our case, we did not use the same student group (e.g.,the original used Carnegie Mellon University's CBDR). To the best of our knowledge, we are the first to conduct a replication study on comparisons among online crowdsourcing platforms and traditional participants with the focus on behavioral research.

As in the original BTT studies, we used a survey to evaluate the data quality of two online crowdsourced productivity platforms, Amazon's Mechanical Turk (MTurk.com) and Prolific Academic (prolific.ac) and compared them with results from a traditional resource outlet, a student participant pool from a set of sections of a Management Information Systems course at a large university in the American southwest.

To the uninitiated, MTurk and its competitors seem to be bizarre and yet amazingly intuitive technologies that could only evolve in an internet-enabled society. By Amazon's statement, "Amazon Mechanical Turk (MTurk) is a marketplace for work that requires human intelligence… [which] ... gives businesses access to a diverse, on-demand, scalable workforce, and gives Workers a selection of thousands of tasks to complete whenever it's convenient" (Amazon, 2018). Created in 2001, MTurk provides a platform for microwork and online outsourcing, though not necessarily for achieving academic goals, and supplying several hundreds of thousands of individuals a means to earn income (Prpić, Taeihagh, & Melton, 2015). MTurk workers, or ‘Turkers’, consist of a diverse population, with varying ages, genders, ethnicities, income levels, and educations (Casler et al., 2013), as well as reputation and performance incentives, encourage Turkers to generally provide a quality service (Peer et al., 2017). Academics have used MTurk for years as a cheap and fast method to acquire survey data whose quality is reasonably comparable to that of traditional methods of survey data collection (Paolacci & Chandler, 2014).

Prolific Academic (ProA) describes itself as the “world’s largest crowdsourcing community of people who love science” (Prolific, 2018), and is known as a popular and research focused alternative to MTurk (Palan & Schitter, 2018). Constructed in 2014, ProA is more than a decade younger than MTurk and built with a focus on academic research. ProA workers are diverse in terms of ethnicity, education, and income, though understandably differently diverse, given its core user distributions are weighted significantly higher in the US and UK versus MTurk’s high concentrations in the US and India (See Appendix Figure A2 for distribution). ProA has a significantly smaller population, 41,000 as of April 2018, up from 35,600 from December 2017 (Palan & Schitter, 2018) compared to 100,000 to 200,000 Turkers (Difallah, Filatova, & Ipeirotis, 2018; Prpić et al., 2015). However, ProA workers are generally paid better than Turkers and are similarly incentivized to provide a quality service (Palan & Schitter, 2018).

Several other online crowdsourcing platforms exist. However, MTurk's benchmark, as a popular choice for microworkers, is a consistent comparison. These studies, including the one we replicate (Peer et al., 2017), compare online crowdsourcing platforms such as MTurk, ProA, CrowdFlower, and others, and against traditional groups of undergraduate students (Palan & Schitter, 2018). Taken together, the vital sample characteristics of reliability, reproducibility, naivety, and dishonesty indicate that ProA is the superior service for accomplishing research goals.

2 Method

2.1 Sampling and Participants

Our study included an online survey conducted with three groups. With varying success, we sought at least 150 successful completions from each group. To use a standard timeframe between each group, we limited recruitment time to the week between March 30, 2018, and April 6, 2018. Table 1 depicts sample sizes, dropout rates, and worker demographics from this study. No restrictions or participant requirements existed for participants of any platform, with the exception that participants from our student group actually be students. Participants were compensated per response as follows: MTurk, $1, ProA, $1.63, and students with course credit. Our compensation on Mturk and for students is consistent with the original study ($1/£1 per participant) (Peer et al., 2017), but due to the increase in minimum payment at ProA, we increased the compensation from $1.47 (£1) in the original study to $1.63 (£1.17). Despite the increase, our ProA compensation maintains the same proportion to the minimum payment as the original study. In addition to the participation compensation, as mentioned in the dishonest behavior section below, MTurk and ProA participants were eligible to also earn a bonus, up to $0.60, by making an appropriate selection in that section of the survey.

We found that samples from three sources are statistically significant different in terms of ethnicity, $p < 0.01$, education, $p < 0.01$, income, $p < 0.01$, and location, $p < 0.01$. Given the restriction of our sample size, we do not observe any individual for some levels. Therefore, the assumption for the Chi-squared test is not satisfied, and we base results above on a Monte Carlo Simulation with 50,000 replicates, which is an alternative for the chi-squared test (Hope, 1968). Each of our samples had participants in a variety of reported ethnic backgrounds, income and education levels, and locations, with none of them appearing to be remotely equal to the other, and with the unsurprising exception that our student group consisted entirely of individuals within the United States. The cumulative histograms of demographics are included in Appendix A. The comparisons to the original study are also provided.

2.2 Procedure and Materials

As a replication study, we sought to adopt the same survey questions and logic (including randomization settings) as the BTT authors used in their original survey (Peer et al., 2017). As in the BTT study, our model included traditional research groups that we exempted from specific questions. The student group was exempted (due to questionable applicability) from answering duration, task, and income-generating related questions concerned with membership with their online crowdsourcing platform or other online crowdsourcing platforms. Reflecting the original BTT survey, our study also included questions attributed to prominent psychology studies, including the Rosenburg Self-Esteem Scale (RSES) (Rosenberg, 1965) and the Need for Cognition scale (NFC) (Cacioppo, Petty, & Feng Kao, 1984), which assessed data quality. These questions intended to collect demographic information and test for dishonest behavior.

Several measures were used to examine participant responses.

- Participant attention was measured using four attention check questions embedded throughout the survey, which helped to determine if participants were reading the questions and appropriately following directions. Nonsensical responses indicated that the participant was not reading and reasonably responding to their questionnaire.
- Naivety was measured using familiarity questions. These followed only the measure questions discussed in this section (i.e., naivety questions did not follow demographic, consent, or honesty questions), and asked if it was the first time that participants had ever seen such a question. Indications of familiarity with the questions indicated that the participant had seen and was at least somewhat prepared to answer, given prior reflection.
- Reproducibility of known effects was measured using four personal judgment tasks. These included the Asian Disease Framing question (Tversky & Kahneman, 1981), the Sunk Cost Fallacy question (Oppenheimer, Meyvis, & Davidenko, 2009), the Retrospective Gambler's Fallacy question (Oppenheimer & Monin, 2009), and the Quote Attribution question (the original BTT study conceptually replicated from Lorge & Curtiss, 1936) and sought to measure and compare known effects from other behavioral studies.
    1. The Asian Disease Framing question asked participants to choose the preferable one of two programs. Each of these involved the outbreak of disease and randomly presented as either a positively (lives saved) or negatively (lives lost) framed outcome. These included Program A, where uneven numbers of lives were saved or lost (200 survive or 400 die), or Program B, where uneven probabilities of lives were saved or lost (i.e., ⅓ probability that all survive or ⅔ probability that all die).
    2. The Sunk Cost Fallacy question asked participants randomly to choose whether to attend a desirable sporting event in an unpleasant climate after being given high-value tickets or to do so after personally spending a large sum for the tickets.
    3. The Retrospective Gambler's Fallacy question describes a scenario where a gambler is witnessed to have rolled three dice, with an outcome that participants were randomly shown one of two alternatives. These alternatives included (1) where each of the three dice displays a six, or (2) where one die displays one and two dice display three. The participant was then asked to input the number of times they imagined the dice were rolled before they walked by.
    4. The Quote Attribution question asked participants the degree to which they agreed with the following quote randomly attributed to either George Washington or Osama Bin Laden: “I have sworn to only live free, even if I find bitter the taste of death” (a statement attributed to both individuals).

The survey ended with three other response tasks, given only to our online platform participants, and not to our student group. These sought to gather the participant's demographic facts (e.g., gender, age, income), their online crowdfunding platform facts (e.g., length and frequency of use, regular platform income level, and whether other platforms are also used), and to measure their dishonesty. Much like with the original BTT studies, to measure dishonesty, participants were given an opportunity to earn an additional completion bonus, composed of the product of a $0.10 base and a multiplier value, which was displayed on a randomly generated die face. Before it was rolled, participants were encouraged to mentally choose the top face or bottom face of the die. The die was then virtually rolled (via random number generation) and displayed. Participants were then given a choice to select a top or bottom die face, based on their prior mental selection, and the number selected would then be multiplied by $0.10 and provided as a bonus. Due to the anonymity inherent with mental selection, participants were incented to cheat by merely selecting the die face with the greater number, as the greater number would grant a larger bonus. Detection of dishonesty could only be determined in the aggregate and not from individual selections, by comparing the total payout to the average payout.

3 Results

3.1 Dropout Rates

Table 1. Completion Distribution - Sample size, dropout rate, and sample demographics
Bold Text = This Study, Normal Text = Original BTT Study (Peer et al., 2017)

Sample	Started the study		Completed		Percentage of dropouts		Percent Males		Median Age	
MTurk	**196**	220	**162**	201	**17.3%**	8.6%	**62.3%**	56.7%	**30 (20 - 77)**	32 (27-38.5)
ProA	**197**	243	**177**	214	**10.2%**	11.9%	**46.3%**	64.5%	**26 (18 - 67)**	27 (23-27)
Students	**267**	215	**232**	195	**13.1%**	9.3%	**47.8%**	29.2%	**21 (19 - 42)**	23.5 (23-37)

The overall dropout rate across all the platforms was around 13%, which was very similar to the 10% found in the original BTT study (Peer et al., 2017) and detected no significant difference between the platforms (χ2(2) = 0.32, p=0.85). The dropout rate information for each platform can be found on Table 1.

In order to fix a broken link in the compensation protocol on ProA, we disabled the survey for 30 minutes. Therefore, we removed the first data point on ProA (which was collected before we disabled the survey) when we analyzed the amount of time it takes to collect responses. As shown in Figure 1, collectively, Turkers provided the fastest overall study completion, followed closely by ProA participants. The relatively slow increase in ProA completions could be attributed to a number of issues inherent to ProA projects that delay their availability to participants. To reach the intended 150 participants in each group, on average, it took 5.00 minutes to collect 10 responses from MTurk, 23.56 minutes to collect 10 responses from ProA, and 436.75 minutes (around 7 hours) to collect 10 responses from the student participants pool.

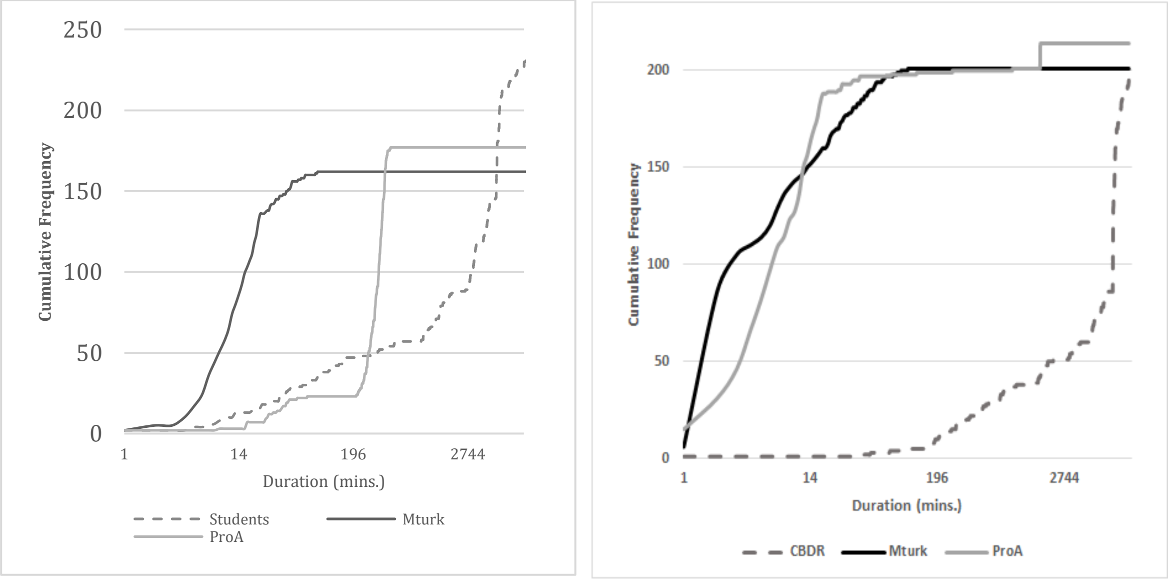

Figure 1. Time to Study Completion - Response Rate (In Minutes) Between Platforms
Left Panel = This Study
Right Panel = Adapted from Original BTT Study #1 (Peer et al., 2017)

Each sample had a different average survey completion duration. The median was lowest for the student group (8.8 minutes), followed by MTurk (10.6 minutes), and ProA (14.7 minutes). The result of the Kruskal-Wallis test shows that these differences were significant. The difference here may be explainable by the lack of incentive for participants to provide a quality response, perhaps due to participants' relative value (or lack thereof) for their lasting reputation as a study participant.

3.2 Attention

To confirm adequate engagement with our survey, we challenged each with four attention check questions (ACQ). Nonsensical responses indicated that the participant was not reading and reasonably responding to their questionnaire. We examined the remaining percentage of participants that passed all ACQs (strict exclusion), and those that failed only one of the four (lenient exclusion). We found significant attention differences between the platforms, for strict rule ($\chi2$ (2) = 16.56, $p < 0.01$), and for lenient rule ($\chi2$ (2) = 19.34, $p < 0.01$). Figure 2 depicts the success rates given these two exclusion policies. Results for MTurk and ProA were dissimilar from the original BTT study, where Turkers generally proved more attentive than participants from ProA (Peer et al., 2017). In our study, ProA participants generally performed better (6% to 12%) than Turkers, but both online crowdsourced participant groups performed significantly better (10% to 41%) than our student group in terms of attentiveness.

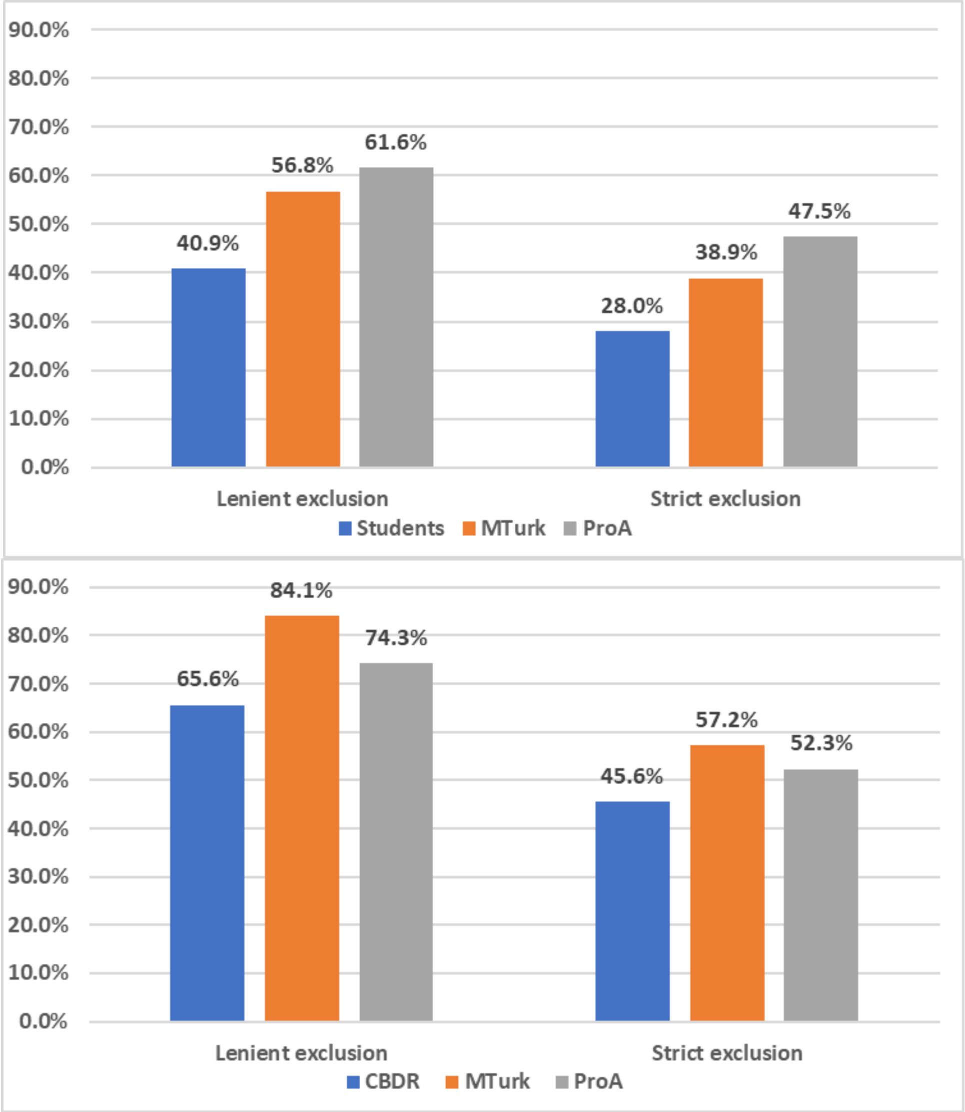

**Figure 2. Attentiveness - Success Rates Given Strict And Lenient Exclusion Policies.
(Lenient = Failure Of More Than One ACQ. Strict = Failure Of Any ACQ)
Top Panel = This Study
Bottom Panel = Adapted from Original BTT Study #1 (Peer et al., 2017)**

The result of ANOVA shows that the average number of failed ACQs differed significantly across the platforms (F (2, 568) = 6.812, p = 0.001). Participants from ProA performed best with the least average number of failed ACQs (mean = 1.23, SD = 1.377), followed by participants from MTurk (mean = 1.54, SD = 1.536), and student participants (mean = 1.75, SD = 1.325). The post-hoc differences show that only the differences between ProA and the student group remain significant after applying Bonferroni's correction (p<0.05). In our replication, ProA participants showed the highest propensity to follow instructions, with MTurk participants earning a close second. As pointed out in the original BTT paper, we checked if English proficiency could account for some of the failing attention check questions (Peer, et al., 2017). We found that across all the platforms, the results replicate the original BTT study. Participants who reported their English proficiency as "Good", "A little", or "Not at all" (N = 31, 5.4%) failed more ACQs (mean = 2.84, SD = 1.167) than participants who reported their English proficiency as "Excellent" or "Very good" (mean = 1.45, SD = 1.397). The difference was significant according to Welch's t test (t(35) = 6.26, p < 0.001). Results suggest (as with CrowdFlower in the original BTT study) that participants may fail more attention check questions because of their unfamiliarity with English (Peer et al., 2017), or based on the assumption that failing more ACQs may also be an indication of a higher degree of naivety and sincerity. Using the same reasoning, student participants may have failed more attention check questions because of this naivety. We also considered the number of failed ACQs in later data analysis.

3.3 Reliability

Following the method from BTT, reliability (Cronbach's alpha) is calculated and compared between the platforms using the Rosenberg Self-Esteem Scale (RSES) (Rosenberg, 1965) and the Need for Cognition Scale (NFC) (Cacioppo et al., 1984). As shown in Figure 3, reliability scores for RSES is at or above 0.90 for all platforms. For NFC, all of the groups performed relatively well, and reliability scores increase with the application of lenient or strict exclusion rule. Based on a Chi-squared test, we didn't find any significant difference between the platforms and their subgroups, which replicates the results in the original BTT study.

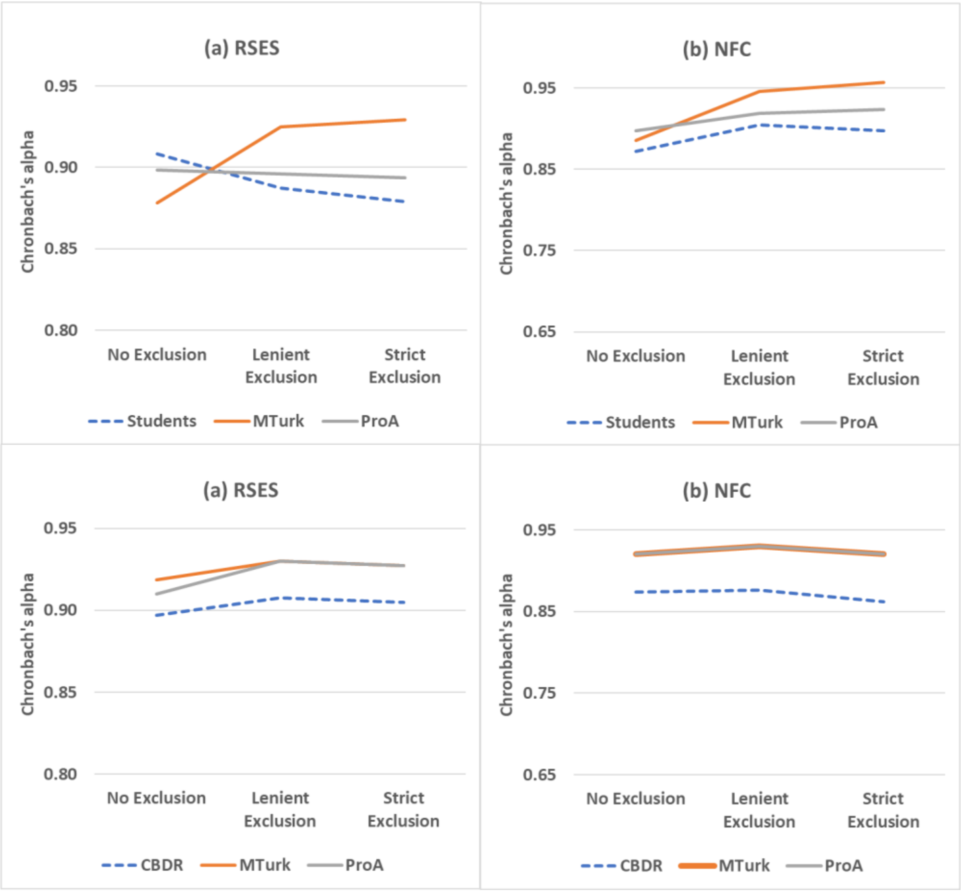

Figure 3. A-B - Reliability - Relative Measures Of Cronbach's Alpha For RSES (3a) and NFC (3b) For Each Platform As Function Exclusions (Lenient = Failure Of More Than One ACQ. Strict = Failure Of Any ACQ). Top Panels = This Study Bottom Panels = Adapted from Original BTT Study #1 (Peer et al., 2017)

3.4 Reproducibility

We examined the reproducibility of known effect sizes of four experimental tasks including Asian Disease, Sunk Cost, Gambler's Fallacy, and Quote Attribution. Studying effect sizes by excluding inattentive participants involves a tradeoff: Culling potentially large numbers of valuable data points diminishes the highly desirable quantity of data. Such culling could affect the generalizability of its analysis by cutting out subjects that are important representatives of the population. However, in turn, this helps us to better understand the marginal effects that these elements have on the data, and it also leaves potentially better quality, more reliable, and even more desirable data points from which to potentially draw better quality conclusions; conclusions that are more generalizable of the population. As demonstrated in Table 2, all effects are statistically significant, with the exception being MTurk participants' Sunk Cost task (under no exclusion). By applying the exclusion policy, the effect size of even this exception was increased and became significant as well.

Overall, exclusion policies increased the effect size of MTurk. For ProA, exclusion policies had a mixed effect, increasing the effect size for each element except for Quote Attribution. For the student group, the exclusion policy also had a mixed effect, with an increased effect size for Gambler's Fallacy and Quote Attribution, but decreased effect size for Asian Disease and Sunk Cost. The effect size of strict and lenient exclusion policies is nearly the same for each of MTurk, ProA, and the student group.

Table 2. Reproducibility - Effect Sizes (Cohen's d) Between Platforms And Exclusion Policies. All Effect Sizes Were Statistically Significant Except Those Indicated With *
Bold Text = This Study, Normal Text = Original BTT Study #1 (Peer et al., 2017)

	Exclusion policy	Asian Disease		Sunk cost		Gambler's Fallacy		Quote attribution	
MTurk	None (all Ps.)	**0.61**	0.82	**0.10 ***	0.27	**0.23**	0.28	**0.31**	0.73
	Lenient exclusion	**0.83**	0.99	**0.36**	0.34	**0.7**	0.29	**0.59**	0.75
	Strict exclusion	**0.84**	0.94	**0.48**	0.24	**0.68**	0.24	**0.85**	0.73
ProA	None (all Ps.)	**0.53**	0.63	**0.24**	0.39	**0.64**	0.29	**0.52**	0.68
	Lenient exclusion	**0.77**	0.74	**0.36**	0.61	**0.67**	0.36	**0.39**	0.66
	Strict exclusion	**0.64**	0.82	**0.28**	0.53	**0.66**	0.31	**0.46**	0.72
Students	None (all Ps.)	**0.69**	0.76	**0.53**	0.42	**0.23**	0.12	**0.66**	0.51
	Lenient exclusion	**0.57**	1.11	**0.54**	0.41	**0.32**	0.14	**0.77**	0.28
	Strict exclusion	**0.46**	1.12	**0.41**	0.56	**0.54**	0.25	**0.85**	-0.01

3.5 Non-naivety

Non-naivety values indicate responding participants to the familiarity of study questions. Therefore, their potential for reducing effect size or making studies using their services non-generalizable, implying that participants prepared for study instruments are generally less desirable than those who are not (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015).

Non-naivety was measured using familiarity questions, which followed many other questions, including ACQ, Asian Disease, Sunk Cost, Gambler's Fallacy, Quote, NFC, and RSES questions, but did not follow demographic, consent, or honesty questions. Participants were asked if exposure to the question was the first time they had seen such a question. We interpreted 'yes' as an indication of naivety, and 'no' or 'not sure' as an indication of familiarity. As shown in Figure 4, the most familiar tasks included NFC and RSES questions.

Using the percentage of unfamiliarity tasks as the overall score of 'naivety', we carried out an ANOVA, which shows statistically significant differences between the platforms (F(2,568) = 9.055, p<0.001). Figure 4 depicts relative differences between groups, with participants from ProA performing best in terms of overall naivety (mean=0.79, SD =0.25), followed by participants from the student group (mean=0.70, SD=0.27), MTurk (mean=0.69, SD=0.28). The original BTT study found that their student group (CBDR) indicated naivety levels at or above participants from each crowdsourced platform. However, results from our study indicated that ProA participants performed consistently more naive in our study, against not only MTurk, but also against our student group.

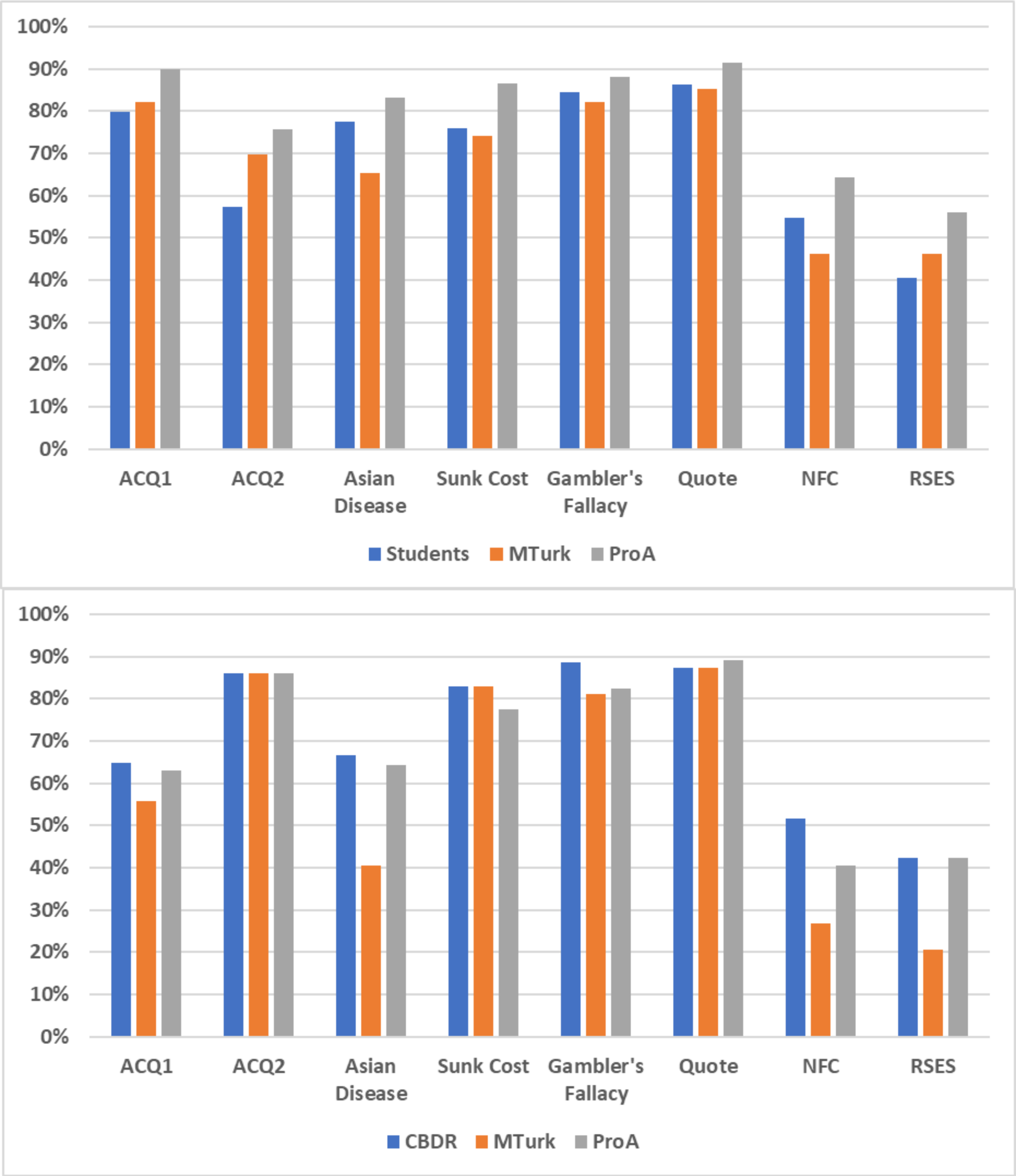

Figure 4. Non-naivety - Percentage Of Naive Participants (Indicated Unfamiliarity) Per Task Per Platform.
Top Panel = This Study
Bottom Panel = Adapted from Original BTT Study #1 (Peer et al., 2017)

3.6 Dishonest behavior

As in the original BTT study, giving participants a chance to earn bonus provided an opportunity to gauge dishonesty. Both MTurk and ProA samples were given this opportunity, but the student group was not; due to a lack of a mechanism of bonus distribution. Following the same design as the original study, the mean bonus claimed by participants on MTurk and ProA should be 35 cents (calculated by an average of each of the six equally probable bonus payments, $0.10 to $0.60). Therefore, to see if there existed an aggregate over-reporting of earned bonuses, we examined the mean bonus claimed by participants. We found significant degrees of over-reporting on both MTurk (mean=44.57, SD=14.54) and ProA (mean=42.60, SD=15.85), which were similar to the original BTT study findings for dishonesty. In the original paper, the mean of MTurk is 46.87, with standard deviation being 12.67, and the mean of ProA is 42.29, with standard deviation being 15.8. The original BTT study found that MTurk had a significantly higher rate of cheating than the other platforms they studied (ProA and CrowdFlower) (Peer et al., 2017). However, in our study, we found there was no statistically significant difference between bonuses claimed by participants of either platform, t(337)=1.193, p=0.234

3.7 Overlap of participants between platforms

To measure the overlap of participants between platforms, we followed the design of original BTT study #1, asking participants to indicate their frequency of use for the platforms, and provided the results in Table 3. As with the BTT original study, we exempted our traditional research pool group from questions in this part of our study. The results show that the degree of overlap between MTurk and ProA is similar to the results found in the original BTT study, with 22% of Turkers also being members of ProA, where only about 14% of ProA members were also Turkers. Rationale for the disparity between MTurk users reporting MTurk usage or ProA users reporting ProA usage could be explained by the recency of their usage or could indicate that participants did not understand the question or the name of their platform.

Table 3. Platform Overlap - Percentage of participants reporting using platforms more than "a few times"
Bold text = this study
Normal text = Adapted from BTT Study #2 (Peer et al., 2017)

	Uses MTurk		Uses ProA	
MTurk participants	**95.06%**	98.50%	**22.84%**	14.5%
ProA participants	**14.12%**	22%	**60.45%**	88.8%

3.8 Usage patterns

As can be seen in Figure 5, 43.8% of Turkers report spending more than 8 hours per week on MTurk. By comparison, 59.9% of ProA participants report spending less than 2 hours per week on ProA. This difference results in the earning difference between the platforms. 60% of Turkers report earning more than 50 dollars per week and about 67% of ProA participants report earning less than 5 dollars per week. This generally replicates the results in the original BTT study. The differences between the two platforms were statistically significant, F(1,298)=5.815, p=0.016. On average, MTurk participants reported much higher earnings per week with greater variance (mean=432.23, SD=2166.50), compared to ProA participants, (mean=4.23, SD=4.00). That Turkers have a wider variety of potential tasks not related to research and that ProA participants focus more on research related tasks may indicate the reason behind these differences.

The median number of tasks participants reported completing on the platform had a stark difference in the typical activity between Turkers and ProA participants. On average, Turkers reported having completed 3,000 tasks, while at ProA's average was a comparably minuscule 20 tasks. Performance reputation, perhaps known as the intrinsic value that participants of online crowdsourcing platforms create with the investment of their active engagement and quality completion of tasks, has a direct effect on the potential for future assigned tasks (and therefore future potential revenue). Unsurprisingly the median approval score for Turkers is 99.05%, which was nearly the same as the average for ProA participants (99.00%).

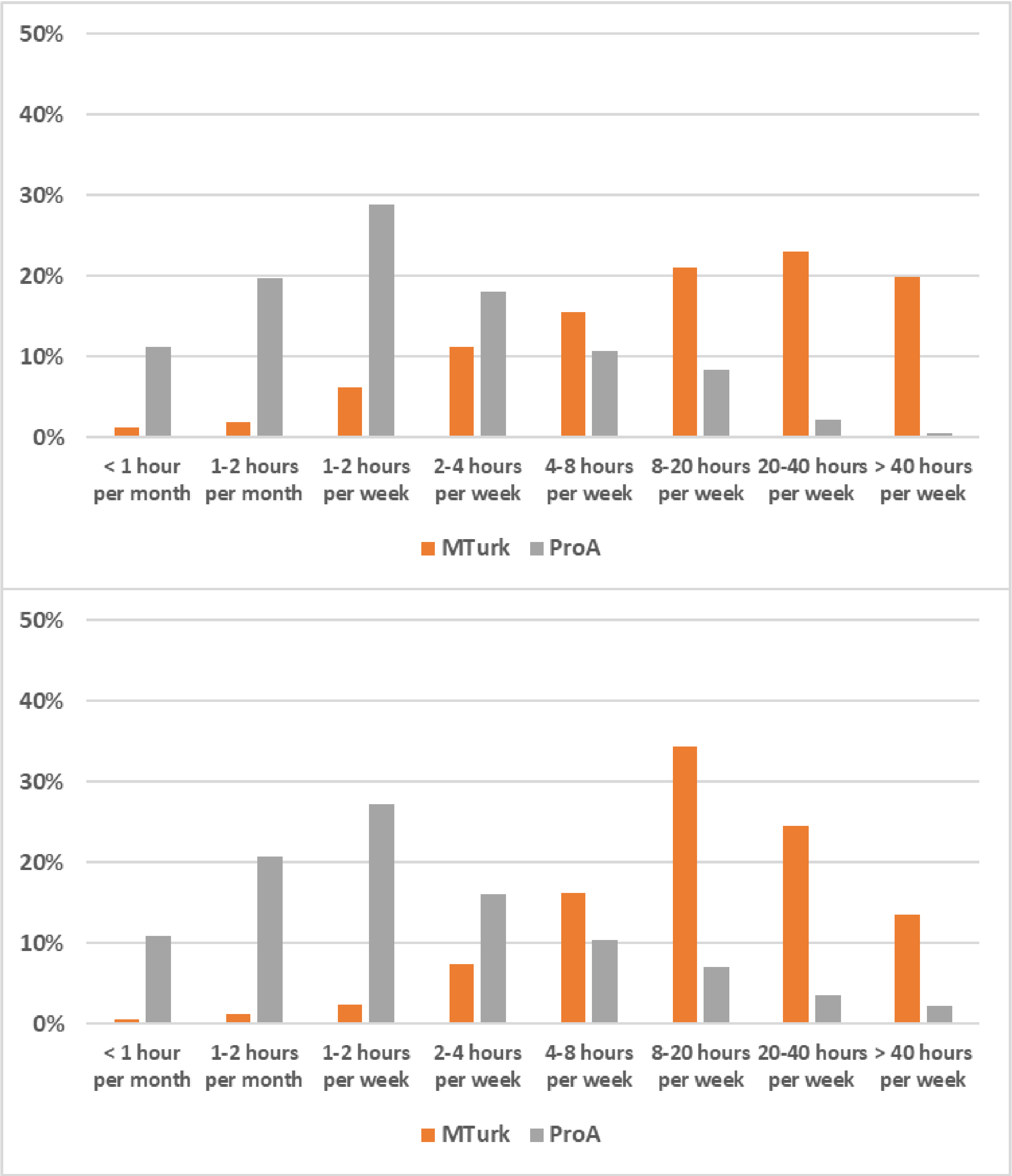

Figure 5. Site Usage - Usage Pattern Distribution Of Online Crowdsourcing Platforms
Top Panel = This Study
Bottom Panel = Adapted from BTT Study #2 (Peer et al., 2017)

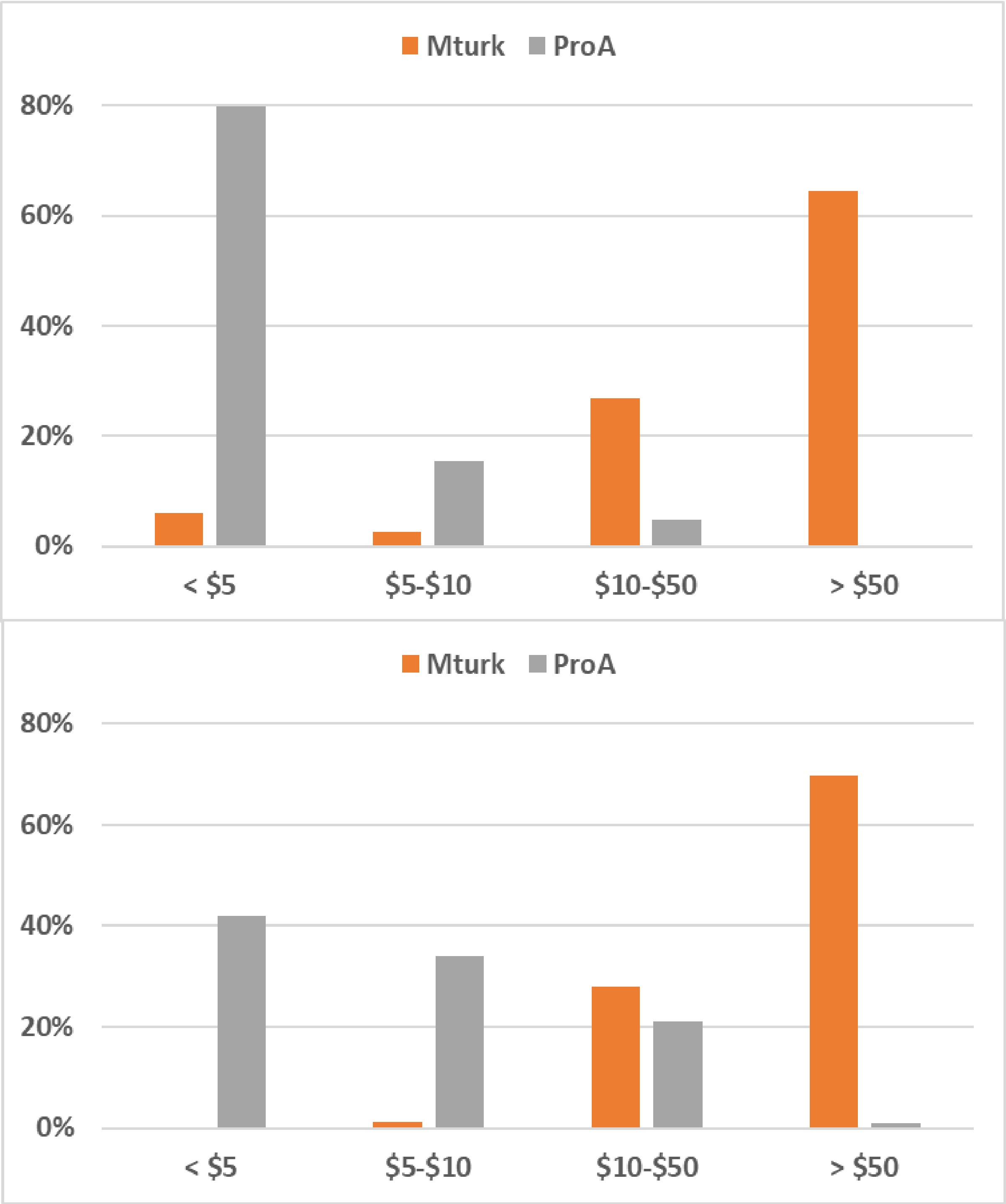

Figure 6. Reported Income - Quartile Percentages Of Online Crowdsourcing Platform Participants
Top Panel = This Study
Bottom Panel = Adapted from BTT Study #2 (Peer et al., 2017)

4 Discussion and Limitations

Our study largely replicates the results in the original BTT paper, and the complete comparison is presented in Table 4. Note that the original paper compared several other platforms, but we only include the three that we did replication on. In the original BTT study #1, the authors found that ProA participants reported higher naivety, lower degrees of dishonest behavior, lower frequencies of weekly participation, and only slightly lower levels of attention compared to MTurk. In our study, we found the same pattern for high naivety and lower frequencies of weekly participation. However, student participants' superiority on naivety became less obvious in our replication. MTurk presented high naivety compared to the original study, and the degree of the naivety of participants on ProA even outperformed that of our student participants. We suspect this to be related to the increasing size and/or the increasing update rate in the participant pool on MTurk and ProA, since the popularity of online survey platforms could have been having an increasing trend, such as MTurk and ProA (Bohannon, 2016; Palan & Schitter, 2018).

Table 4. Overall Comparison Between Platforms
Bold Text = This Study
Normal Text = Original BTT Study #1 (Peer et al., 2017)

	Mturk		ProA		Students	
Dropout rate	**Low**	Low	**Low**	Low	**Low**	Low
Response rate	**Fastest**	Fast	**Fast**	Fast	**Slowest**	Slowest
ACQs failure rate	**Low**	Lowest	**Lowest**	Low	**High**	Medium
Reliability	**High**	High	**High**	High	**High**	High
Reproducibility	**Good**	Good	**Good**	Good	**Good**	Fair
Naivety	**High**	Lowest	**Highest**	High	**High**	High
Dishonesty	**High**	Highest	**High**	Medium	**-**	-
Ethnic diversity	**Medium**	Low	**Low**	Low	**Low**	Medium
Geographic origin	**Mostly U.S.**	Mostly U.S.	**Mostly U.S.**	Mostly U.S.	**Mostly U.S.**	Mostly U.S.
English fluency	**High**	High	**High**	High	**High**	High
Income level	**High**	Low	**Medium**	Medium	**Low**	Low
Median education level	**High School**	Bachelor's	**Bachelor's**	Bachelor's	**Bachelor's**	Bachelor's
Usage frequency	**High**	High	**Medium**	Medium	**-**	Lowest
Overlap with other	**Some (ProA)**	Some (ProA)	**Some (Mturk)**	Some (Mturk)	**-**	Few

A few areas were found to deviate from the original study include the significance of dishonest behavior and levels of attention. Although we found a slightly lower degree of dishonest behavior on ProA, in this area the difference between ProA and MTurk was not found to be statistically significant, while in the original study, dishonesty is found to be significantly higher on MTurk. We found participants from ProA had the highest level of attention compared to the other participant groups. However, one potentially minor limitation for using ProA over MTurk is the apparent longer time required to collect sufficient quantities of data. In the opening hour of our survey, an error was found that stopped the bonus payment provided to ProA participants. The survey was taken offline and corrected within 30 minutes. Despite this error, on average, it took more than 20 minutes to collect 10 responses on ProA, while the time for MTurk was only about 5 minutes. One limitation in our work is that we recruited less than 200 participants on MTurk and ProA respectively. In our analysis, although on average ProA took a longer time to collect responses, the delay occurred close to the beginning and ending of the collection. ProA responses accelerated faster than MTurk for a time period in the middle of the collection. Therefore, in large-scale surveys with large sample sizes, the timeliness limitation of ProA might become less obvious, and this would remain to be explored for future work with larger sample size.

Consistent with the original study, we find that between the two online crowdsourcing platforms, MTurk and ProA, ProA consistently performed better in most areas, including successful completions, gender and ethnic diversity, participant internationality, attention, and naivety. MTurk and ProA performed comparably in areas including reliability, dishonest behavior, reproducibility, and response rate. Further, given the stark earnings and usage frequency distributions between Turkers and ProA participants, of the two platforms, ProA appears preferable. Given that researchers would often prefer a more attentive, naïve, honest and diversified group of participants, the advantages mentioned above in ProA's population make it a preferable platform from which to conduct research, despite MTurk's speed advantage. As an opposing viewpoint, however, some studies indicate that recruiting from a population that tends to use survey participation as a core means of revenue generation is problematic and may introduce bias that invalidates any finding (Chandler, Mueller, & Paolacci, 2014; Mason & Suri, 2012). Therefore, our findings indicate that ProA can be a much better alternative in general over MTurk, especially when necessity requires no strict constraint on timeliness.

Similar patterns occur in our replication paper between online platforms and the traditional research group. We found that some online platforms perform better in terms of attention, reproducibility and even naivety. Conclusions made from our findings should include consideration for the same limitations indicated by the original BTT paper, to include potential allocation bias (Peer et al., 2017).

Comparing results between the student groups is also an item of curiosity. Knowing that these two student groups answered their questionnaires at somewhat to significantly different contemporary climate or cultural environments, mixes of locale, demographics, and diversity may help describe the disparity found between them in almost every measurement.

5 Conclusion

Our findings reinforce the findings found in the original BTT study. Our study helps to validate the impression that online crowdsourcing platforms may be suitable and often a superior alternative to traditional research pools in many of the areas that behavioral researchers hope to see in their subjects, including successful completions, attention, and naivety. This study and the several that have come before (Bentley et al., 2017; Casler et al., 2013; Peer et al., 2017) indicate that both online crowdsourcing platforms continue to provide useful sources of fast, inexpensive, and relatively reliable research subjects. This was especially true for ProA, which appears to have marginally increased its performance over MTurk in the two years since the original BTT study.

Acknowledgements

We would like to thank AIS-TRR reviewers for their helpful assistance in revisions to our manuscript, and Dr. Susan Brown for encouraging us to get started on our journey.

Appendix A: Additional Figures

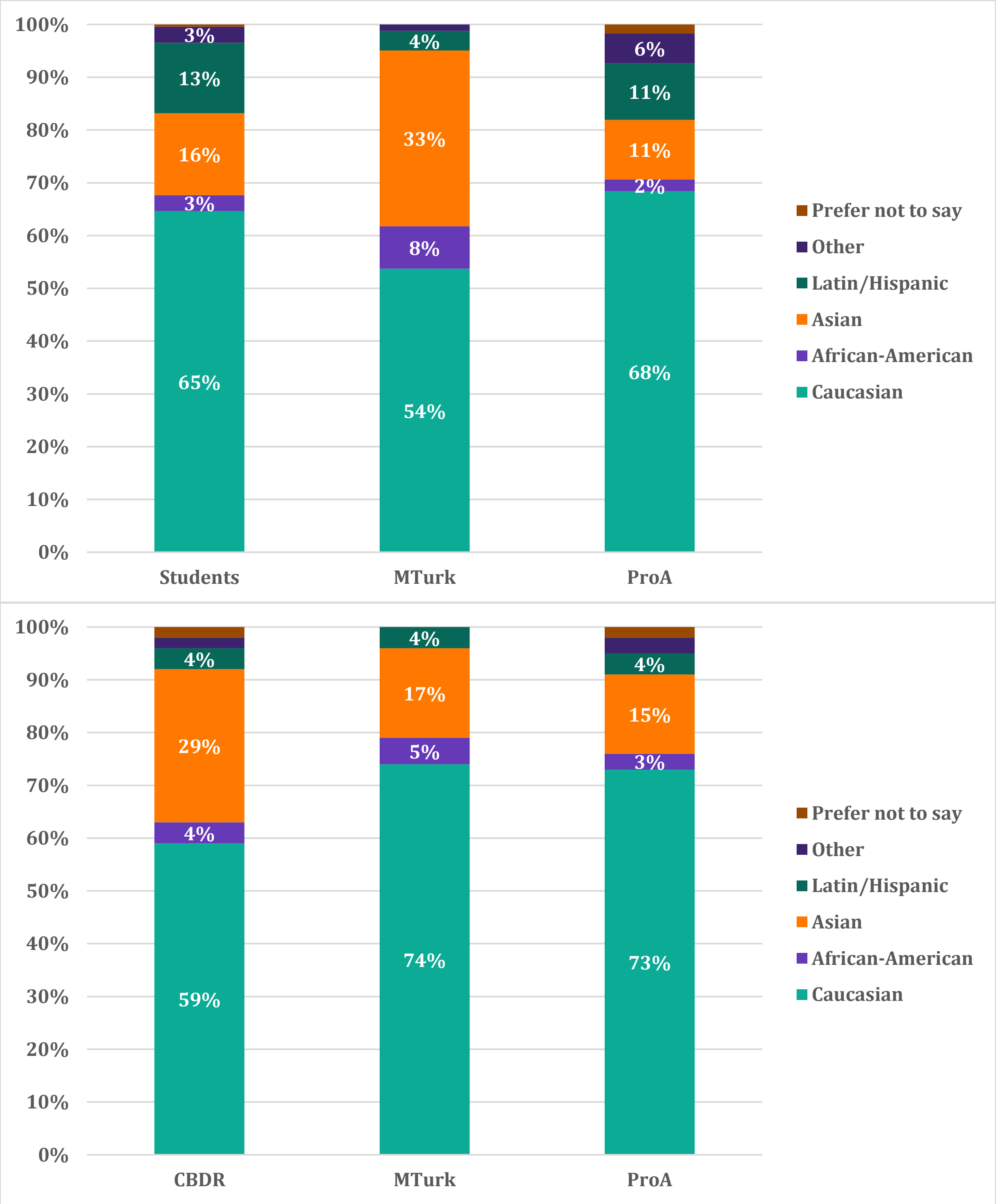

Figure A1. Reported Ethnicity Distributions
Top Panel = This Study
Bottom Panel = Adapted from Original BTT Study #1 (Peer et al., 2017)

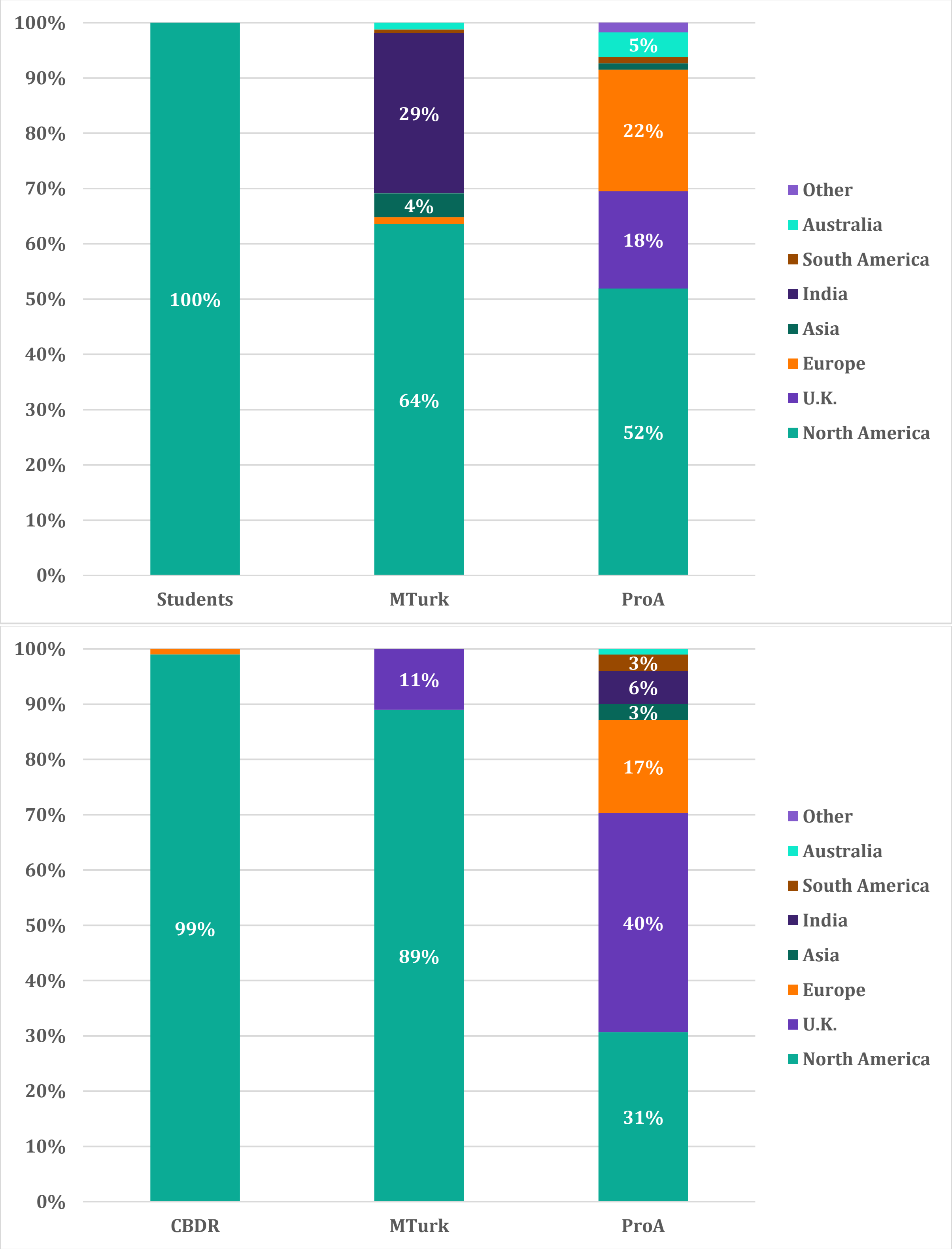

Figure A2. Reported Location Distributions
Top Panel = This Study
Bottom Panel = Adapted from Original BTT Study #1 (Peer et al., 2017)

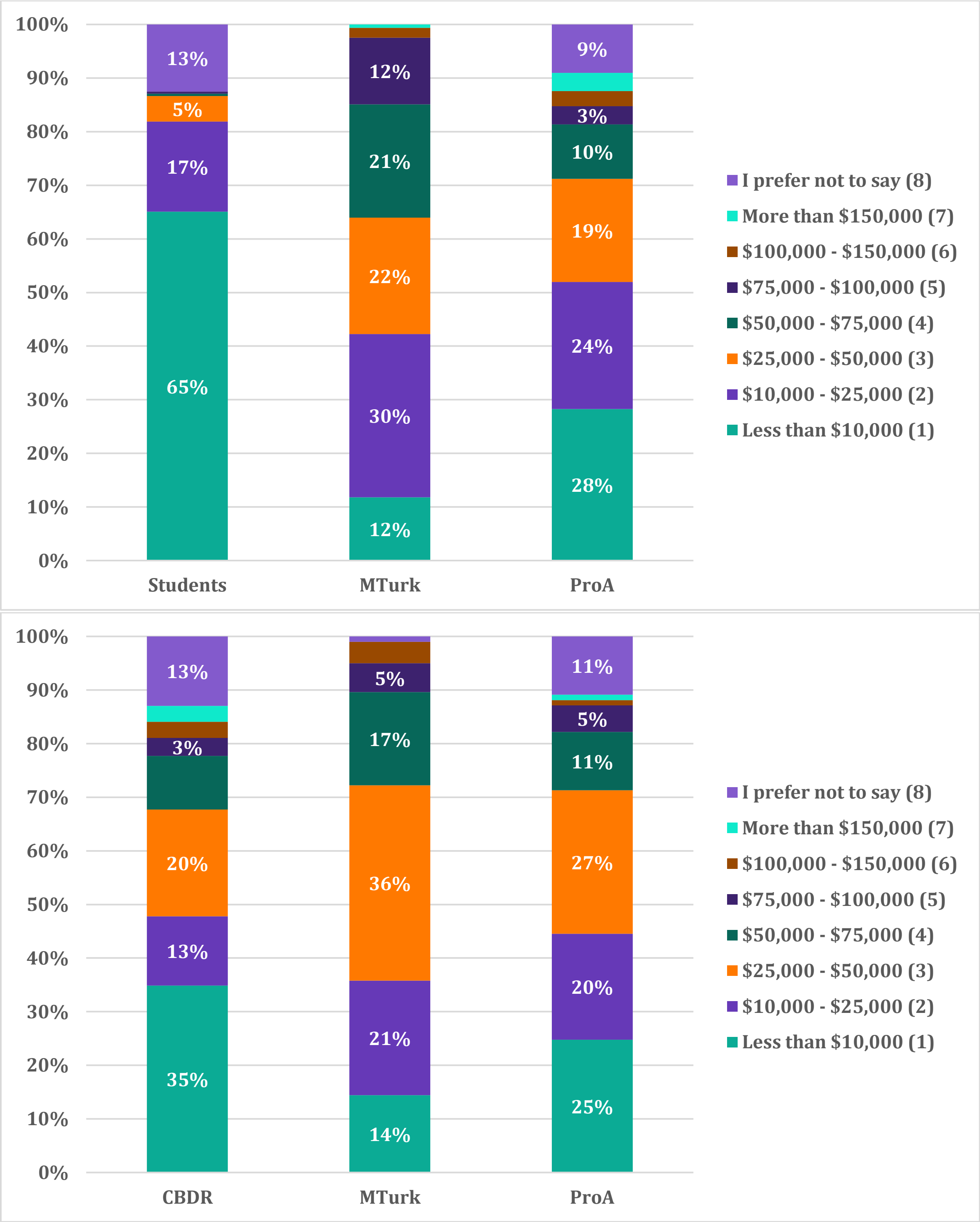

Figure A3. Reported Income Distributions
Top Panel = This Study
Bottom Panel = Adapted from Original BTT Study #1 (Peer et al., 2017)

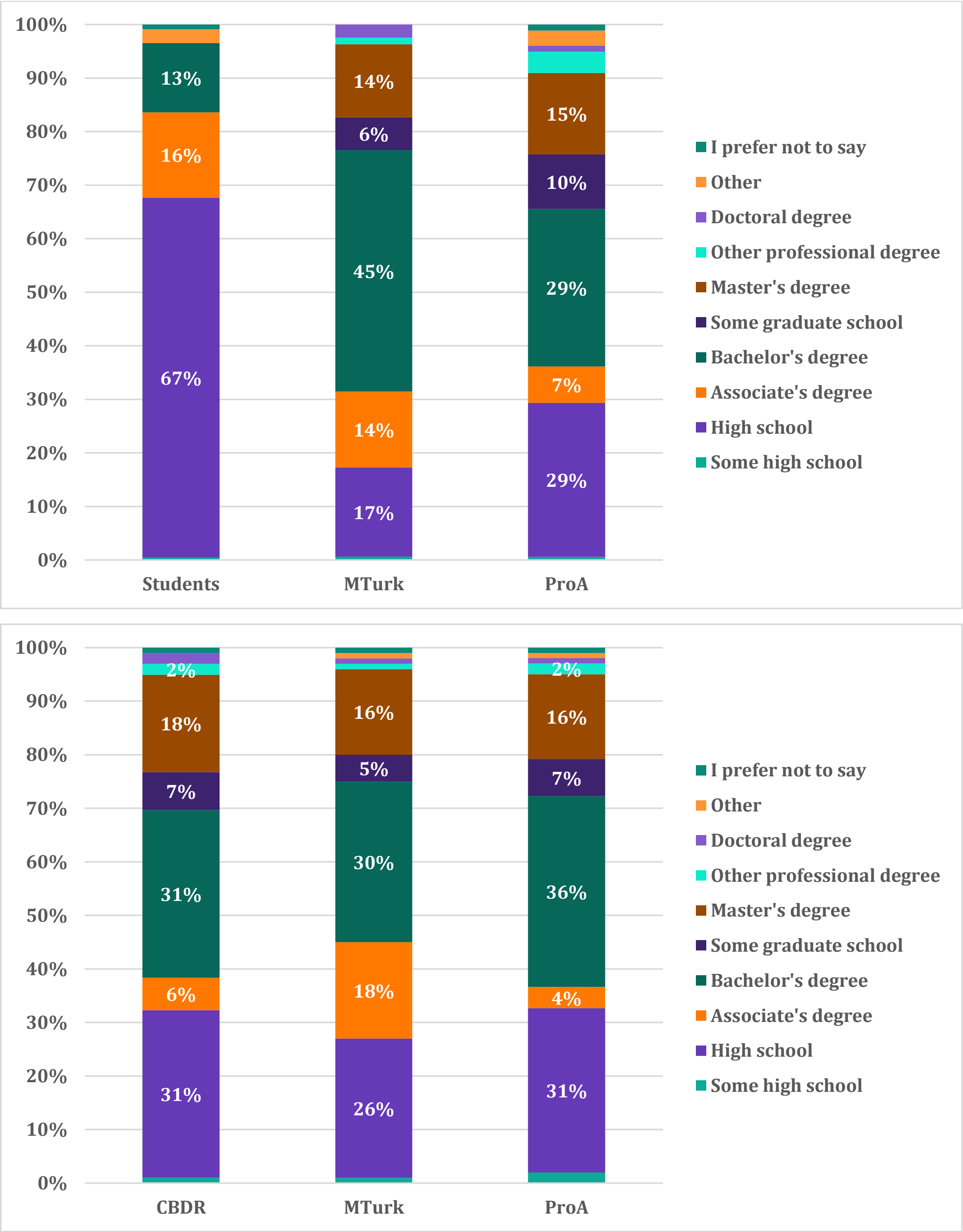

Figure A4. Reported Education Level Distributions
Top Panel = This Study
Bottom Panel = Adapted from Original BTT Study #1 (Peer et al., 2017)

Appendix B: Additional Tables

Table B1. Comparison Of Average Time To Collect 10 Responses
Bold Text = This Study, Normal Text = Original BTT Study #1 (Peer et al., 2017)

MTurk (5.00 minutes)	MTurk (5.62 minutes)
ProA (23.56 minutes)	ProA (12.94 minutes)
Students (around 7 hours)	CBDR (around 9 hours)

Table B2. Comparison Of Average Number Of Failed ACQs
Bold Text = This Study, Normal Text = Original BTT Study #1 (Peer et al., 2017)

MTurk (mean = 1.53, SD = 1.54)	MTurk (mean = 0.67, SD = 0.96)
ProA (mean = 1.22, SD = 1.38)	ProA (mean = 0.81, SD = 1.01)
Students (mean = 1.74, SD = 1.33)	CBDR (mean = 1.04, SD = 1.14)

T-test for number of failed ACQs between the BTT original paper and our replication shows there is significant difference for MTurk, t(363) = 7.071, p < 0.01. Significant difference also exists for ProA, t(391) = 16.207, p < 0.01. Difference between student groups is not significant, t(427) = 3.543, p = 0.270.

Table B3. Comparison Of Bonus Over Reporting
Bold Text = This Study, Normal Text = Original BTT Study #1 (Peer et al., 2017)

MTurk (mean = 44.57, SD = 14.54)	MTurk (mean = 46.87, SD = 12.67)
ProA (mean = 42.60, SD = 15.85)	ProA (mean = 42.29, SD = 15.80)

References

Amazon Mechanical Turk. (2018). About Amazon Mechanical Turk. Retrieved from https://www.MTurk.com/worker/help.

Bentley, F. R., Daskalova, N., & White, B. (2017). Comparing the reliability of Amazon Mechanical Turk and Survey Monkey to traditional market research surveys. Paper presented at the *2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*.

Bohannon, J. (2016). Mechanical Turk upends social sciences. *Science*, 352(6291), 1263-1264.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48(3), 306-307.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6), 2156-2160.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. Behavior Research Methods, 46(1), 112-130.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological Science*, 26(7), 1131-1139.

Difallah, D., Filatova, E., & Ipeirotis, P. (2018). Demographics and dynamics of mechanical Turk workers. Paper presented at the *Eleventh ACM International Conference on Web Search and Data Mining*.

Lutz, J. (2015). The validity of crowdsourcing data in studying anger and aggressive behavior. *Social Psychology*, 47(1), 38–51.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1), 1-23.

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.

Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, 4(5), 326.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22-27.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23(3), 184-188.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153-163.

Prolific Academic. (2018). What is Prolific Academic? Retrieved from https://support.prolific.ac/article/42-what-is-prolific-academic.

Prpić, J., Taeihagh, A., & Melton, J. (2015). The fundamentals of policy crowdsourcing. *Policy & Internet*, 7(3), 340-361.

Rosenberg, M. (2006). Rosenberg self-esteem scale (RSE). In Ciarrochi, J. & Bilich, L., *Acceptance and Commitment Therapy, Measures Package: Process Measures of Potential Relevance to ACT*, (pp. 61).

Smith, S. M., Roster, C. A., Golden, L. L., & Albaum, G. S. (2016). A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research*, 69(8), 3139-3148.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481), 453-458.

About the Authors

Troy L Adams is an Instructor at Arizona State University's Ira A Fulton Colleges of Engineering. He holds master's degrees in Management Information Systems and Cybersecurity from the University of Arizona's Eller College of Management and is a graduate of the AZSecure SFS Cybersecurity Fellowship program. He is a federal employee of the US Department of Health and Human Services, as a Cyber Engagement Specialist at the Health Sector Cybersecurity Coordination Center (HC3), and President of the (ISC)^2 Southern Arizona Chapter.

Yuanxia Li is a PhD student from Eller College of Management, University of Arizona. She joined the Department of Management Information Systems in 2017. Her research interests lie in conceptual modeling, social media analytics, and business intelligence.

Hao Liu is a PhD candidate in the Department of Management Information Systems at the Eller College of Management at University of Arizona. He received his master's degree in Computer Science from University of North Carolina at Charlotte in 2017 and Bachelor's degree in Computer Science and Technology from University of Science and Technology of China in 2013. His research interests include data mining, machine learning, deep learning, and their applications in recommender systems, professional talent analytics, and healthcare informatics.