TPM®

PAYING PARTICIPANTS: THE IMPACT OF COMPENSATION ON DATA QUALITY

KIMBERLY R. MORE
UNIVERSITY OF DUNDEE

KAYLA A. BURD
UNIVERSITY OF WYOMING

CURT MORE
UNIVERSITY OF DUNDEE

L. ALISON PHILLIPS
IOWA STATE UNIVERSITY

Poor-quality data has the potential to increase error variance, reduce statistical power and effect sizes, and produce Type I or Type II errors. Paying participants is one technique researchers may use in an attempt to obtain high-quality data. Accordingly, two secondary datasets were used to examine the relationship between participant payment and data quality. The first dataset revealed that data quality did not differ between paid and unpaid undergraduates. Similarly, the second dataset showed that data quality did not differ between unpaid community participants and MTurkers. A comparison across studies showed that undergraduate students engaged in lower levels of careless responding than the community samples but the unpaid community sample outperformed the MTurk sample and both undergraduate samples. Taken together, the current findings suggest that offering financial incentives to undergraduate or community samples does not improve data quality but may improve data collection rates and increase the diversity of participants.

Keywords: Participant compensation; Data quality; Surveys; Research methods; Data collection.

Correspondence concerning this article should be addressed to Kimberly More, Department of Psychology, University of Dundee, Nethergate, Dundee, DD1 4HN, Scotland. Email: kmore001@dundee.ac.uk

Psychological research is largely dependent upon undergraduate participant pools that offer course credit as compensation (Rocchi et al., 2016). This form of data collection is inexpensive and, within larger universities, results in relatively quick data collection (Kees et al., 2017). Unfortunately, undergraduate participants have been criticized for being inattentive (Hauser & Schwarz, 2015) and responding carelessly (Maniaci & Rogge, 2014). Consequently, this response style produces lower-quality data with increased error variance as well as reduced statistical power and effect sizes in both observational and experimental studies. Research has shown that even low levels of random responding can greatly increase Types I and II error rates (Credé, 2010). Additionally, the size of participant pools may hinder data collection efforts at smaller universities (Crump et al., 2013).

Recently, researchers have begun utilizing online data collection platforms, such as Mechanical Turk (MTurk), as a relatively expeditious means of collecting data. However, the quality of data obtained from such sources has been called into question (Chmielewski & Kucker, 2019). Therefore, the purpose of the present research was (1) to examine data quality among undergraduate participants who were financially compensated in addition to receiving course credit compared to those who were not, (2) to compare indices of data quality between subjects drawn from the general population who were either paid (i.e., MTurk sample) or unpaid (i.e., community sample obtained through snowball sampling), and (3) to compare the proportion of correct memory questionnaire scores and longstring responding across samples to determine the sample type and method that produces the highest index of data quality.

To date, research has yet to examine the potential impact of receiving monetary payment in addition to extra credit versus extra credit alone on data quality among undergraduate student samples. Thus, it is unclear whether financial compensation would enhance motivation to perform well on research-related tasks among student participants. Prior research suggests that monetary incentives increase participation rates (e.g., Church, 1993; Singer & Ye, 2013), and may improve item nonresponse (Porter & Whitcomb, 2003). Further, research examining the use of extra credit to incentivize research participation has explored the impact of such incentives on meeting educational goals (i.e., exposing students to research practices). Findings indicated that extra credit alone may not be sufficient to induce participation in research to levels that meet educational goals, and may limit the generalizability of research findings, as students incentivized with extra credit to participate in research differed systematically from students who were not motivated by extra credit (Padilla-Walker et al., 2005). However, no known research to date has compared data quality among participants who received extra credit versus extra credit in addition to a monetary incentive. Rather, research on data quality has largely focused on the comparison between undergraduate samples and samples drawn from online platforms (e.g., MTurk) or on the comparison between community samples and samples drawn from online platforms.

Research has shown that undergraduate participants drawn from research pools are more careless in their responding in comparison to participants from MTurk. Specifically, MTurkers were significantly more likely to pass a manipulation check (25.5-96%) compared to undergraduate students (2.2-26%) across three studies (Hauser & Schwarz, 2015). Such findings are robust and highlight the inattentiveness of participants from undergraduate research pools. Further, Kees and colleagues (2017) found that data quality from an MTurk sample either matched or outperformed undergraduate samples collected online or in person. These results were replicated using data from the Many Labs 3 project, which collected participant data from 20 universities (*N* = 2,696) and MTurk (*N* = 737). A large effect size found between groups suggests that MTurk participants paid more attention during the surveys than did undergraduate students (Capaldi, 2017).

MTurk samples have also been shown to provide similar or higher quality data than other general population samples (Kees et al., 2017). For instance, Necka and colleagues (2016) found that respondent behaviors were similar across MTurkers and community participants collected through email snowball sampling. Additionally, research has shown MTurk samples outperform participants recruited through Qualtrics or Lightspeed on many metrics, including reliability of measures, manipulation checks, and attention checks (Kees et al., 2017). Moreover, data from experimental samples collected through MTurk have replicated results from more traditional means of data collection in experimental economics as well as in cognitive, clinical, and social psychology (e.g., Amir et al., 2012; Crump et al., 2013; Sharpiro et al., 2013; Summerville & Chartier, 2013). Specifically, meta-analytic estimates have shown that results obtained from MTurk correlate with results obtained from population-based probability samples (i.e., *r* = .75-.83; Coppock, 2019; Mullinix et al., 2015). In addition, data produced by MTurkers has relatively high indices of internal, convergent, and test-retest reliability (Sharpiro et al., 2013). Taken together, these results suggest that MTurkers are not only more attentive when participating in research activities, but that MTurk represents an efficient means of collecting representative data with relative ease, which is lacking in undergraduate and other community samples. However, the payment associated with recruiting participants via MTurk and other paid participant platforms, such as Qualtrics or Lightspeed, has been shown to lead to fraudulent responses to gain access to a survey with specific inclusion criteria, especially when the reward is high.

The purpose of this study was two-fold. First, based on previous findings, the present study was conducted to evaluate the potential effects of participant payment on metrics of data quality. Specifically, the purpose was to examine data quality among undergraduate participants receiving extra credit who were paid or unpaid and the general population who were either paid or unpaid. Second, the present study examined the potential effects of participant payment on memory questionnaire scores concerning important study-specific content. Although indices of memory are distinct from both direct and statistical measures of careless responding (Goldammer et al., 2020), assessing participants' knowledge of study-specific content is important for understanding which participants were able to retain knowledge of the information presented to them. These scores may be influenced by carelessness (e.g., not paying attention to or skipping over the intervention content).

Two independent samples (Study 1 = undergraduate participants enrolled in a self-affirmation study; Study 2 = general population participants enrolled in a jury decision making study), upon which secondary analyses were conducted, were utilized to examine the impact of financial compensation on data quality. In both studies, payment was only offered to part of the sample. To this end, the present research makes use of two existing datasets that allowed for the opportunity to evaluate the relationship between participant payment and data quality.

STUDY 1

Method

Participants and Procedure

Women ($N = 310$) were recruited from an undergraduate research participant pool at a Midwestern University in the United States to participate in an experiment regarding self-affirmation and exercise intentions (More et al., 2022). Participants were eligible to enroll in the study for two course credits if they did not regularly engage in moderate or vigorous exercise. For the first semester of data collection, the study description also highlighted that participants would be eligible to earn five U.S. dollars if they scored at least 80% on a memory questionnaire after reading a brief message ($n = 160$; 51.6%). However, participant payment led to cheating. Specifically, several students ($n = 5$) were caught cheating on the memory questionnaire by the research assistant. Therefore, during the second semester of data collection participants earned two course credits but were not paid based on memory scores ($n = 150$; 48.4%). Participants who cheated were not included in the aforementioned sample size or in any analyses.

Participants were randomized at baseline to either an intervention (i.e., self-affirmation) or control group. The self-affirmation group wrote at least 250 characters about how a value of personal importance to them guided their behavior. The control group wrote at least 250 characters about how a value that was not of personal importance to them may be important to the average college student. Both groups read the same message concerning the consequences of physical inactivity and responded to an 11-item questionnaire assessing how well they remembered the message. In addition to the memory questionnaire, random response checks and whether participants followed the writing task instructions were used to evaluate data quality.

Measures of Data Quality

Whether participants lied to gain access to the survey (Chandler & Paolacci, 2017) was assessed by evaluating whether participants met the inclusion criteria of not regularly engaging in moderate or vigorous exercise. Engagement was pre-screened using the Stage of Change Measure (Prochaska & DiClemente, 1982). Participants were *not* eligible for participation if they were in either the action (regular exercise engagement for less than six months) or maintenance (regular exercise engagement for more than six months) phases of behavior change.

Participants responded to the 44-item Big Five Inventory (John & Srivastava, 1999) using a 5-point Likert scale ranging from *strongly disagree* to *strongly agree* to assess the longest string of any given response category for each participant. The subscales for the Big Five Inventory were interspersed among the scale.

Two random response checks were embedded within the survey, which prompted participants to select a particular response to indicate that they were reading and attending to the task (i.e., "If you are reading this, answer 2," "If you are reading this, answer agree a little"). Such random responding items with only one correct response have been used in research (Buchanan & Scofield, 2018) and are recommended as good practice given the problems associated with even a small amount of random responding (Credé, 2010).

The intervention group was asked to write about how their self-reported most important value has guided their behavior in the past. The control group was asked to write about how their self-reported least important value could be relevant to the average college student. Adherence was assessed by two independent coders.

Memory Measure

The memory questionnaire was comprised of 11 questions that directly assessed knowledge of the message content received by participants in both groups. Items directly pertained to the message content delivered (e.g., "Which type of cancer was discussed in the message that you just read?" and "Is physical activity associated with Type 1 or Type 2 diabetes?"). Participants filled in the blank for each item using pencil and paper. Higher scores were indicative of greater memory of the message content. Eight participants had 100% missing data on the memory questionnaire and were excluded from analyses using the memory questionnaire.

Results

Participants were 19.20 years of age on average (SD = 1.49) with the majority self-reporting as European American (71.6%). Other reported ethnicities included Asian (9.4%), Latin American (7.7%), African American (5.5%), Middle Eastern (4.5%), and Native American (1.3%). The paid and unpaid groups did not significantly differ in age, $t(306) = –1.40$, $p = .164$ (Levene's test for equality of variances, $F = 0.01$, $p = .925$), or ethnicity, $\chi^2(5, n = 310) = 5.81$, $p = .325$.·

Categorical outcomes (i.e., ethnicity, lying to gain access to the study, random responding, adherence to the writing task) were analyzed using contingency tables with chi-square values. Longstring screening was used on the longest measure presented to participants (i.e., the Big Five Inventory), which is composed of 44 items. The first 23 items were presented to participants together, followed by a page break, after which the remaining 21 items were presented. As recommended by Johnson (2005), longest strings of the same response category were calculated using syntax provided by Johnson and Mayer (2019). There were no missing data on the Big Five Inventory. Syntax was adapted to reflect a 44-item scale, a 23-item scale, and a 21-item scale, respectively. A 5-point Likert scale was specified. Full instructions for use, including annotated syntax, are provided by Johnson and Mayer (2019). Continuous outcomes (i.e., age, longstring responding, and memory questionnaire score) were analyzed using independent samples *t*-tests. In both cases, outcomes were compared across those who had the opportunity to receive payment and those who did not.

Adherence to the writing prompt within groups was assessed on two dimensions by two independent coders. First, it was determined whether participants in the control group and participants in the intervention group actually wrote about their lowest and highest ranked values, respectively. This was dichotomously coded as 0 = *failed to write about appropriate value* and 1 = *succeeded in writing about appropriate value*. Second, it was determined whether participants in the control group only wrote about an average college student (i.e., did not write about themselves) and whether participants in the intervention condition actually wrote about themselves. This was dichotomously coded as 0 = *failed to write about the appropriate subject* and 1 = *succeeded in writing about the appropriate subject*. A holistic adherence-to-writing-task score was computed where 0 = *failing one or both dimensions* and 1 = *succeeding on both dimensions*. The coders agreed on 95.87% of cases with the remaining cases resolved through discussion. An accuracy score, which ranged from 0 to 1, was calculated for the memory questionnaire by dividing the total number of correct memory check items by the total number of memory check questions (i.e., 11).

Measures of Data Quality

Whether payment influenced lying to gain access to the survey was assessed. Ineligible participants (n = 66) were not included in the remainder of analyses. Across groups that were offered payment and those that were not, there was no significant difference in the proportion of participants that lied to gain access to the study, $\chi^2(1, n = 376) = 0.47$, $p = .493$.

Whether payment influenced random responding was analyzed in five ways. First, a trichotomous outcome of missing both response checks, missing one response check, or missing no response checks was assessed. Second, a dichotomous outcome of either passing both response checks or not was assessed. Across groups that were offered payment and those that were not, there was no significant difference in the proportion of participants randomly responding for the trichotomous outcome variable, $\chi^2(2, n = 310) = 1.41$, $p = .494$, or for the dichotomous outcome variable, $\chi^2(1, n = 310) = 0.22$, $p = .639$. Third, longstring was compared for the total Big Five Inventory across those who had the opportunity to receive payment ($M = 4.21$, $SD = 4.65$) and those who did not ($M = 4.07$, $SD = 3.58$). Longstrings ranged from 2 to 44. Levene’s test for equality of variances was met, $F = 0.19$, $p = .661$, and there was no significant difference between the two groups, $t(308) = –0.29$, $p = .769$. Fourth, longstring was compared across groups for the first 23 items of the Big Five Inventory, which were presented to participants on a single survey page. Longstrings ranged from 1 to 23. Similar to the total scale, Levene’s test for equality of variances was met, $F = 0.01$, $p = .924$, and those who had the opportunity to receive payment ($M = 3.23$, $SD = 2.44$) did not differ from those who did not have the opportunity to receive payment ($M = 3.10$, $SD = 2.04$), $t(308) = –0.49$, $p = .626$. Finally, longstring was compared for the final 21 items of the Big Five Inventory, which were presented to participants on a single survey page. Longstrings ranged from 2 to 21. Levene’s test for equality of variances was met, $F = 0.16$, $p = .691$, and those who had the opportunity to receive payment ($M = 3.51$, $SD = 2.23$) did not differ from those who did not ($M = 3.47$, $SD = 1.99$), $t(308) = –0.13$, $p = .894$.

The impact of payment on adherence to the writing task was then analyzed. Results indicated that there was no statistical difference in the proportion of individuals who adhered to the task instructions between individuals who were offered payment and those who were not, $\chi^2(1, n = 310) = 2.26$, $p = .133$.

Memory Measure

Finally, whether participants who were offered payment scored significantly different on the memory questionnaire in comparison with participants who were not offered payment was analyzed. As indicated above, all participants who were found to have cheated, as well as participants who did not complete any items on the memory questionnaire, were removed from the sample and subsequent analyses. Levene's test for equality of variances was met, $F = 0.61$, $p = .434$, with the group who was offered payment ($M = 0.74$, $SD = 0.19$) scoring significantly better on the memory questionnaire in comparison with the group who was not offered payment ($M = 0.68$, $SD = 0.21$), $t(300) = –2.22$, $p = .027$, with a small effect size, $d = 0.25$.

Discussion

Indices of data quality and memory were compared between two groups of undergraduate students. The first group was offered payment for participation, whereas the second group was not. Analyses indicated that the opportunity to receive payment did not alter indices of data quality (i.e., lying to gain access to the study, longstring, random responding and adherence to writing task instructions). However, the opportunity to receive payment did influence memory questionnaire scores. Specifically, the group that had the opportunity to receive financial compensation scored significantly higher on the memory questionnaire in comparison with the group that did not. It should be noted that this difference was small and corresponded to only a 6% improvement. Importantly, both groups still scored below 80% on the memory questionnaire on average, which was the requirement of payment (i.e., a score of nine or higher), and the difference between groups was small in accordance with Cohen's *d*.

STUDY 2

Method

Participants and Procedure

Study 2 was comprised of 291 participants, which included 157 (53.95%) participants recruited via email, social media, and word of mouth using snowball sampling. These participants recruited via snowball sampling volunteered their time and were not compensated. In addition, 134 (46.05%) participants were recruited via MTurk and TurkPrime (Litman et al., 2016). Individuals in the MTurk sample who completed the full survey were compensated with two U.S. dollars for their participation.

Participation was completed entirely online. Participants were eligible if they were 18 years of age or older, U.S. citizens, and living in the United States. All materials and measures were administered online utilizing Qualtrics. Survey links were disseminated to the general population sample via email, social media, and Amazon's MTurk via the TurkPrime platform. Participants acted as mock jurors and were tasked with reading a brief summary of a civil trial along with judicial instructions. Participants then responded to three memory attention checks followed by an emotion questionnaire and case-related dependent measures. The memory attention check was composed of three items, which directly pertained to the content of the trial summary. Items were answered using a multiple-choice response format with four response options each.

Measure of Data Quality

Participants responded to 17 selected items from the Positive and Negative Affect Schedule-X (PANAS-X; Watson & Clark, 1994) and two additional items, Anxious and Contempt, using a 5-point Likert scale ranging from *strongly disagree* to *strongly agree* to assess the longest string of any given response category for each participant. The subscales of negative and positive affect were interspersed among the scale.

Memory Measure

Memory was assessed using a memory questionnaire containing three items that tested how well participants remembered key pieces of the mock trial evidence. For each of the three memory questionnaire items participants received a score of 0 for incorrect answers and a score of 1 for correct answers. An accuracy score ranging from 0 to 1 was then calculated by dividing the total number of correct memory check items by the total number of memory check questions. Higher scores were indicative of greater memory for case-related details. Four hundred and sixteen participants began the survey (general sample: *n* = 239; MTurk sample: *n* = 177). Analyses were conducted on those participants who completed the survey in its entirety, including the memory questionnaire (general sample: *n* = 157; MTurk sample: *n* = 134; *N* = 291).

Results

Participants across the sub-samples were dissimilar in terms of age, gender, education, and race/ethnicity (Table 1). Similar to Study 1, categorical outcomes (i.e., demographics) were analyzed using contingency tables with chi-square values. Continuous outcomes (i.e., age, longstring, and the memory questionnaire) were analyzed using independent samples *t*-tests. In both cases, outcomes were compared across those who had the opportunity to receive payment (i.e., MTurk sample) and those who did not (i.e., community sample). Longstring screening was used on the longest measure presented to participants (i.e., the emotion items from the PANAS-X, plus two additional emotion items) for a total of 19 items. There were no missing data on the PANAS-X. Syntax was adapted to reflect a 19-item scale. A 5-point Likert scale was specified.

Measure of Data Quality

Longstring was compared for the 19 emotion items, which were presented on a single survey page, and ranged from 1 to 19. Levene's test for equality of variances was not met, F = 5.72, p = .017, for an independent samples *t*-test. Therefore, a Welch ANOVA was performed. Results showed that the unpaid community (M = 4.94, SD = 2.97) and MTurk (M = 4.85, SD = 3.71) samples did not differ, $F(1, 253.72) = 0.05$, p = .832.

TABLE 1
Characteristics of Study 2
Final sample

Variable	General sample (*n* = 157)		MTurk sample (*n* = 134)		*p*
	Range / *n*	Mean / %	Range / *n*	Mean / %	
Age	19-71	38.99	21-66	33.02	< .001
Race/Ethnicity					< .001
African American	6	3.8%	32	24.1%	
Hispanic	17	10.8%	10	7.5%	
Asian/Pacific Islander	2	1.3%	10	7.5%	
Native American	1	0.6%	2	1.5%	
European American	125	79.6%	76	57.1%	
Other	6	3.8%	3	2.3%	
Gender identification					< .001
Cis-male	34	22.1%	70	52.6%	
Cis-female	103	66.9%	46	34.6%	
Trans-male	0	0.0%	5	3.8%	
Trans-female	1	0.6%	0	0.0%	
Nonbinary	2	1.3%	5	3.8%	
Other	14	9.1%	7	5.3%	
Education					.013
< High school	0	0.0%	1	0.7%	
High school/GED	9	5.7%	10	7.5%	
Some college	44	28.0%	33	24.6%	
College graduate	40	25.5%	57	42.5%	
Some graduate school	21	13.4%	8	6.0%	
Graduate degree	43	27.4%	25	18.7%	

Note. Please note that there is one missing value within the MTurk sample on both gender identification and race/ethnicity and that the general sample has three missing values on gender identification and one missing value on age.

Memory Measure

Memory was assessed by using an independent samples *t*-test to examine whether the unpaid community sample and the MTurk sample differed regarding accuracy on the memory questionnaire. A Levene's test for equality of variances was not met, $F = 65.19$, $p < .001$, for an independent samples *t*-test. Therefore, a Welch ANOVA was performed. The unpaid community sample (M = 0.90, SD = 0.19) was significantly more accurate on the memory questionnaire compared to the paid MTurk sample (M = 0.74, SD = 0.33), $F(1, 203.20) = 24.57$, $p < .001$, with a medium-to-large effect size, $\omega^2 = .08$.

Exploratory Analysis

An exploratory analysis was conducted to examine whether demographic differences between the samples contributed to the differential accuracy on the memory questionnaire. As noted above, the two subsamples were dissimilar regarding age, gender, education, and ethnicity. To examine whether the above factors influenced memory questionnaire accuracy, a multiple regression was performed. Gender, education, and ethnicity were dummy coded. As the sample was primarily comprised of white, cis-women, unpaid community members with a college degree, these groups served as the reference group for the following analysis (see Field, 2009). Age, gender, education, and race/ethnicity were entered at Step 1 (Adjusted R^2 = .20), whereas sample (unpaid community vs. paid MTurk) was entered at Step 2 (adjusted R^2 = .22). The R^2 change between models was equivalent to approximately 2% and was significant, *p* = .007.

Sample (unpaid community vs. paid MTurk) remained a significant predictor of memory questionnaire accuracy even after controlling for group differences. Specifically, the unpaid community sample still significantly outperformed the paid MTurk sample, *b* = .09, *SE* = .03, *p* = .007 (Table 2).

TABLE 2
Exploratory regression

	b	*SE*	*B*	*t*	*p*
Model 1					
Constant	.78	.06		13.68	< .001
Age	.003	.001	.14	2.35	.020
Reference group white					
Black	–.25	.05	–.30	–5.40	< .001
Hispanic	–.07	.05	–.07	–1.26	.210
Native American	–.27	.15	–.10	–1.86	.064
Asian/Pacific Islander	–.07	.07	–.05	–0.92	.358
Other	–.02	.09	–.01	–0.26	.793
Reference group cis-female					
Cis-male	–.001	.03	–.002	–0.03	.977
Other	–.08	.06	–.08	–1.37	.172
Trans male	–.39	.11	–.19	–3.42	.001
Trans woman	.06	.25	.01	0.23	.821
Nonbinary	–.30	.10	–.17	–3.08	.002
Reference college degree					
High school degree/GED	–.05	.06	–.05	–0.79	.430
Some college	.06	.04	.10	1.55	.121
< four years of high school	.14	.25	.03	0.55	.581
Some graduate school	.07	.06	.08	1.26	.207
Graduate school	–.03	.04	–.05	–0.81	.419
Model 2					
Constant	.74	.06		12.76	< .001
Age	.002	.001	.12	1.97	.049

(table 2 continues)

Table 2 (continued)

	b	*SE*	*B*	*t*	*p*
Reference group white					
Black	–.22	.05	–.27	–4.65	< .001
Hispanic	–.08	.05	–.09	–1.53	.127
Native American	–.25	.14	–.09	–1.72	.086
Asian/Pacific Islander	–.04	.07	–.03	–0.50	.619
Other	–.02	.09	–.02	–0.28	.778
Reference group cis-female					
Cis-male	.02	.03	.04	0.71	.481
Other	–.08	.06	–.08	–1.44	.152
Trans male	–.33	.11	–.16	–2.87	.004
Trans woman	.03	.25	.01	0.12	.903
Nonbinary	–.28	.10	–.16	–2.84	.005
Reference group college degree					
High school degree/GED	–.04	.06	–.04	–0.68	.501
Some college	.05	.04	.08	1.32	.187
< four years of high school	.16	.25	.04	0.67	.503
Some graduate school	.04	.06	.05	0.75	.454
Graduate school	–.05	.04	–.08	–1.25	.213
Sample (MTurk vs. unpaid community)	.09	.03	.17	2.70	.007

Note. *b* = unstandardized regression coefficient; *SE* = standard error; *B* = standardized regression coefficient.

Discussion

Two sub-samples participated in the described study, including unpaid participants recruited via snowball sampling and paid participants recruited via MTurk. Participants across both subsamples were demographically dissimilar and were differentially compensated. Analyses revealed that the unpaid community sample still outperformed the MTurk sample on a memory questionnaire even after controlling for key sample differences. However, it should be noted that this difference was small as it only accounted for 2% of the variance in memory questionnaire scores. Further, no differences were found regarding longstring responding among the paid and unpaid samples.

Comparison of Studies 1 and 2

Comparing the four groups on longstring responding was done in two ways: (1) with the first 21 items of the Big Five, all of which were displayed on one page to participants, and (2) with the final 23 items of the Big Five, all of which were displayed on one page to participants (Table 3). For the comparison of the community groups with the undergraduate groups using the first 21 items of the Big Five, Levene’s test for equality of variances was not met, $F = 20.20$, $p < .001$. Therefore, a one-way Welch ANOVA was used and revealed that the groups significantly differed on accuracy, $F(3, 315.81) = 21.16$, $p < .001$, with a small effect size, $\omega^2 = .03$ (Table 3). Specifically, both undergraduate groups had significantly shorter longstring responding patterns than both community groups (*ps* < .001), which indicates lower rates of careless responding. The community groups did not differ from one another, $p = .997$, and the undergraduate groups did not differ from one another, $p = .999$, using a Games-Howell post-hoc test.

TABLE 3
Proportion of correct responses on memory questionnaire and longstring across studies

Sample	*N*	*M*	*SD*	*p*
Longstring with 21 Big Five items (Study 1) and 19 emotion items (Study 2)				
Study 1 (Paid)	160	0.17	0.11	< .001
Study 1 (Unpaid)	150	0.17	0.09	
Study 2 (MTurk)	134	0.26	0.20	
Study 2 (Community)	157	0.26	0.16	
Longstring with 23 Big Five items (Study 1) and 19 emotion items (Study 2)				
Study 1 (Paid)	160	0.14	0.11	< .001
Study 1 (Unpaid)	150	0.13	0.09	
Study 2 (MTurk)	134	0.26	0.20	
Study 2 (Community)	157	0.26	0.16	
Memory questionnaire				
Study 1 (Paid)	160	0.74	0.19	< .001
Study 1 (Unpaid)	150	0.68	0.21	
Study 2 (MTurk)	134	0.74	0.33	
Study 2 (Community)	157	0.90	0.19	

Note. The proportion of correct responses was calculated by dividing the total number of correct responses to memory questionnaire or longstring items by the total number of items. Group differences were tested using a one-way Welch ANOVA.

These results were replicated using the final 23 items of the Big Five. Levene's test for equality of variances was not met, $F = 28.05$, $p < .001$. Therefore, a one-way Welch ANOVA was used and revealed that the groups significantly differed on accuracy, $F(3, 314.02) = 37.60$, $p < .001$, with a medium effect size, $\omega^2 = .06$ (Table 3). Compared to the community groups, both undergraduate groups engaged in lower levels of careless responding in accordance with their longstring pattern, $p < .001$. There were no differences between the two community groups, $p = .997$, or between the two undergraduate groups, $p = .961$, using a Games-Howell post-hoc test. However, different questionnaires were utilized to assess longstring responding across Studies 1 and 2, which should be taken into consideration when interpreting these results.

When comparing the four groups on memory check accuracy, Levene's test for equality of variances was not met, $F = 32.80$, $p < .001$. Therefore, a one-way Welch ANOVA was used and revealed that the groups significantly differed on accuracy, $F(3, 315.33) = 35.14$, $p < .001$ (Table 3). Given the unequal variances between groups, a Games-Howell post-hoc test was conducted, which corrects for family-wise error (Sauder & DeMars, 2019). Results showed that unpaid community participants in Study 2 performed significantly better than both the Study 2 MTurk sample, $p < .001$, and both Study 1 undergraduate samples,

p < .001, on the memory questionnaire. Specifically, unpaid community participants outperformed MTurkers by 16%, and the undergraduate samples by 16-22% on a memory test on average with a small-to-medium effect size, ω^2 = .04. After correcting for family-wise error, the other samples did not significantly differ on their memory questionnaire scores, *ps* = .122-.998.

General Discussion

Data quality impacts effect sizes, statistical power, error variance, and Types I and II error rates. Therefore, it is paramount to determine which methods produce data of the highest quality. To that end, different approaches to subject recruitment and compensation were examined. Specifically, two secondary datasets were used to compare data quality between undergraduate research participants receiving extra credit, who varied in their opportunity for monetary compensation, and between community samples, who were unpaid or paid for their participation.

The present study took an exploratory approach. Results revealed that undergraduate participants who varied in their opportunity to receive monetary compensation (extra credit vs. extra credit + monetary payment) did not differ on measures of lying to gain access to the study, random responding, longstring, or adherence to a writing task but did perform slightly better on a memory questionnaire. Additionally, the unpaid community sample outperformed the MTurk sample on the memory questionnaire when demographic differences were accounted for. However, this difference was small as it only accounted for 2% of the variance in scores. There were no differences between the unpaid community sample and the MTurk sample on longstring responding (i.e., careless responding).

When comparing across the four samples, the undergraduate samples engaged in lower levels of careless responding in comparison with both community samples. There were no differences in careless responding between the undergraduate samples or between the community samples. Additionally, the community sample outperformed the MTurk sample, and both undergraduate samples, on their respective memory questionnaires. It should be noted that the differences between the community sample and the undergraduate sample may be due to the differing lengths of the memory questionnaires across samples and that the underperformance from undergraduates may simply be due to the increased complexity of the task.

The current results highlight novel findings as well as similarities regarding previous research. First, research has yet to evaluate the influence of monetary compensation on data quality for undergraduate participants who are already receiving extra credit. Results from the present study indicated that financial compensation was not associated with higher quality data in this population — but was associated with a slight improvement on a memory questionnaire performance by paid undergraduate participants. This initial finding is consequential as it highlights that payment may not be a requirement to yield the highest quality data possible in single-session experimental studies collected in university settings. Second, the unpaid community sample outperformed the MTurk sample on a memory questionnaire, but did not differ on longstring (i.e., careless) responding. Thus, single-session experimental studies collected in community samples may not require participant payment to collect high-quality data — though it should be noted that this method of data collection may result in slow recruitment. However, in contrast with previous studies (e.g., Capaldi, 2017; Hauser & Schwarz, 2015), the undergraduate samples outperformed the community samples in terms of engaging in lower levels of careless responding (i.e., longstring responding). In contrast with previous studies, the community sample outperformed both undergraduate samples on their respective memory questionnaires (Kees et al., 2017). Moreover, in contrast with previous studies, MTurkers did not outperform undergraduate samples and did not perform as well as the community sample on the memory questionnaire (Capaldi, 2017; Kees et al., 2017). This is not in alignment with previous research, which has shown that samples collected on MTurk outperform undergraduate samples (e.g., Capaldi, 2017; Hauser & Schwarz, 2015) and other community sampling methods (Kees et al., 2017). However, the current findings are in alignment with previous research showing that samples obtained on MTurk perform as well as other community sampling methods based on longstring responding (Kees et al., 2017).

Researchers should always choose a sample (i.e., undergraduate or community) for which the research question under investigation is most appropriate. If the research question is appropriate for both undergraduate and community samples, then the research team may consider collecting data through an undergraduate sample to obtain data on the phenomenon of interest, yielding data with lower levels of careless responding. Conversely, researchers investigating questions that may best be answered with a general (non-undergraduate) sample should view unpaid community members as a valuable resource. However, MTurk may be a useful option given the potential difficulty associated with recruiting large samples from the general population. Further, MTurk may allow for a more diverse participant pool compared to accessible community samples.

The present study is not without limitations. First, the present research was exploratory in nature and used secondary datasets. Thus, the experiments conducted varied between undergraduate and community participants. This variation could underlie the differences found between the paid undergraduate and community samples. In addition, the use of secondary datasets to assess data quality necessitates that the present study is not causal in nature. To that end, it cannot be concluded that differences in payment caused differences in data quality. However, the use of secondary data allowed for the examination of the relations between subject recruitment, payment, data quality, and memory questionnaire performance, which may guide future experimental studies. Second, in Study 2, there were significant differences in demographic variables between the two community samples including age, ethnicity, gender, and education. In addition, there are likely other unassessed demographic differences between participants who engage in research that is paid versus unpaid (e.g., income). Conversely, paid and unpaid participants in Study 1 did not significantly differ on any demographic variables and no differences were found in data quality between groups. To this end, demographic differences did not seem to influence data quality across studies. However, researchers should be cognizant that by not offering an incentive to participate, the types of people who volunteer may differ from those who would volunteer if they were offered compensation for their time (e.g., in terms of socioeconomic status). Therefore, researchers should be mindful of their research question when choosing their means of data collection. Finally, while longstring variables can detect careless responding in the form of repeatedly selecting the same response option, it should be noted that this type of statistical index cannot detect careless responding where the participant alternates between response options randomly (Johnson, 2005). In addition to this, the measures utilized to detect longstring responding were not consistent across Studies 1 and 2. Rather, the longest measure from each dataset was chosen. Future research will need to validate these findings by comparing careless responding, as detected by longstring, between community and undergraduate participants using the same measures and methodology.

Future research will be needed to validate the current findings. Specifically, a quasi-experimental study where both an undergraduate sample and a community sample are randomly assigned to a payment condition (paid vs. unpaid) is needed to draw causal, within-sample conclusions as well as stronger conclusions between samples of the association between payment and data quality. Moreover, future research will be needed to evaluate the influence of payment on data quality in prospective and longitudinal research.

The present study found that undergraduates had higher data quality in comparison with MTurkers and an unpaid community sample. Additionally, undergraduate students who were offered financial compensation did not have overall higher quality data in comparison with their peers who were not offered financial compensation. Moreover, MTurkers did not vary from their unpaid counterparts on careless responding. Additionally, the present study found that unpaid community participants had higher memory questionnaire scores in comparison with MTurkers as well as undergraduate students who were either offered a monetary incentive or not. Therefore, compensation may be unnecessary when utilizing undergraduate and community samples in one-session experimental research.

References

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the Internet: The effect of 1$ stakes. *PLOS ONE*, *7*(2), Article e31461. https://doi.org/10.1371/journal.pone.0031461

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*, 2586-2596. https://doi.org/10.3758/s13428-018-1035-6

Capaldi, C. (2017). *Graduating from undergrads: Are Mechanical Turk workers more attentive than undergraduate participants?* Fourth Psychology Outside the Box Conference, Ottawa, Canada. https://doi.org/10.13140/RG.2.2.19038.46401

Chandler, J. J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, *8*(5), 500-508. https://doi.org/10.1177/1948550617698203

Chmielewski, M., & Kucker, S. C. (2019). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, *11*, 464-473. https://doi.org/10.1177/1948550619875149

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, *57*(1), 62-79. https://doi.org/10.1086/269355

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, *7*, 613-628. https://doi.org/10.1017/psrm.2018.10

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596-612. https://doi.org/10.1177/0013164410366686

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

Field, A. P. (2009). *Discovering statistics using SPSS* (3rd ed.). SAGE Publications.

Goldammer, P., Annen, H., Stöckli, L. P., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), Article 101384. https://doi.org/10.1016/j.leaqua.2020.101384

Hauser, D. J. & Schwarz, N. (2015). Attentive turkers: MTurk participants perform better on online attention checks than do subjective pool participants. *Behaviour Research Methods*, *48*, 400-407. https://doi.org/10.3758/s13428-015-0578-z

John, O. P., & Srivastava, S. (1999). The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (Vol. 2, pp. 102-138). Guilford Press.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103-129. https://doi.org/10.1016/j.jrp.2004.09.009

Johnson, J. A., & Mayer, J. D. (2019). *SPSS code for longstring screening*. University of New Hampshire. https://mypages.unh.edu/sites/default/files/jdmayer/files/longstring-responding-2020-06-17-jdm-3rded.pdf

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising*, *46*(1), 141-155. https://doi.org/10.1080/00913367.2016.1269304

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavioral Research Methods*, *49*, 433-442. https://doi.org/10.3758/s13428-016-0727-z

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

More, K. R., Phillips, L. A., Green, Z., & Mentzou, A. (2022). Examining self-affirmation as a tactic for recruiting inactive women into exercise interventions. *Applied Psychology: Health and Well-Being*, *14*, 294-310. https://doi.org/10.1111/aphw.12303

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, K. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109-138. https://doi.org/10.1017/XPS.2015.19

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, G. T. (2016). Measuring the prevalence of problematic respondent behaviours among MTurk, campus, and community participants. *PLOS ONE*, *11*(6), Article e0157732. https://doi.org/10.1371/journal.pone.0157732

Padilla-Walker, L. M., Zamboanga, B. L., Thompson, R. A., & Schmersal, L. A. (2005). Extra credit as incentive for voluntary research participation. *Teaching of Psychology*, *32*, 150-153. https://doi/org/10.1207/s15328023top3203_2

Porter, S. R., & Whitcomb, M. E. (2003). The impact of lottery incentives on student survey response rates. *Research in Higher Education*, *44*, 389-407. https://doi.org/10.1023/A:1024263031800

Prochaska, J. O., & DiClemente, C. C. (1982). Transtheoretical therapy: Toward a more integrative model of change. *Psychotherapy: Theory, Research & Practice*, *19*(3), 276-288. https://doi.org/10.1037/h0088437

Rocchi, M., Beaudry, S. G., Anderson, C., & Pelletier, L. G. (2016). The perspective of undergraduate research participant pool nonparticipants. *Teaching of Psychology*, *43*(4), 285-293. https://doi.org/10.1177/0098628316662756

Sauder, D. C., & DeMars, C. E. (2019). An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science*, *2*(1), 26-44. https://doi.org/10.1177/2515245918808784

Sharpiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213-220. https://doi.org/10.1177/2167702612469015

Singer, E., & Ye, C. (2013). The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, *645*, 112-141. https://doi.org/10.1177/0002716212458082

Summerville, A., & Chartier, C. R. (2013). Pseudo-dyadic "interaction" on Amazon's Mechanical Turk. *Behaviour Research Methods*, *45*, 116-124. https://doi.org/10.3758/s13428-012-0250-9

Watson, D., & Clark, L. A. (1994). *The PANAS-X. Manual for the positive and negative affect schedule — expanded form.* The University of Iowa. https://www2.psychology.uiowa.edu/faculty/clark/panas-x.pdf