Computers in Human Behavior 58 (2016) 354–360

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Psychological research in the internet age: The quality of web-based data

Sarah R. Ramsey, Kristen L. Thompson, Melissa McKenzie, Alan Rosenbaum*

Department of Psychology, Northern Illinois University, DeKalb, IL, United States

ARTICLE INFO

Article history:
Received 16 July 2015
Received in revised form
15 December 2015
Accepted 18 December 2015
Available online 21 January 2016

Keywords:
Environmental effects
Human–computer interaction
Performance
Mass media
Test validity

ABSTRACT

The internet is increasingly used in psychological research to solicit participants and collect data. This paper includes two studies examining the quality of data obtained via web-based methods administered either inside or outside the lab. Both studies used item recognition accuracy as a proxy for attention to questions. Study 1 examined the extent to which undergraduate participants ($N = 504$) read and attended to questions either inside or outside the lab. Study 2 ($N = 744$) replicated Study 1, added a Mechanical Turk sample, and examined attention to non-intuitive survey instructions. Results indicated that participants demonstrated good item recognition, regardless of locale or sample; however, small sex effects on accuracy were found in both studies. Specifically, women were more accurate at identifying previously seen items than men in both Study 1 and Study 2. In Study 2, Mechanical Turk participants were more likely to read instructions than undergraduate participants, regardless of whether they participated inside or outside of the lab. The findings support the use of the internet for sampling purposes as well as survey administration, and suggest that researchers use care when studies include non-intuitive instructions.

1. Introduction

Psychological research commonly utilizes undergraduate, subject pool participants, raising questions about both the internal and ecological validity of the research. Are subject pool participants, who are “voluntarily” compelled to participate, giving thoughtful responses? Or are they answering randomly and as quickly as possible? Are they even reading the questions? In traditional research, participants typically provide data in a supervised, laboratory setting (i.e., on-site). This allows investigators to monitor participant impairment (e.g., fatigue) or carelessness and observe whether participants complete surveys independently and without distraction (e.g., multi-tasking). The laboratory also confers a scientific aura that might result in more conscientious participation. Additionally, investigators are available to address questions or provide clarifications.

While much research is still collected in laboratory settings, web-based methodologies are proliferating (Prince, Litovsky, & Friedman-Wheeler, 2012). Prior to the internet, off-site study administration typically involved distributing and receiving materials via mail. However, the growth of internet usage has created new opportunities for researchers. Crowdsourcing services such as Amazon's Mechanical Turk (MTurk) permit recruitment of geographically and culturally diverse participants, enhancing external validity. Data is collected more quickly and prepared more accurately with web-administered surveys than via traditional methods (Gosling, Vazire, Srivastava, & John, 2004). Printing and mailing costs are reduced and participants are paid less, rendering these studies more eco-friendly and economical. Additionally, the flexibility and freedom regarding when and where one participates increases participant convenience and anonymity.

Despite the advantages, the quality of web-based data must be demonstrated. The lack of investigator oversight of off-site, web-based studies makes the amount of attention and care participants exercise in completing them unknown (Rosenbaum & Langhinrichsen-Rohling, 2006). However, some studies have begun to examine whether off-site, web-based methodologies produce comparable results to those administered on-site via traditional methods (e.g., Weigold, Weigold, & Russell, 2013; Zhang et al., 2012). Most, but not all, have demonstrated that on and off-site data are statistically comparable.

Chuah, Drasgow, and Roberts (2006) found no differences in the

* Corresponding author.
E-mail address: arosenbaum@niu.edu (A. Rosenbaum).

http://dx.doi.org/10.1016/j.chb.2015.12.049
0747-5632/

statistical equivalency of data collected via a proctored, on-site, web-administered survey and an un-proctored, off-site, web-administered survey. Templar and Lange (2008) also compared on-site to off-site web-based survey responses and found no differences between sites. However, some studies suggest important site differences beyond statistical equivalency. Paré and Cree (2009) identified differences in participants' item response latency when completing web-based tasks in on-site versus off-site conditions. When rating characteristics of images, off-site participants were significantly slower than on-site participants. Dandurand, Shultz, and Onishi (2008) compared participant accuracy on complex computer-administered problem-solving tasks and found that on-site participants were significantly more accurate than off-site participants. The authors of both studies speculated that decreased attention and/or multi-tasking in the off-site condition may explain these results.

While early studies assessing the comparability of on-site and off-site data collection utilized undergraduate samples, recent research has examined the comparability of research using undergraduate samples to those conducted entirely using web-based methods. The results of these studies have been mixed. Buhrmester, Kwang, and Gosling (2011) found consistency in questionnaire reliability scores across undergraduate and MTurk samples. Paolacci, Chandler, and Ipeirotis (2010) also found that undergraduate and MTurk participants responded similarly on judgment and decision-making tasks. Further, using one item to assess attention to survey items (i.e., "While watching the television, have you ever had a fatal heart attack?"), the authors found no significant differences in incorrect responses across groups. However, a study conducted by Goodman, Cryder, and Cheema (2013) incorporated complex instructions at the end of their survey and found undergraduate participants were significantly more likely to follow instructions than MTurk participants.

The internet has changed the way we live and it is starting to change the way we do research. Given that recent studies suggest there may be important differences in the extent to which on-site versus off-site participants attend to web-based surveys, we conducted two studies to further examine web-based research. Study 1 examined whether undergraduate participants (in both on-site, and off-site, web-based conditions) attended to survey questions and whether there were differences attributable to site. Study 2 replicated Study 1 and expanded upon it by adding an MTurk comparison group and also assessing the extent to which participants attended to survey instructions.

2. Study 1

Study 1 employed a novel strategy to examine the extent to which undergraduates attend to (i.e., read and process) survey items presented via the internet, either in the lab or at an off-site location (and time) of their choosing. Recognition accuracy of previously seen items served as a proxy for attending. Presumably, participants that read survey questions would be better at recognizing those questions (embedded in a second survey) compared with participants who responded without carefully reading the questions. Given the dearth of research regarding participant attending between on-site and off-site conditions, no formal hypotheses were proposed.

2.1. Method

2.1.1. Participants

The G*Power formula (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that a total sample size of 158 was necessary to obtain a power of .95, assuming a small effect size (.25) and setting a significance level of $p < .05$ based on the statistical analyses conducted. Data was collected at a large, public Midwestern university for the duration of one semester, resulting in a sample of 504 undergraduate students ranging in age from 18 to 37 ($M = 19.35$; $SD = 2.02$). Additional descriptive statistics are presented in Tables 1 and 2. Participants received research credit applied to a course research requirement.

2.1.2. Measures

2.1.2.1. Demographic questionnaire. All participants completed a short demographic questionnaire assessing age, sex, race/ethnicity, and socioeconomic status. Off-site participants were also asked to specify the device they used (i.e., desktop computer, laptop, smart phone, other).

2.1.2.2. Questionnaires. Questionnaire 1 (Q1) included 50-items drawn from a variety of existing measures assessing both sensitive and non-sensitive topics. These included: the Balanced Time Perspectives Scale (Webster, 2011), the Attitudes Toward Emotions Scale (Harmon-Jones, Harmon-Jones, Amodio, & Gable, 2011), the Sexual Risk Survey (Turchik & Garske, 2009), the Short Sadistic Impulse Scale (O'Meara, Davies, & Hammond, 2011), the Behavioral Undercontrol Questionnaire (Stice, Myers, & Brown, 1998), the Caffeine Expectancy Questionnaire (Huntley & Juliano, 2012), the Alcohol Use Disorders Identification Test: Self-Report (Saunders, Aasland, Babor, De La Fuente, & Grant, 1993), and the Revised

Table 1
Participant descriptives: Study 1 and Study 2.

	On-site ($n = 250$)		Off-site ($n = 254$)		M-Turk –		Total ($N = 504$)	
	n	%	*N*	%	*n*	%	*N*	%
Study 1								
Gender								
Male	89	35.6	110	43.3	–	–	199	39.5
Female	161	64.4	144	56.7	–	–	305	60.5
Race/Ethnicity								
Caucasian	124	49.6	150	59.1	–	–	274	54.4
African American	62	24.8	50	19.7	–	–	112	22.2
Hispanic	24	9.6	26	10.2	–	–	50	9.9
Asian American	9	3.6	12	4.7	–	–	21	4.2
Native American	2	.8	0	0	–	–	2	.4
Indian	3	1.2	2	.8	–	–	5	1.0
Multiracial	25	10.0	13	5.1	–	–	38	7.5
Missing	1	.4	1	.4	–	–	2	.4
	On-site ($n = 251$)		Off-site ($n = 247$)		M-Turk ($n = 246$)		Total ($N = 744$)	
Study 2								
Gender								
Male	101	40.2	101	40.9	105	42.7	307	41.3
Female	150	59.8	144	58.3	141	57.3	435	58.5
Education								
Some high school	0	0	0	0	3	1.2	3	.4
High school diploma	0	0	0	0	24	9.8	24	3.23
Some college	242	96.4	228	92.3	73	29.7	543	72.98
Bachelor's/Assoc.	9	3.6	18	7.3	106	43	133	17.9
Graduate education	0	0	0	0	40	16.3	41	5.5
Race/Ethnicity								
Caucasian	145	57.8	143	57.9	188	76.4	476	64.0
African American	51	20.3	43	17.4	19	7.7	113	15.2
Hispanic	16	6.4	27	10.9	5	2.0	48	6.5
Asian American	16	6.4	13	5.3	14	5.7	43	5.8
Native American	1	.4	0	0	4	1.6	5	.7
Indian	1	.4	0	0	1	.4	2	.3
Pacific Islander	1	.4	1	.4	0	0	2	.3
Multiracial	16	6.4	17	6.9	13	5.3	46	6.2
Other	4	1.6	2	.8	1	.4	7	.9
Missing	0	0	1	.4	1	.4	2	.3

Table 2
Participant age descriptives: Study 1 and Study 2.

	On-site ($n = 250$)			Off-site ($n = 254$)			M-Turk			Total ($N = 504$)		
	M	*SD*	R	*M*	*SD*	R	*M*	*SD*	R	*M*	*SD*	R
Study 1	19.36	1.81	18–35	19.33	2.21	18–37	–	–	–	19.35	2.02	18–37
	On-site ($n = 251$)			Off-site ($n = 247$)			M-Turk ($^*n = 245$)			Total ($N = 743$)		
	M	*SD*	R	*M*	*SD*	R	*M*	*SD*	R	*M*	*SD*	R
Study 2	19.26	1.75	18–28	19.32	2.76	18–48	34.79	12.17	18–73	24.40	10.27	18–73

Note. R = Range. * = One participant in the M-Turk sample elected to not enter their specific age, resulting in an *n* of 245 instead of 246.

Attitudes Toward Violence Scale (Anderson, Benjamin, Wood, & Bonacci, 2006). Of the 50 items, 25 reflected content of a sensitive nature (e.g., sexual behaviors, violent behaviors) and 25 reflected content of a non-sensitive nature (e.g., caffeine consumption, study habits). Half of each type utilized a multiple choice response format and half utilized a Likert-type response format in order to assess whether there were any effects due to question content or response format.

Questionnaire 2 (Q2) also contained 50 items, including varying numbers of items that appeared in Q1 and novel items that were not in Q1. A pool of 30 new items were written to be similar enough in content and style to Q1 items so as not to be too easily identified, but not so similar as to trick participants who had actually attended to the questions on Q1. For example, the Q1 item, "Do you typically vote in elections?" was replaced by the novel Q2 item, "How would you identify your current political affiliation?" Three different forms of Q2 were created differing only in the number of previously viewed items (i.e., 20, 30, or 40 of the 50 Q1 items) along with the requisite number of previously unseen items to bring the total to 50 (i.e., 30, 20, or 10 new items). A random number generator was used to determine which questions would be included on each form. Repeated items from Q1, as well as the requisite number of novel items in each form of Q2, were randomly selected. As participants' actual responses to the content of the items on Q1 and Q2 were not of interest, there were no concerns regarding the psychometric properties of the questionnaires. Additionally, given that the questionnaires were compiled from a number of other existing surveys, reliability scores for the study questionnaires would not provide any meaningful information and are not reported.

2.1.3. Procedure

Participants were recruited online through SONA Systems, the University's research catalog. The study was posted twice, differing only with regard to site (i.e., on-site or off-site). Participants could only participate once and were allowed to decide whether to participate in the on-site or off-site condition (as opposed to being randomly assigned to a condition). This arguably increases the ecological validity of the study, in that participants who choose to complete studies off-site may perform differently from students who choose to complete studies in the lab. Those who signed up to complete the survey off-site were sent a username, password, and link to the survey and completed the survey at a time and location of their choosing. On-site participants used SONA to schedule a time to participate and were administered the survey, in groups of one to four, on a computer in a laboratory. Experimenters were present while participants completed the surveys in the lab.

All participants first completed Q1. The instructions led participants to believe that the content of the questions was the subject of the experiment. After they completed Q1 they were immediately presented with Q2 and instructed to identify those questions they had just seen, or not seen, on Q1 by answering "yes" or "no," respectively, to each item. Participants were randomized to complete one of the three forms of Q2 (containing 20, 30, or 40 of the Q1 items) in order to control for potential differences in recognition accuracy based on participant guessing and tendencies toward yea-, or nay-saying. It was presumed that if participants engaged in yea-saying or nay-saying, then accuracy would be higher for versions of Q2 that included 40 previously administered items in the case of yea-saying, or 20 previously administered items in the case of nay-saying. Further, if participants were merely guessing, recognition accuracy would appear higher in versions of Q2 that included more of the Q1 items, as participants would have a higher likelihood of being correct based on chance alone.

2.2. Study 1 results

Data analyses were performed using SPSS 22.0 statistical software. Two participants elected not to disclose information about their race. No other missing values were identified and no extreme outliers were detected. Accordingly, data from all 504 participants were included. Recognition accuracy variables were not normally distributed; however, according to the Central Limit Theorem, normally distributed data are not essential when data sets are large, as in the present case (Tabachnick & Fidell, 2013). Therefore, the recognition accuracy variables were not transformed. All on-site participants used a desktop computer. Of those participating off-site, 84% used a laptop, 15% used a desktop computer, and 1% used a smart phone.

Accuracy was indexed as participants' ability to correctly identify items on Q2 as having been seen or not seen on Q1. Across sites, participants had an 89% hit rate regarding whether items on Q2 had been previously seen on Q1 ($M = 44.57$, $SD = 8.57$). Recognition accuracy was assessed in relation to sex and age. Results indicated that sex, but not age, was related to recognition accuracy; therefore, sex was included as a control variable in subsequent analyses. Race was not evaluated as a covariate because there were not enough participants from any of the non-Caucasian racial/ethnic groups to make meaningful comparisons.

The results of a 2 × 2 × 3 (site × sex × form [20, 30, or 40 previously seen items]) analysis of variance (ANOVA) revealed a significant site by sex interaction, $F(1, 494) = 4.20$, $p = .04$, $\eta_p^2 = .008$. When completing the questionnaires off-site, women ($M = 45.43$, $SD = 7.65$) were significantly more accurate than men ($M = 41.90$, $SD = 10.75$), $F(1, 494) = 10.67$, $p = .001$, $\eta_p^2 = .021$, whereas on-site there were no differences between women ($M = 45.31$, $SD = 8.00$) and men ($M = 45.16$, $SD = 7.38$), $F(1, 494) = .07$, $p = .79$, $\eta_p^2 < .001$. The main effects of site, $F(1, 494) = 4.18$, $p = .04$, $\eta_p^2 = .008$, and sex, $F(1, 494) = 5.90$, $p = .02$, $\eta_p^2 = .012$, were also significant. According to Cohen's (1988) guidelines for interpreting effect sizes, (i.e., $.01 \leq \eta_p^2 \leq .08$ is

considered a small effect), the simple effect for sex differences off-site was small but significant. Finally, there was no effect of Q2 form, $F(2, 494) = 2.09$, $p = .12$, $\eta_p^2 = .008$, nor an interaction between site and form $F(2, 494) = .40$, $p = .67$, $\eta_p^2 = .002$, indicating that the proportion of questions from Q1 included in Q2 was not significantly related to recognition accuracy (See Table 3 for means).

Site differences in recognition accuracy based on question content (sensitive vs. non-sensitive) and response option type (Likert-type vs. multiple choice) were also assessed. The within-subjects effect of question content was significant, $F(1, 502) = 4.02$, $p = .05$, in that participants were more accurate at correctly identifying the non-sensitive items compared to the sensitive items (See Table 4 for means). The between-subjects effect of group (on-site vs. off-site) was trending, $F(1, 502) = .30$, $p = .08$, in that on-site participants were marginally more accurate than off-site participants (See Table 4 for means). However, in both cases, the effect size ($\eta_p^2 = .008$ and .006, respectively) did not exceed the accepted threshold for a small effect (Cohen, 1988). The interaction between site and question content, $F(1, 502) = .03$, $p = .87$, $\eta_p^2 < .001$, was not significant.

The within-subjects effect of response option type was not significant, $F(1, 502) = 1.57$, $p = .21$, $\eta_p^2 = .003$ (See Table 4 for means). The interaction between site and response option type, $F(1, 502) = .15$, $p = .70$, $\eta_p^2 < .001$, was not significant. The between-subjects effect of group (on-site vs. off-site) was also not significant, $F(1, 502) = 1.01$, $p = .32$, $\eta_p^2 = .002$ (See Table 4 for means).

2.3. Study 1 discussion

Study 1 examined the extent to which undergraduates attended to survey questions and whether there were site differences in recognition accuracy. Overall, participants identified 89% of the questions correctly. This result is substantially superior to chance and far exceeds what would be possible if participants were not attending to questions. The study design also provides assurances that participants were not answering randomly, as recognition accuracy on the three forms of Q2 did not differ across, or between, sites. While these findings do not guarantee honest participant responding, the fact that they recognized a high percentage of previously seen items suggests they attended to the questions.

Interestingly, women's recognition accuracy was higher than men's in the off-site, but not on-site, condition, suggesting that women may participate more conscientiously than men when unsupervised. Further, participants were more likely to correctly identify non-sensitive, compared to sensitive, items. However, the small effect sizes of both these findings would advise cautious interpretation. Study 1 findings are encouraging, as it appears undergraduates read items before responding to them, regardless of whether they complete questionnaires in or outside of the lab.

Table 3
Average participant overall and between-site accuracy based on Q2 format.

	On-site ($n = 250$)		Off-site ($n = 254$)		M-Turk		Total ($N = 504$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Study 1								
Q2 format								
10 unseen items	44.77	8.86	42.83	11.21	–	–	43.79	10.13
20 unseen items	45.60	7.48	45.31	7.74	–	–	45.45	7.59
30 unseen items	45.32	6.90	43.23	8.78	–	–	44.27	7.94
	On-site ($n = 251$)		Off-site ($n = 247$)		M-Turk ($n = 246$)		Total ($N = 744$)	
Study 2								
Q2 format								
10 unseen items	45.82	7.95	46.52	6.75	43.56	9.89	45.35	8.30
20 unseen items	47.45	4.59	45.47	8.35	45.48	7.35	46.16	6.94
30 unseen items	45.84	6.48	46.01	6.74	45.76	6.62	45.87	6.59

Note. Format means represent the number of correctly categorized (i.e., as seen or unseen) items out of a possible 50 items.

Table 4
Average participant overall and between - site accuracy based on question content and response option format.

	On-site ($n = 250$)		Off-site ($n = 254$)		M-Turk		Total ($N = 504$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Study 1								
Question content								
Sensitive	22.54	3.84	21.84	4.62	–	–	22.19	4.26
Non-sensitive	22.72	4.18	22.06	4.94	–	–	22.39	4.58
Response option								
Likert-type	13.51	4.55	13.11	4.90	–	–	13.31	4.73
Multiple choice	13.45	4.67	13.00	4.95	–	–	13.23	4.81
	On-site ($n = 251$)		Off-site ($n = 247$)		M-Turk ($n = 246$)		Total ($N = 744$)	
Study 2								
Question content								
Sensitive	22.96	3.31	22.94	3.63	22.42	3.97	22.77	3.65
Non-sensitive	23.40	3.38	23.09	3.83	22.60	4.24	23.03	3.84
Response option								
Likert-type	13.49	4.51	13.79	4.70	12.96	4.78	13.41	4.67
Multiple choice	13.30	4.59	13.65	4.85	12.96	4.67	13.31	4.70

Note. Question content means represent the number of correctly categorized items out of a possible 25 items, as Q2 included 25 sensitive items and 25 non-sensitive items. Response option format means represent the number of correctly categorized items out of a possible 10, 15, or 20 items, depending on which form of Q2 each participant received. Q2 accuracy for response option format could only be calculated for correct identification of previously seen items because unseen items included in Q2 did not include response options.

3. Study 2

As noted, crowdsourced participants are increasingly replacing undergraduate samples in research. Study 2 replicated and extended Study 1 by utilizing the same design and including an additional off-site, MTurk sample. Study 2 also explored the extent to which participants across sites and sample groups attended to survey instructions.

3.1. Method

3.1.1. Participants

A total of 744 participants completed the study. Participants ranged in age from 18 to 73 ($M = 24.40$; $SD = 10.27$). Additional descriptive statistics for Study 2 are presented in Tables 1 and 2.

3.1.2. Measures

All measures from Study 1 were included in Study 2. A question regarding education level and an additional set of instructions were added. At the beginning of the instructions, participants were given a non-intuitive directive regarding how to complete the first item of the study. Participants were asked, “Please enter the current time of day, as you are beginning this survey. To assist us with coding our data, please include the time followed by the number one in parentheses if you are completing this survey before noon, or the number two in parentheses if you are completing this survey in the afternoon or evening.” Compliance with this instruction was used as an indicator of whether participants read and followed directions.

3.1.3. Procedure

All undergraduate participants enrolled in the study using the SONA system and received participation credit as described in Study 1. MTurk participants were recruited using Amazon's Mechanical Turk and included only U.S. residents. MTurk participants were compensated $.25.

3.2. Study 2 results

Study 2 data were screened and cleaned in the same manner as Study 1 data. No missing values or extreme cases were identified; data from all 744 participants were included in analyses. All on-site undergraduate participants completed the questionnaires on a desktop computer. In the off-site undergraduate group, 85% used a laptop, 11% used a desktop computer, and 5% used a smart phone or tablet. Among MTurk participants, 57% used a laptop, 39% used a desktop computer, and 5% used a smart phone or tablet.

Across sites, participants were able to correctly identify whether items on Q2 had been previously seen on Q1 92% of the time ($M = 45.81$, $SD = 7.25$). When indicated, age and sex were controlled for in the on-site versus off-site undergraduate comparisons and the MTurk versus off-site undergraduate comparisons. Race was not evaluated as a covariate because there were not enough participants from any of the non-Caucasian racial/ethnic groups to make meaningful comparisons.

3.2.1. On-site, versus off-site, undergraduate comparisons

A 2 × 2 × 3 (group [on-site vs. off-site] × sex × form [20, 30, or 40 previously seen items]) ANOVA was conducted to assess for between-groups differences in recognition accuracy among undergraduates. Those who completed the study on-site ($M = 46.36$, $SD = 6.45$) did not significantly differ from those who completed it off-site ($M = 46.01$, $SD = 7.28$) with regard to their ability to correctly identify Q2 items seen on Q1, $F(1, 490) = .30$, $p = .59$, $\eta_p^2 = .001$. Across groups, accuracy was also not significantly related to form, $F(2, 490) = .27$, $p = .76$, $\eta_p^2 = .001$ (See Table 4 for means). In contrast, sex was significantly related to accuracy, $F(1, 490) = 7.7$, $p = .01$, $\eta_p^2 = .016$, such that women ($M = 46.90$, $SD = 5.80$) were more accurate than men ($M = 45.15$, $SD = 8.07$). However, in contrast to Study 1, there was no interaction between site and sex, $F(1, 490) < .001$, $p = .99$, $\eta_p^2 < .001$.

The effects of question content (sensitive vs. non-sensitive items) and response option type (Likert-type vs. multiple choice) were also assessed. The within-subjects effect of question content was significant, $F(1, 496) = 14.06$, $p < .001$, in that participants were more accurate in identifying the non-sensitive items compared to the sensitive items (See Table 4 for means). The effect size was small, but meaningful ($\eta_p^2 = .028$; Cohen, 1988). The interaction between site and question content, $F(1, 496) = 3.16$, $p = .08$, was trending. However, the effect size ($\eta_p^2 = .006$) did not exceed the accepted threshold for a small effect (Cohen, 1988). The between-subjects effect of group (on-site vs. off-site) was not significant, $F(1, 496) = .30$, $p = .59$, $\eta_p^2 = .001$ (See Table 4 for means).

The within-subjects effect of response option was significant, $F(1, 496) = 5.93$, $p = .02$, in that participants were more accurate at correctly identifying the Likert-type items compared to the multiple choice items (See Table 4 for means). The effect size ($\eta_p^2 = .012$) was small, yet meaningful (Cohen, 1988). The interaction between site and response option, $F(1, 496) = .17$, $p = .68$, $\eta_p^2 < .001$, was not significant. The between-subjects effect of group (on-site vs. off-site) was not significant, $F(1, 496) = .61$, $p = .44$, $\eta_p^2 = .001$ (See Table 4 for means).

With regard to instruction compliance, 14.7% (n = 37) of on-site undergraduates and 8.5% (n = 21) of off-site undergraduates followed directions. Logistic regression was used to determine whether participation location significantly predicted participants' likelihood of complying with instructions. For the off-site versus on-site undergraduate comparison, the full model tested against the null model was statistically significant, indicating that location significantly predicted the likelihood of following directions (χ^2 (1, $N = 498$) = 4.77, $p = .03$). The overall predictive success of the model was 88.4%. Nagelkerke's R^2 was .02, indicating that the model explained 2% of the variance. The Wald statistic (4.60) indicated that location significantly predicted following directions ($p = .03$). The odds ratio for location (on-site vs. off-site) was 1.86, indicating that following instructions increased by a factor of 1.86 if the participant was on-site compared to off-site. Table 5 provides the raw score binary logistic regression coefficients, Wald statistics, odds ratios (i.e., Exp(B)), and 95% confidence intervals for the odds ratios.

3.2.2. Mechanical Turk versus off-site undergraduate comparison

Results of a 2 × 2 × 3 (group × sex × form) ANOVA indicated that MTurk participants ($M = 45.02$, $SD = 7.96$) and off-site undergraduates ($M = 46.01$, $SD = 7.28$) did not significantly differ from each other regarding their accuracy at identifying previously seen questionnaire items, $F(1, 485) = 2.18$, $p = .14$, $\eta_p^2 = .004$. Consistent with findings from Study 1 and from the on-site/off-site undergraduate comparison in Study 2, accuracy was not significantly related to form, $F(2, 485) = .37$, $p = .69$, $\eta_p^2 = .002$ (See Table 4 for means). Once again, the effect of sex was significant, $F(1, 485) = 7.68$, $p = .01$, $\eta_p^2 = .016$, such that women ($M = 46.34$, $SD = 6.59$) were more accurate than men ($M = 44.37$, $SD = 8.77$). The interaction between participant group and sex was not significant, $F(1, 485) = .09$, $p = .76$, $\eta_p^2 < .001$, indicating that women were more accurate than men across groups.

The effects of question content (sensitive vs. non-sensitive items) and response option type (Likert-type vs. multiple choice) on recognition accuracy were also assessed. The within-subjects effect of question content was significant, $F(1, 491) = 4.00$, $p = .05$, in that participants were more accurate at correctly identifying the non-sensitive items compared to the sensitive items (See Table 4 for means). However, the effect size ($\eta_p^2 = .008$) did not exceed the accepted threshold for a small effect (Cohen, 1988). The interaction between site and question content, $F(1, 491) = .02$, $p = .88$, $\eta_p^2 < .001$, was not significant. The between-subjects effect of group (off-site vs. MTurk) was also not significant, $F(1, 491) = 2.16$, $p = .14$, $\eta_p^2 = .004$ (See Table 4 for means). The within-subjects effect of response option type was not significant, $F(1, 491) = .96$, $p = .33$, $\eta_p^2 = .002$ (See Table 4 for means). The interaction between site and question type, $F(1, 491) = 1.22$, $p = .27$, $\eta_p^2 = .002$, was also not significant. The between-subjects effect of group (off-site vs. MTurk) was trending, $F(1, 491) = 3.22$, $p = .07$, in that off-site undergraduates were marginally more accurate than MTurk participants (See Table 4 for means). However, once again, the effect size ($\eta_p^2 = .007$) did not exceed the accepted threshold for a small effect (Cohen, 1988).

With regard to instruction compliance, 8.5% (n = 21) of off-site undergraduates and 49.6% (n = 122) of MTurk participants

Table 5
Study 2 logistic regression predicting the likelihood of participants following study instructions.

	B	S.E.	Wald	df	*p*	Exp(B)	CI lower	CI upper
On-site vs. Off-site								
Location	.62	.29	4.60	1	.03	1.86	1.06	3.28
M-Turk vs. Off-site								
Location	2.11	.31	47.06	1	<.001	8.24	4.51	15.05
Age	−.02	.01	2.53	1	.11	.98	.96	1.00

followed directions. Hierarchical binary logistic regression was used to determine whether participant group (i.e., off-site undergraduates versus MTurk participants) significantly predicted participants' likelihood of complying with instructions, while controlling for age. For the off-site undergraduate versus MTurk participant comparison, the full model tested against the null model was statistically significant, indicating that group significantly predicted the likelihood of following directions (χ^2 (2, $N = 492$) = 112.33, $p < .001$). The overall predictive success of the model was 73.4%. Nagelkerke's R^2 was .30, indicating that the full model explained 30% of the variance. The Wald statistic indicated that group (47.06) significantly contributed to the predictive value of the model ($p < .001$), whereas age (2.53) did not ($p = .11$). The odds ratio for group (MTurk vs. off-site) was 8.24, indicating that following instructions increased by a factor of 8.24 if the participant was in the MTurk, compared to off-site, undergraduate group. Table 5 provides the raw score binary logistic regression coefficients, Wald statistics, odds ratios (i.e., Exp(B)), and 95% confidence intervals for the odds ratios.

3.3. Study 2 discussion

Consistent with Study 1, participants were quite accurate in their recognition of previously seen items (92%) and there were no significant differences in recognition accuracy between on-site and off-site undergraduates or when comparing off-site undergraduates to MTurk participants. In both comparisons, the number of repeated questions (i.e., form) did not significantly impact recognition accuracy, suggesting that participants across conditions were not randomly responding. As in Study 1, women were more accurate than men across the two comparisons and all participants were significantly more accurate at identifying non-sensitive items. Unlike Study 1, there was no site-by-sex interaction. Additionally, in the on-site versus off-site undergraduate comparison, participants more accurately identified Likert-type compared to multiple choice questions.

Among undergraduates, location was significantly related to instruction compliance; on-site participants were more likely to comply with instructions. However, the small effect size suggests the on-site/off-site distinction may not be important. Contrary to Goodman et al.'s (2013) findings that undergraduates were more likely to follow instructions than MTurk participants, MTurk participants were far more likely than off-site undergraduates to comply with survey instructions in the current study. Present findings suggest MTurk participants may take a more conscientious, careful approach to research compared with undergraduates. This is consistent with previous findings that MTurk participants report an interest in completing surveys because of an intrinsic interest in research (Kaufmann, Schulze, & Veit, 2011).

While MTurk participants were more compliant with survey instructions, the overall low compliance rate across groups (24.2%) is notable. In general, it may be unnecessary to read instructions; however, some studies may include atypical instructions (e.g., “Answer these questions as your wife would.”). If participants assume they know what to do and do not read instructions, the impact on data validity could be significant when directions are not intuitive. Researchers administering surveys with complicated or non-intuitive instructions should be concerned by these results and adjust methods accordingly.

4. Discussion

The weakest link in psychological research may be the quality of data provided by participants, many of whom are undergraduates “coerced” into participation as one way to satisfy class research requirements. Questions remain regarding whether these participants are honestly engaging in the research and if they are even reading the questions and/or instructions. The internet increasingly permits research to shift from the supervised laboratory environment to less controlled locations (e.g., at home), making the question of participant attention to surveys even more important. Further, as the popularity of crowdsourcing sites increases, we are obligated to question the validity of such data.

The present studies demonstrate that, regardless of completion locale, participants attend to survey items sufficiently to subsequently recognize approximately 90% of them. Thus, while the contexts which off-site participants completed surveys are unknown, off-site recognition accuracy was not degraded compared to accuracy in the lab. Given this, we might speculate that if participants read questions carefully enough to remember seeing them, they may also participate conscientiously. Nevertheless, we must acknowledge that participant accuracy in recognizing previously seen items does not necessarily equate to honest responding.

The news is not as good regarding whether participants attend to instructions. Anecdotally, several Study 1 on-site participants asked questions suggesting they had not read the somewhat atypical instructions. This prompted us to examine instruction compliance in Study 2 and the results were concerning. Fewer than half of the MTurk participants and less than 15% of undergraduates followed the non-intuitive instruction. This is not surprising. People often skip instructions and only consult them if necessary. In support of this, Q2 had atypical instructions, yet recognition accuracy was high. If participants did not initially read the instructions to answer “yes” or “no” to whether an item was seen on Q1, it seems unlikely that they could have completed the questionnaire as accurately as they did without going back to read them.

When surveys have non-standard instructions, researchers should take steps to ensure participants have read and understood them (e.g., preventing participants from starting surveys until they correctly answer a question about the directions). Researchers can also present instructions using an audio track, emphasizing atypical instructions. Additionally, questions might be framed more explicitly (e.g., Would your wife agree that you participate equally in child-rearing?).

Significant sex differences were found in both studies. In Study 1, men had less accurate item recognition off-site than on-site, while women's accuracy was not significantly different across sites, suggesting that men do not pay as close attention to questions when unsupervised. This interaction did not replicate in Study 2, but we found that women had significantly better recognition accuracy in all conditions, again possibly suggesting that women exhibit greater conscientiousness. In interpreting these findings, it is important to note both that the effects sizes for sex differences were small, but that women's greater accuracy was consistent.

Both studies demonstrated a small effect for better recall of non-sensitive, compared to sensitive, questions. While the reason for this is unclear, it does enhance our confidence that participants were attending to the questions, in that recall errors were non-random. Similarly, Study 2 demonstrated a small advantage for recognizing Likert-type, compared to multiple choice, questions. This may be because Likert-type scales provide consistent response options, meaning there is no additional substantive content in the response options to distract from item recognition.

Given the increase in crowdsourcing, it is meaningful that MTurk participants were statistically equal to undergraduates on the item recognition task and far superior on the instruction task, with almost 50% following the non-intuitive instruction. These results should increase our confidence in data provided by crowdsourced participants. However, while the results suggest participants attended to questions, we have no assurance they

answered truthfully. Further assessment for honest responding in web-based studies is necessary, as this also relates to the quality of web-based data. Another possible limitation to the current findings is that, in both Study 1 and Study 2, undergraduate participants were allowed to self-select whether to complete the study on-site or off-site as opposed to being randomly assigned to a condition. While we asserted that this increases the ecological validity of the studies, it might also be argued that this confounds site differences with individual differences.

Like everything, research methods evolve as new technologies become available. Web-based survey administration and crowdsourcing are powerful new tools that expand the reach of researchers, improving external validity. They save resources and are convenient, benefiting both participants and researchers. However, this matters little if the data are not valid. The studies in this paper provide support for web-based recruitment and survey administration. Although we are concerned about poor instruction compliance, we believe researchers can take simple steps to remediate this problem.

References

Anderson, C. A., Benjamin, A. J., Wood, P. K., & Bonacci, A. M. (2006). Development and testing of the velicer attitudes toward violence scale: evidence for a four-factor model. *Aggressive Behavior, 32*(2), 122–136. http://dx.doi.org/10.1002/ab.20112.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of cheap, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. http://dx.doi.org/10.1177/1745691610393980.

Chuah, S. C., Drasgow, F., & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality, 40*(4), 359–376. http://dx.doi.org/10.1016/j.jrp.2005.01.006.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods, 40*(2), 428–434. http://dx.doi.org/10.3758/BRM.40.2.428.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. http://dx.doi.org/10.3758/BF03193146.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213–224. http://dx.doi.org/10.1002/bdm.1753.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59*(2), 93–104. http://dx.doi.org/10.1037/0003-066X.59.2.93.

Harmon-Jones, E., Harmon-Jones, C., Amodio, D. M., & Gable, P. A. (2011). Attitudes toward emotions. *Journal of Personality and Social Psychology, 101*(6), 1332–1350. http://dx.doi.org/10.1037/a0024951.

Huntley, E. D., & Juliano, L. M. (2012). Caffeine Expectancy Questionnaire (CaffEQ): Construction, psychometric properties, and associations with caffeine use, caffeine dependence, and other related variables. *Psychological Assessment, 24*(3), 592–607. http://dx.doi.org/10.1037/a0026417.

Kaufmann, N., Schulze, T., & Veit, D. (2011, August). More than fun and money. Worker motivation in crowdsourcing – A study on Mechanical Turk. In *Paper presented at the seventeenth Americas conference on information systems, Detroit, Michigan*. Abstract retrieved from http://aisel.aisnet.org/amcis2011_submissions/340/.

O'Meara, A., Davies, J., & Hammond, S. (2011). The psychometric properties and utilities of the Short Sadistic Impulse Scale (SSIS). *Psychological Assessment, 23*(2), 523–531. http://dx.doi.org/10.1037/a0022400.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411–419. Retrieved from http://journal.sjdm.org.

Paré, D. E., & Cree, G. S. (2009). Web-based image norming: How do object familiarity and visual complexity ratings compare when collected in-lab versus online? *Behavior Research Methods, 41*(3), 699–704. http://dx.doi.org/10.3758/BRM.41.3.699.

Prince, K. R., Litovsky, A. R., & Friedman-Wheeler, D. G. (2012). Internet-mediated research: Beware of bots. *The Behavior Therapist, 35*(5), 85–88. Retrieved from http://www.abct.org/Journals/?m=mJournal&fa=TBT.

Rosenbaum, A., & Langhinrichsen-Rohling, J. (2006). Meta-research on violence and victims: The impact of data collection methods on findings and participants. *Violence and Victims, 21*(4), 404–409. http://dx.doi.org/10.1891/0886-6708.21.4.404.

Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R., & Grant, M. (1993). Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO collaborative project on early detection of persons with harmful alcohol consumption: II. *Addiction, 88*(6), 791–804. http://dx.doi.org/10.1111/j.1360-0443.1993.tb02093.x.

Stice, E., Myers, M. G., & Brown, S. A. (1998). A longitudinal grouping analysis of adolescent substance use escalation and de-escalation. *Psychology of Addictive Behaviors, 12*(1), 14–27. http://dx.doi.org/10.1037/0893-164X.12.1.14.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Indianapolis: Pearson.

Templar, K. J., & Lange, S. R. (2008). Internet testing: Equivalence between proctored lab and unproctored filed conditions. *Computers in Human Behavior, 24*(3), 1216–1228. http://dx.doi.org/10.1016/j.chb.2007.04.006.

Turchik, J. A., & Garske, J. P. (2009). Measurement of sexual risk taking among college students. *Archives of Sexual Behavior, 38*(6), 936–948. http://dx.doi.org/10.1007/s10508-008-9388-z.

Webster, J. D. (2011). A new measure of time perspective: Initial psychometric findings for the Balanced Time Perspective Scale (BTPS). *Canadian Journal of Behavioural Science, 43*(2), 111–118. http://dx.doi.org/10.1037/a0022801.

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods. *Psychological Methods, 18*(1), 53–70. http://dx.doi.org/10.1037/a0031607.

Zhang, S., Wu, Q., van Velthoven, M. H., Chen, L., Car, J., Rudan, I. … Scherpbier, R. W. (2012). Smartphone versus pen-and-paper data collection of infant feeding practices in rural China. *Journal of Medical Internet Research, 14*(5), 156–167. http://dx.doi.org/10.2196/jmir.2183.