ASSOCIATION FOR CONSUMER RESEARCH

Labovitz School of Business & Economics, University of Minnesota Duluth, 11 E. Superior Street, Suite 210, Duluth, MN 55802

Non-Naïve Participants Can Reduce Effect Sizes

Jesse Chandler, University of Michigan, USA
Gabriele Paolacci, Erasmus University Rotterdam, The Netherlands
Eyal Peer, Bar-Ilan University, Israel
Pam Mueller, Princeton University, USA
Kate Ratliff, University of Florida, USA

Although researchers often assume their participants are naïve to experimental materials, this is not always the case. We investigated how results in a set of two-condition experiments depended on whether participants completed them once already. Non-naivety markedly decreased effect sizes, especially when participants were assigned to a different condition.

[to cite]:

Jesse Chandler, Gabriele Paolacci, Eyal Peer, Pam Mueller, and Kate Ratliff (2015) ,"Non-Naïve Participants Can Reduce Effect Sizes", in NA - Advances in Consumer Research Volume 43, eds. Kristin Diehl and Carolyn Yoon, Duluth, MN : Association for Consumer Research, Pages: 18-22.

[url]:

http://www.acrwebsite.org/volumes/1020052/volumes/v43/NA-43

[copyright notice]:

This work is copyrighted by The Association for Consumer Research. For permission to copy or use this work in whole or in part, please contact the Copyright Clearance Center at http://www.copyright.com/.

Data Quality in Online Research: Challenges and Solutions

Chairs: Neil Brigden, Miami University, USA
Gabriele Paolacci, Erasmus University Rotterdam, The Netherlands

Paper #1: Beyond the Turk: An Empirical Comparison of Alternative Platforms for Crowdsourcing Online Research

Eyal Peer, Bar-Ilan University, Israel
Sonam Samat, Carnegie Mellon University, USA
Laura Brandimarte, Carnegie Mellon University, USA
Alessandro Acquisti, Carnegie Mellon University, USA

Paper #2: Using Nonnaive Participants Can Reduce Effect Sizes

Jesse Chandler, University of Michigan, PRIME Research, USA
Gabriele Paolacci, Erasmus University Rotterdam, The Netherlands
Eyal Peer, Bar-Ilan University, Israel
Pam Mueller, Princeton University, USA
Kate Ratliff, University of Florida, USA

Paper #3: Research Design Decisions and Their Effects on Task Outcomes: An Illustration Using Free-sorting Experiments

Simon J. Blanchard, Georgetown University, USA
Ishani Banerji, University of Texas at San Antonio, USA

Paper #4: Interactivity and Data Quality in Computer-Based Experiments

Neil Brigden, Miami University, USA

SESSION OVERVIEW

Consumer researchers are increasingly relying on online samples to conduct their empirical investigations. While crowdsourcing markets and online panels (e.g., Amazon Mechanical Turk) have improved the efficiency of behavioral research, they also introduced unique and novel challenges for data quality (for recent reviews see Gosling & Mason 2015; Paolacci & Chandler 2014). In particular, both the recruitment and execution stages of research are inherently less controllable when they happen online than offline (e.g., in a physical laboratory). As a result, several questions arise that directly impact the reliability of online behavioral investigations and ultimately the trustworthiness of research findings: Is there such thing as an "online population", or are some online samples more reliable than others? What are the consequences of research participants self-selecting into research studies as they wish? How can researchers ensure that unsupervised participants will truly engage in the tasks they perform? This session brings together four papers that assess these concerns, and suggest practical solutions for consumer researchers to ensure high data quality in their online investigations.

The first two papers focus on the recruitment stage of online research. Peer and colleagues compare a variety of online panels, finding vast differences in data quality as well as participant demographics and psychometrics. Chandler and colleagues investigate how participating twice in the same study—a phenomenon which previous research documented as prevalent—affects results, and find that including non-naive participants consistently reduces observed effect sizes. The last two papers focus on how features of the research execution can be changed to increase data quality in online research. Blanchard and Banerji comprehensively examine free sorting tasks in online research and find evidence that researcher decisions (e.g., topic, number of items, pre-task video tutorials) substantially affect both participants' experience and the quality of the data they provide. Brigden examines the effect of interactive study elements on participants' attentiveness, and finds that participants' attentiveness is significantly improved in presence of interactive elements. Altogether, the four papers highlight specific data quality concerns for online researchers, test the effectiveness of attempts to address these concerns, and identify important avenues for future research in this area.

The Internet democratized science by lowering the barriers to its consumption and dissemination. Recently, opportunities to conduct research online have allowed for a more democratic *production* of scientific knowledge. However, online data collection also has many undetected pitfalls, that this session uncovers and examines. Prior ACR sessions on crowdsourcing and online research (e.g., Goodman & Paolacci, 2014) demonstrated great interest in the community for a more thorough understanding of the consequences of relocating empirical investigations online. We expect this session to be equally successful, and to contribute substantially – by providing attendees with actionable methodological implications – to improve the practices of online data collection.

Beyond the Turk: An Empirical Comparison of Alternative Platforms for Crowdsourcing Online Research

EXTENDED ABSTRACT

In recent years, a growing number of researchers have been using Amazon Mechanical Turk (MTurk) as an efficient platform for crowdsourcing online human-subjects research. A large body of work has shown MTurk to be a reliable and cost-effective source for high-quality and representative data, for various fields and research purposes (e.g., Buhrmester, Kwang, & Gosling, 2011; Chandler, Mueller, & Paolacci, 2014; Crump, McDonnell, & Gureckis, 2013; Fort, Adda, & Cohen, 2011; Goodman, Cryder, & Cheema, 2013; Litman, Robinson, & Rosenzweig, 2014; Mason & Suri, 2012; Paolacci, & Chandler, 2014; Paolacci, Chandler, & Ipeirotis, 2010; Peer, Vosgerau, & Acquisti, 2013; Rand, 2012; Simcox, & Fiez, 2014; Sprouse, 2011). In parallel, several other alternative platforms now offer similar services, with distinct differences from MTurk: they offer access to new and more naïve populations than MTurk's, and have fewer restrictions on the types of assignments researchers may ask participants to do (see Vakharia & Lease, 2014, for an overview). These alternative services for crowd-sourced research could be highly beneficial for researchers interested in conducting online surveys and experiments, as long as these new sites prove to provide high-quality data. We conducted an empirical investigation of the data quality (in terms of response rates, attention, dishonesty, reliability and replicability) of several alternative online crowdsourcing platforms, and compared those to both MTurk and a university-based online participants pool.

At first, we focused on six services that we found by searching for crowdsourcing websites on the web, which are similar in purpose and general design to MTurk. These services included CrowdFlower, MicroWorkers, RapidWorkers, MiniJobz, ClickWorker and ShortTask. However, we were able to run our study only on the first three sites due to various problems with the other sites (MiniJobz rejected our

study with no explanation or response to our questions; ClickWorker required a high set-up fee of about $840 for 200 participants; and ShortTask failed to process our payment method several times and no support could be reached). In addition to the above three sites and MTurk, we also ran our study on a university-based online participant pool (CBDR) as another comparison group. We aimed to sample 200 participants from each site using a week for sampling. We obtained 200 responses from MTurk and CrowdFlower in less than 2 hours (101.01 and 108.55 responses per hour, respectively). With a considerable difference, CBDR showed the third fastest response rate (1.42 responses per hour), followed by MicroWorkers (1.08 responses per hour) and RapidWorkers (0.63 responses per hour) – from which we could only sample 105 completed responses in a week. Eventually, we obtained a total sample of 890 participants.

Our online study included several parts designed to examine different aspects of data quality. For brevity, we describe these parts here alongside their results. In one part, participants completed several validated questionnaires to examine differences in reliability between the sites: the Internet User Information Privacy Concerns scale (Malhotra, Kim, & Agarwal, 2004), the Need For Cognition scale (Cacioppo, Petty, & Kao, 1984) and the Rosenberg Self-Esteem Scale (Rosenberg, 1979). Overall, we found that MTurk participants showed the highest reliability scores on all three scales, followed by CrowdFlower participants, CBDR and Microworkers, all of whom performed adequately well on all scales (except a somewhat lower score for CrowdFlower participants on the NFC scale). RapidWorkers participants showed high reliability on the IUIPC scale, but very low reliability on the NFC and mediocre reliability on the RSES scales. We used Hakistan & Whalen's (1976) method to compare between independent reliability coefficients and found no statistically significant differences between the samples (using all participants from each sample) for the IUIPC, $\chi^2(4) = 6.63$, $p = .17$, but we did find statistically significant differences for the NFC and the RSES, $\chi^2(4) = 127.07, 75.69$, $p < .01$.

Two attention-check questions (Peer et al., 2014), embedded at different points of the study, checked participants' attention and compliance with written instructions. Whereas only 14% of MTurk participants failed both questions, almost half of the CBDR participants failed them, and the majority of the participants in all other sites failed them as well. Interestingly, CrowdFlower participants (who showed the fastest response rate) had a failure rate of almost 75%.

Another part of the study examined replicability of known findings using tasks from the judgment and decision-making literature (following Chandler et al., 2010): the Asian-disease gain vs. loss framing, the sunk-cost fallacy, and four anchoring-and-adjustment questions. We found the expected effects in both CrowdFlower and MicroWorkers, in levels comparable to MTurk, whereas RapidWorker's results were less than adequate. In another part, which used a die-throwing task, we found no differences in the propensity for dishonest behavior between the different sites.

To conclude, we found that, at the time of writing, both CrowdFlower and MicroWorkers sites, but not the RapidWorkers site, could be potential alternatives to MTurk. Additional examinations revealed that there was a very small overlap between participants from the different sites, and some individual differences between the sites. The most pronounced, and probably most practical, of those was that CrowdFlower's participants included much more Asian participants and non-U.S. citizens than MicroWorkers or MTurk. We believe additional research is required to understand the origins of these differences between the sites, and to further explore other aspects of data quality between these sites in comparison with MTurk, as well as in comparison with other, more traditional samples.

Using Nonnaive Participants Can Reduce Effect Sizes

EXTENDED ABSTRACT

When conducting a study, researchers often assume that participants are naive to the research materials, either because the pool of participants is large (e.g., Internet samples) or because participants' prior exposure to research is limited (e.g., in the case of first year college students). This assumption, however, is often violated. People can belong to a participant pool for several years, and some members are disproportionately likely to be sampled (Chandler, Mueller, & Paolacci, 2014). Moreover, researchers with overlapping interests rely on the same undergraduate subject pools, and participants may easily share information with each other (Edlund et al, 2009). People may also gain knowledge of research materials through college courses or media coverage.

Some research suggests that familiarity with research materials might impact findings. Prior knowledge may increase the likelihood of hypothesis guessing and potentially lead to demand effects (Weber & Cook, 1972). Relatedly, earlier conditions in within-subject experiments inform subsequent conditions, causing effects observed in between-subjects designs to be inflated, attenuated, or reversed (see Charness, Gneezy, & Kuhn, 2012). Recently, researchers have noted that responses to psychological measures correlate with proxies of prior participation in similar experiments, such as memory of prior participation (Greenwald & Nosek, 2001), chronological order of studies themselves (Rand et al., 2014), measures of the total number of completed experiments (Chandler et al., 2014), or naturally varying levels of prior experience with a task (Mason, Suri & Watts, 2014). Although these findings suggest that non-naivety may influence observed effect sizes more generally, this possibility has not been directly tested. To address this gap, we examine how prior exposure to study materials affects responses.

Method

We conducted a two-stage study on Amazon Mechanical Turk (for a review see Paolacci & Chandler, 2014). One thousand participants completed a set of eleven two-condition experiments in Wave 1 (W1), testing phenomena such as anchoring, framing, retrospective gambler's fallacy, etc. (full details about W1 are reported in Klein et al., 2014). In Wave 2 (W2), these participants were invited to participate in a study including the same experiments with the exclusion of two (that were not successful in W1). For each experiment, participants were randomly assigned to the same condition as in W1 or in the alternative condition. Additionally, we manipulated two factors that that should affect whether participants recall previous materials and potentially moderate the effect of non-naivety. *Visual similarity* was manipulated by randomly assigning participants to complete the experiments on the same platform as W1 or on a different, visually distinct platform. *Time Delay* was manipulated by re-contacting participants a few days, about a week, or about a month after W1. This resulted in a 3 (Time Delay) X 2 (Visual Similarity) X 2 (Condition) between-participants design.

Results

We tested the effect of non-naivety on the responses of participants who participated in both W1 and W2 (N = 638; 55% women; M_{age} = 36, SD = 12.8). Overall, effect sizes declined from W1 (weighted d = 0.82) to W2 (weighted d = 0.63) by d = 0.19, a drop of about 25%. Only one effect size increased from W1 to W2 (low vs. high scales task; Schwarz et al.,1985) and all others showed 17% to 83% declines. 9 of the 12 effects (we analyzed the four anchoring tasks separately) exhibited declines, and 5 of these declines were statistically significant.

To examine whether the attenuation of effects was stronger when recalling information from previous participation was easier, we regressed W2 effect sizes on same vs. different Condition, Visual Similarity, Time Delay, and on dummy variables for experiment (accounting for differences in attenuation across experiments). There was a significant effect of Condition, reflecting that participants exposed to different conditions demonstrated greater decline of effects from W1 to W2. There were no main effects or interactions.

After study completion, participants reported for each experiment whether they remembered participating in it. Memory for participation in each experiment depended on Time Delay and ranged between 35% and 80% of participants. We regressed W2 effect sizes on whether they represented those who did or did not report remembering the prior experiment, same vs. different condition dummy, and dummies for the different experiments. There was a significant effect of being assigned to a different condition, and neither memory of prior participation nor its interaction with same vs. different condition were significant. This suggests self-reported memory for prior participation is at best a poor indicator of whether participants will display attenuated effect sizes because of prior participation.

Our findings show that prior exposure to research materials can reduce the effect size of true research findings. Effect sizes decreased of 40% on average, although the reduction was different for different experiments. Future research should examine whether and how some paradigms (e.g., those that require participants to generate numerical estimates) are more susceptible to non-naivety. Effects were particularly attenuated when participants were exposed to alternative conditions of an experiment, highlighting that decreased might be a function of information. However, they were also attenuated among participants exposed to the same condition twice, which might be explained by repeated exposure leading to more elaboration and decreased reliance on intuition (Sherman, 1980). Self-reported participation does not identify all prior participants, or even those who demonstrated a particularly large non-naivety effect. This may be explained if participants quickly forget the source from which the information was learned, but do not forget the information itself (Johnston, Hashtroudi & Lindsay, 1993).

Non-naivety is a serious concern for behavioral researchers that cannot be solved controlling for self-reported previous participation. When directly monitoring prior participation not possible, researchers should design procedures and stimuli that differ from those known to the tested population (Chandler et al., 2014), or increase their sample size to offset the anticipated decrease in power.

Research Design Decisions and Their Effects on Task Outcomes: An Illustration Using Free-sorting Experiments

EXTENDED ABSTRACT

Numerous research areas within psychology and marketing have relied on free-sorting tasks, wherein participants allocate a set of objects into groups of their own choosing to study the natural cognitive processing of information that consumers encounter in their lives (e.g., Blanchard 2011; Ross & Murphy 1999). Unfortunately, there is little systematic empirical research that provides guidance on how researchers should design sorting tasks in order to minimize unwanted consequences such as contaminated process data, and depleted or dissatisfied participants. As different studies have provided differing recommendation, we provide an empirical investigation of sorting task researcher-driven design decisions on a variety of outcomes.

To do so, we created an experimental design that systematically varies the decisions that a researcher may face when adopting a sorting task using a fractional factorial design to test the main effect of each factor (i.e., researcher decisions) on various dependent measures (Collins, Dziak, & Runze, 2009). We then provide guidance as to best practices and potential pitfalls. The factors, along with the final design (involving 36 orthogonal tasks), are presented in a table that can be downloaded here: http://tinyurl.com/blanchardtable

Participants & Procedure

We requested 720 participants (20 per task) from Amazon Mechanical Turk (mTurk) for a "consumer perceptions study." Participants were paid $0.50 and were allocated evenly/sequentially using Qualtrics' quota functions.

Once a participant clicked the survey link on mTurk, each participant was randomly assigned to one of the 36 task configurations via an initial Qualtrics survey whose sole purpose was to randomly assign participants. If the participant was assigned to a task design for which pre-task examples were to be provided (*Using pre-task tutorials; yes/no*), Qualtrics displayed a short video that demonstrated musical instruments being sorted along (adapted from Lickel et al., 2000). Participants then proceeded to the online sorting task interface (Cardsorting.net), which affords researchers the ability to present a varying number of objects to be sorted (*Number of objects; 20, 40, 60*), to customize instructions (*Providing a criteria for the sorts; similarity/dissimilarity*), to use either a single or a multiple cards sorting task (*Type of sorting task*), to require that participants sort all the objects at least once (*Requiring to use all the cards at least once*), and the option to ask participants to label the piles during the task, after the task, or not at all (*If and when to ask for pile labels*). After submitting their sorts, all participants proceeded to the same post-task Qualtrics survey, which contained additional dependent measures.

Results

For our analyses, we use mixed-effect linear models where the task number is a random effect and all researcher decisions are fixed effects. We provide the following recommendations:

Researchers should keep the number of objects manageable, but there is no need to severely constrain the number of objects. Although participants prefer tasks with fewer objects (20), we find no evidence that participants cannot properly follow instructions that require them to sort 40 or even 60 objects. As the number of objects increases, the biggest impact seems to be on completion time, perceptions of the effort required, and the extent to which the task was enjoyable.

Researchers should be mindful of what they are asking participants to sort. The type of objects being sorted has a significant impact. Using "groups types" instead of food objects led to a significant increase in dropout rates, completion time (for an equivalent number of objects), and participant depletion.

There's no harm in allowing participants to sort objects into multiple piles. Even when given the option, participants were assigned objects to multiple piles only when their perceptions dictate it. Allowing participants to do so when it will not negatively impact research goals and may result in less attrition.

If researchers want labels for the piles, they should ask participants to do so after *they have submitted their sorts as complete.* Asking participants to name the piles during the task led to significantly greater dropout rates, a greater number of cards left unused, and higher completion times.

If researchers are trying to increase the frequency at which participants assign objects to multiple piles, they should consider providing pictures along with the objects' names. Doing so allows participants to visualize the objects, and tends to lead to a greater number of objects used more than once in the sorts.

Requiring participants to use all the cards has little effect on task and satisfaction measures. If you suspect that participants will be familiar with the majority of the objects to be sorted, then requiring they sort all objects is more likely to result in fully complete data without any detrimental effects on the sorting process or participants' experience.

Sorting interfaces are sufficiently intuitive. Expansive instructions may not be necessary. We found that participants followed the instructions, and providing them with videos that illustrated the features of the sorting interface (e.g., adding/removing objects from piles, labeling, etc.) did not have much of an impact other than to speed up the sorting process once participants got started.

General Discussion

Free-sorting is a popular data collection methodology. With online panel data and computerized platforms for free-sorting becoming more readily available, it is now possible to conduct these experiments even more easily than before. Nevertheless, researchers have many decisions to make regarding how the task should be presented to the participants and the present research shows that some of these decisions can have important consequences on the task's outcomes, and participants' perceptions of the experience.

In general, we suggest that researchers who have been wary of using sorting tasks thus far consider using online interfaces that provide a great deal of flexibility and feasibility to the researcher, while simultaneously providing participants with a reasonably pleasant and interesting task. Moreover, online interfaces allow rapid data collection thus making it possible to collect large amounts of data cheaply and quickly via online panels such as mTurk. Our findings also suggest that it may be worthwhile to revisit existing finding while varying the researchers' decisions thus introducing theoretically interesting nuances to extant literature.

Interactivity and Data Quality in Computer-Based Experiments

EXTENDED ABSTRACT

Research participants are often not as diligent and engaged as researchers would like. Instructional manipulation checks – IMCs (Oppenheimer, Meyvis, & Davidenko 2009), allow researchers to detect respondents who are not being attentive, but these checks do not address the underlying issue. Having detected inattentive respondents the researcher can discard their responses, likely reducing noise and improving statistical power, but reducing the sample size and possibly skewing the sample in the process. As an alternative, we propose that researchers might improve participant engagement by making studies more engaging, reducing the need to discard data after the fact.

The need to validate new online research pools has led to several studies examining the problem of inattentiveness among research participants. Some have suggested that the problem is comparable in magnitude (Paolacci, Chandler, & Ipeirotis, 2010), or worse among MTurk workers (Goodman, Cryder, & Cheema, 2013), while two recent studies have suggested that it is now less severe among MTurk workers than among other subject pools (Hauser & Schwarz, 2015; Klein et al., 2014).

The higher performance on IMCs among MTurk workers has multiple causes. IMCs are commonly used as a criterion for worker payment on MTurk (Chandler, Mueller, & Paolacci, 2014). Also, MTurkers in general are more experienced with these checks (Peer, Vosgerau, & Acquisti, 2013). Lastly, researchers tend to only include high reputation workers on MTurk, whereas undergraduate student samples tend to be more unfiltered (Hauser & Schwarz, 2015).

While MTurk workers generally perform better than undergraduates on IMCs, there remains room for improvement, with pass rates as low as 26% on novel IMCs (Hauser & Schwarz, 2015). Such low pass rates on these novel IMCs suggest that MTurk workers may be scanning for familiar IMCs rather than diligently reading all instructions.

Studies that are more interactive are, almost by definition, more engaging for participants. They may feature motion, sound, and choices that meaningfully change the participants' experience. These interactive elements were predicted to significantly improve participant attentiveness as measured by IMC pass rates.

Method

To assess the impact of interactivity on participant attentiveness, a meta-analysis was run on a set of 17 prior studies (total N=2384). The analysis compared IMC pass rates between studies that were either interactive or not, across three subject pools (MTurk, undergraduate students, and an in-house online panel). For the purposes of the analysis a study was only classified as interactive if it met two criteria: It had to contain audio or video and it had to feature some consequential choices. All studies included the same IMC, which asked participants to scroll down and click on a button off screen if they were reading the instructions, rather than clicking a "continue" button already visible on screen. For the MTurk studies, participation was limited to workers who had previously completed a minimum of 100 HITs with a minimum 95% approval rating.

Results

The effect on interactive elements on the pass rate for the IMC was assessed separately within each subject pool, as prior research has found dramatic differences in IMC pass rates between different subject pools. The largest difference in IMC pass rates between interactive and non-interactive studies was in the in house online subject pool, where interactive studies had a significantly higher IMC pass rate (87.3% than non-interactive studies (59.5%), χ^2 (1, N = 415) = 229.8, $p < .001$, $\varphi = .27$. The difference was also significant in the undergraduate subject pool, where again the interactive studies had a significantly higher IMC pass rate (94.1%) than non-interactive studies (78.2%), χ^2 (1, N = 1035) = 56.8, $p < .001$, $\varphi = .24$. However, in the MTurk subject pool there was no difference in IMC pass rates between interactive studies (91.2%) and non-interactive studies (93.9%) χ^2 (1, N = 934) = 2.1, $p = .15$.

Discussion

The results indicate that more interactive studies may motivate participants to read instructions more carefully leading to better compliance. Interestingly, the improvement in attention appears to generalize to parts of the study that are not interactive. The IMC, in all studies, occurred outside of the interactive portion of the experiment.

Interactive studies were no better than non-interactive studies on MTurk. However, it may be difficult to improve on the high IMC pass rate this panel generally exhibited for this particular IMC. Further research using a novel IMC or a less restricted pool of MTurk workers may reveal differences in this subject pool as well.

Interactive elements are appealing from the researchers perspective as they increase engagement without necessarily requiring additional payments to participants. This keeps research costs down and can also avoid the issue of creating a productivity mindset, which may be counter to the objective of studying many consumption experiences that are not typically accompanied by this mindset. Interactive computer based experiments may also offer new avenues of examining consumption experience, an area with great potential for significant new discoveries within consumer research (Janiszewski, 2010).

REFERENCES

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*, *6*(1), 3-5.

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

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

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1-8.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, *8*(3), e57410.

Edlund, J. E., Sagarin, B. J., Skowronski, J. J., Johnson, S. J., & Kutter, J. (2009). Whatever happens in the laboratory stays in the laboratory: The prevalence and prevention of participant crosstalk. *Personality and Social Psychology Bulletin*, *35*(5), 635-642.

Fort, K., Adda, G., & Cohen, K. B. (2011). Amazon mechanical turk: Gold mine or coal mine?. *Computational Linguistics*, *37*(2), 413-420.

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.

Goodman, J., & Paolacci, G. (2014), Questioning the Turk: Conducting High Quality Research with Amazon Mechanical Turk. *Workshop held at the Association for Consumer Research North American Conference*; Baltimore, MD.

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology*, 66, 877-902.

Greenwald, A. G., & Nosek, B. A. (2001). Health of the Implicit Association Test at age 3. *Zeitschrift für Experimentelle Psychologie*, *48*(2), 85-93.

Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 1-8.

Janiszewski, C. (2010). The Consumer Experience. In *Advances in Consumer Research* (Vol. 37, pp. 1–9). Duluth, MN: Association for Consumer Research.

Johnson, M. K., Hashtroudi S. & Lindsay S. D. (1993), Source Monitoring, *Psychological Bulletin , 114*, 3–28.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology, 45,* 142-152.

Litman, L., Robinson, J., & Rosenzweig, C. (2014). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, 1-10.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. *Information Systems Research*, *15*(4), 336-355.

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

Mason, W., Suri, S., & Watts, D. J. (2014). Long-run Learning in Games of Cooperation.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867-872.

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

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, *5*(5), 411-419.

Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods*, 1-9.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of theoretical biology*, *299*, 172-179.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T. , Newman, G. E., Wurzbacher, O., Nowak, M. A., Greene, J. D. (2014). Social Heuristics Shape Intuitive Cooperation. *Nature Communications*.

Rosenberg, M. (1979). *Rosenberg self-esteem scale*. New York: Basic Books.

Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 26, pp. 123-162). San Diego, CA: Academic Press.

Schwarz, N., Hippler, H. J., Deutsch, B., & Strack, F. (1985). Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, *49*(3), 388-395.

Sherman, S. J. (1980). On the self-erasing nature of errors of prediction. *Journal of Personality and Social Psychology*, *39*(2), 211.

Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior research methods*, *46*(1), 95-111.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, *43*(1), 155-167.

Vakharia, D. & Lease, M. (2014). Beyond Mechanical Turk: An Analysis of Paid Crowd Work Platforms, *iConference 2014*. Retrieved January 26th 2015 from http://www.ischool.utexas.edu/~ml/papers/donna-iconf15.pdf

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin*, *77*(4), 273.