Journal of Business Research 69 (2016) 3139–3148

Contents lists available at ScienceDirect

Journal of Business Research

# A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples

Scott M. Smith [a,*], Catherine A. Roster [b], Linda L. Golden [c], Gerald S. Albaum [b]

[a] Brigham Young University, United States
[b] University of New Mexico, United States
[c] University of Texas at Austin, United States

ARTICLE INFO

*Article history:*
Received 1 March 2015
Received in revised form 1 June 2015
Accepted 1 October 2015
Available online 25 January 2016

*Keywords:*
Consumer panels
Respondents
Data quality
Mechanical Turk
Internet surveys
Survey research

ABSTRACT

With the exploding use of Internet surveys, research efforts and data quality are increasingly subject to the effects of respondents who do not give the required attention to survey questions and who speed through the survey, or who intentionally cheat with their answers. We investigate respondent integrity and data quality for samples drawn from a "Regular" online panel and from Amazon's MTurk. New metrics for assessing sample integrity and online data quality are introduced. Overall, MTurk respondents in both respondent groups took less time to answer questions. The non-USA MTurk group deviated most from correct answers in attention filter questions and had more duplicate IP addresses. In addition, the results from the three Internet sample sources are substantively different. The choice of an Internet survey sample vendor is critical, as it can impact sample composition, respondent integrity, data quality, data structure and substantive results.

© 2015 Published by Elsevier Inc.

## 1. Introduction and purpose

Online, or Internet-based, surveys are now the predominant delivery method for survey data collection worldwide, eclipsing traditional survey data collection methods used more frequently in the past, such as mail, face-to-face interviews, and landline telephone surveys. According to recent industry statistics compiled by IBIS*World* (Morea, 2014), online research now accounts for 32% of global market research revenues as a percentage of total expenditures.

The widespread adoption of Internet surveys has fueled an increased reliance on online Internet panels as a source of potential respondents for market research—for all forms of business and social science research (Brick, 2011; Callegaro et al., 2014). A whole industry has grown recently and flourished around using online panels as a source of respondents for Internet survey delivery.

Online panels are possible because of technological innovation, most specifically the Internet, and quickly grew in the 1990s, beginning in the USA, and extending to Europe. They filled the need for a way to access respondents, as there was a void left by lists, directories, and RDD telephone sampling techniques that were either non-existent or non-applicable (and no longer useful) in an increasingly mobile and Internet-driven world (Poynter, 2010). As a matter of necessity, modern-day researchers have little choice but to use online panels to address academic and applied issues when relatively large scale primary data and/or diverse samples are needed.

There are now many different forms of online panels (Couper, 2000). We use the term "online panelists" to refer to a pool of individuals who have volunteered to participate in discontinuous consumer surveys via their Internet panel membership. These survey opportunities are emailed to the panelists as "invitations" if they meet the general specifications (and are randomly selected to be invited). The incentive for panel membership and survey participation are various rewards, such as entry into prize drawings, points, or direct monetary payments. We use this definition because it is descriptive of the characteristics of many of the vendors in the Internet respondent marketplace.

Online panels provide convenient access to a large pool of potential respondents at a relatively low cost and with a potentially quick response time. The two main concerns about use of online panels relate to sample integrity and data quality. Public opinion researchers have questioned the inherent validity of extending broad population statistical inferences to data drawn from non-probability samples (Baker et al., 2010; Brick, 2011), as panel membership inherently involves self-selection bias including only people who have at least some Internet access.

Business and social scientist researchers, driven by pragmatic needs, often accept convenience samples as long as these sample sources fit the particular research expectations and are reasonably representative of a defined target market (e.g., Murray, Rugeley, Mitchell, & Mondak,

* Corresponding author at: 1377 S. 1140 East, Orem, UT 84097, United States. Tel.: +1 801 376 1339.
*E-mail address:* smsmith@byu.edu (S.M. Smith).

http://dx.doi.org/10.1016/j.jbusres.2015.12.002
0148-2963/© 2015 Published by Elsevier Inc.

2013). This acceptance has allowed for the growth of online panel data bases, as has the need for respondent access created by changing communication lifestyles (e.g., telephone surveys to reach consumers at home is nearly an impossibility in today's changing world).

There have been a few studies of specific motivations for joining panels (e.g., Brüggen, de Ruyter, & Wetzels, 2005; Brüggen, Wetzels, de Ruyter, & Schillewaert, 2011). Poynter and Compley (2003) report a mix of respondent motives in their study, with incentives being the most cited, followed by curiosity, enjoyment, and wanting to have their views heard. In another study, Comley (2005) used results from an online study of panelist motivation to assign respondents to one of four segments: (1) *opinionated*, wanting to have their views heard and they enjoy surveys; (2) *professionals*, who do many surveys and generally will not respond unless there is an incentive; (3) *incentivized*, who are attracted by incentives, but may respond when there is not one; and (4) *helpers*, who enjoy doing surveys and like being part of the online community.

A concern surrounding Internet surveys is that "professional survey takers" who participate primarily to seek rewards are more likely to engage in inattentive, satisficing, or even fraudulent response behaviors to qualify and receive their incentives (Golden & Brockett, 2009). Not all panel-based respondents participate for the incentive that might be offered. Intrinsic motivators for participating in online panels can include interest, enjoyment, curiosity, helping, giving opinions, incentives, obligation, and need for recognition (Brüggen et al., 2005), as well as commitment and involvement to the sponsor or research community in general (Albaum, Evangelista, & Medina, 1998). However, "professional survey takers do exist" and if present in large numbers, the inclusion of "professional survey takers" can negatively impact overall data quality and, possibly, sample integrity (Callegaro et al., 2014; Hillygus, Jackson, & Young, 2014; Karminska, McCutcheon, & Billet, 2010; Menictas, Wang, & Fine, 2011). We address ways in which these data quality factors can be assessed.

As the Internet survey respondent marketplace spawns new entrants, methods of accessing respondents is expanding beyond the online panel databases. Recently, the pool of potential online respondents has enlarged to include an even cheaper and readily accessible sample source, Amazon's Mechanical Turk (MTurk), which is being more widely used.

The overall objective of this paper is to discuss findings from an Internet survey designed to compare and contrast respondent and data quality and respondent integrity measures from an online panel sample and from a sample drawn through Amazon's MTurk. This paper makes unique contributions to data quality issues in Internet survey research and fills a current void in the literature with a comparison of MTurk respondent data quality with that from a "Regular" online consumer panel.

The first sample source, which we call "Regular," results from a general USA population online panel administration. It was drawn from a reputable commercially maintained Internet survey panel.

The second sample is comprised of crowdsourced MTurk respondents. It was discovered post-hoc that the MTurk sample group diverged geographically from the USA consumer online panel, thus, we split the MTurk sample into two subsamples that diverged geographically, one with USA respondents and one with non-USA respondents. This provided an MTurk USA respondent sample that could be compared to the Regular USA respondent panel holding the geographic component fixed, and secondarily a non-USA respondent MTurk sample group which could be compared to the USA respondent-based MTurk group holding the sample source fixed. In the findings, we compare and contrast differences in respondent characteristics and data quality between the USA and non-USA subgroups from the MTurk sample source and the Regular USA online consumer panel.

MTurk is Amazon's crowdsourcing Internet marketplace of workers created in 2005 to allow for crowdsourced solutions to certain labor intensive activities that could not be done robotically. According to Howe (2006), crowdsourcing can be used to promote a job opportunity—including participating in research studies—outsourced to an undefined group of people in the form of an open call. Employers or "requesters" post tasks, called HITs (Human Intelligence Tasks), on Amazon's website (www.MTurk.com) to recruit anonymous "workers" in exchange for a small monetary wage (called a "reward"), typically in the range of a nickel or dime for a 5 to 10 minute task (cf., Buhrmester, Kwang, and Gosling (2011) for details on using MTurk in social science research).

Consumer researchers have already begun to explore use of MTurk samples as an alternative to college student samples, long criticized for inherent limitations (cf., Peterson and Merunka, 2014). However, researchers have only recently begun to investigate the quality of MTurk data and samples. Buhrmester et al. (2011) report that MTurk samples are demographically similar to standard online panel samples, are more diverse than typical college student samples, and produce data comparable in quality at a significantly lower cost. Paolacci, Chandler, and Ipeirotis (2010) review strengths and potential uses of MTurk samples as research subjects, but express concerns about data quality and sample representativeness in part due to recent demographic shifts in MTurk samples. These researchers, and others (e.g., Kazai, Kamps, & Milic-Frayling, 2012; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010), observe trends indicating that the MTurk population is becoming increasingly international and disproportionately populated by highly educated Indian workers. Our data support this International trend observation for MTurk.

By comparison, Internet panel email address vendors generally (but not always) seek to tightly control and maintain the identity and characteristics of their members, so the panel company itself is one potential agent for quality control of Internet subject populations' data quality. Not all panels are created equal, however. Internet panel firms often provide sample frames with specific characteristics that are allegedly accurate (e.g., a male is a male respondent and a female is a female respondent, not someone switching to qualify for a particular study). Many panels will also limit the number of surveys an individual can do in a specific period of time for sample quality control purposes.

Internet sample acquisition decisions can have a tremendous impact on data quality and ultimately the decisions made by managers based on the results (Lee, Lindquist, & Acito, 1997). It is critical for researchers to thoroughly evaluate and compare Internet sample acquisition sources for potential data quality and respondent integrity differences. Prior research has examined Internet surveys and online panels since these methods emerged and has compared them to the more traditional survey data collection techniques of the past (e.g., Braunsberger, Wybenga, & Gates, 2007; de Rada & Dominquez-Alvarez, 2013; Heerwegh & Loosveldt, 2008). Alternatively, and going beyond, our research focuses on the previously unresearched data quality metrics that can impact data integrity in its comparison of online data sources.

Our multi-group analysis involves replications of the same study with different subsamples and demonstrates the importance of data quality standards to the research process. In the discussion, we offer suggestions for improving Internet survey design and ways to increase confidence in both Internet sample integrity and data quality. We conclude with recommendations for the potential use of MTurk samples in Internet-based survey research going forward.

## 2. Research questions

The central purpose of this paper is to report findings that compare and contrast respondent quality and data quality across two modalities of Internet respondent samples, a USA traditional online panel and an MTurk sample, which is split into two subgroups based on geographic origin of respondents for further comparisons and contrasts. Respondent quality is assessed by examining subgroup members' demographic profile and survey-taking experience (i.e., number of panels respondents belong to and average number of surveys taken per week).

Data quality is assessed using measures to identify response behaviors indicative of respondents whose data is of sufficient low quality that validity of responses is questionable or suspect (e.g., Downs, Holbrook, Sheng, & Cranor, 2010; Golden, Larson, & Smith, 2011; Golden & Smith, 2010; Greszki, Meyer, & Schoen, 2014; Hillygus et al., 2014; Kittur, Chi, & Suh, 2008). One questionable response behavior is "responding too fast," suggesting that respondents did not give sufficient thought to their responses. Response speed behavior is typically measured by analysis of response times, and measurement of responses that reveal inappropriate levels of variance. The latter would include evidence of "straight-lining" (i.e., respondents giving an identical response to all items presented in a grid) or rote keystroke responses, both of which are typically measured by assessing inter-item variance within and across question blocks.

Other impacts on data quality are evidenced by inattentiveness or fraudulent behaviors that can be revealed when respondents provide incorrect responses to questions inserted into the survey flow that require specific responses (e.g., "Answer 'yes' to question 7" as an instruction ignored by the respondent) or respondents who do not have knowledge that would be typically expected for the respondent group (e.g., "President Obama is the first American President" or "The Sun rotates around the Earth"). These threats to data quality are generally measured by correct responses to attention filter or "trap" questions.

We contend that threats to data quality are created by respondents who engage in two distinct but potentially overlapping response styles, which we classify as "speeders" and "cheaters." We define a *speeder* as a respondent who does not thoroughly read the questions and uses minimal cognitive effort to provide answers that satisfy the question (to collect their incentive with as little time spent as possible). We define a *cheater* as a respondent who intentionally answers survey questions dishonestly and in a fashion that maximizes their opportunity for participation and subsequent rewards.

A major aspect of response quality for the study presented in this paper is the extent of "cheaters" and "speeders" in the resulting respondent groups. Our analysis of data quality includes a comparison of differences in data structure and substantive findings among the three web-based sample groups.

Given the objectives of this study, the following research questions ground our analysis:

RQ1: Are there differences in respondents' demographic characteristics and survey-taking experience among the three sample groups (*sample integrity* issues)?

RQ2: Are there differences in response behavior among the three sample groups for the following *data quality* characteristics: (1) speeding, (2) cheating, and (3) data structure and results?

## 3. Methodology

### 3.1. Sample acquisition and composition

The total number of participants in the study were 1543 respondents obtained from two sources: (1) a commercially purchased "general household" panel ("Regular") with sample size of 707, and (2) a sample generated using Mechanical Turk (MTurk), Amazon's crowdsourced Internet marketplace of workers (n = 836). Regular panel respondents were randomly selected from members of a general household panel maintained by a commercial panel company used and purchased through Qualtrics, whose platform software was used for this study. No demographic quotas were established beyond an equal male/female split and USA for the Regular panel, as investigating additional sample characteristics that emerged naturally was an empirical question for our comparative analyses of samples. The rationale for choosing a general population national sample was to introduce variability regarding participants' familiarity with the topics and rating scales which were intentionally designed to be varied and generally applicable.

The MTurk sample was purchased directly from and recruited and incentivized by MTurk. The Regular panel company, as is usual practice, did their recruitment email invitations to its panel members, as researchers only pay for access and never see the actual email addresses which are proprietary to the panel vendor (after data is collected, researchers can see respondents' IP addresses). Both samples were compensated in their usual manner by either MTurk or the panel company and we paid the flat fee required to be able to access both sample data sources for the final sample size requested. The researcher does not generally know what exactly is paid or incentivized to participants, and payments and incentives can vary by panel company (points for product purchases, money, etc.) or sample source.

For both samples there was a general invitation to participate of a specific type relevant to the sample source. There was a general post of the survey participation opportunity in the case of MTurk and, in the case of a Regular online Internet panel, a set of "clients" who signed up to be panel members and who have given their email addresses to be part of Internet surveys from time-to-time for various "rewards"/incentives (Regular panel companies). The Regular panel company would send an email to a subgroup of these clients inviting participation and a survey link to open to do so.

Given the MTurk process, the post resulted in an international sample which was subsequently decomposed into the two separate MTurk subsamples via IP addresses (non-USA and USA MTurk) post hoc. Unlike a Regular panel, participant characteristics cannot be pre-specified to the same extent with MTurk when ordering the sample access (which is a cost saving offered by regular panel companies, eliminating the need for a series of screener questions to focus the sample characteristics, avoiding the purchase of a larger sample frame than might actually be needed).

### 3.2. Survey development and measurement

The research instrument was an online survey developed to focus on collecting varied and general opinions, administered using the Qualtrics Internet survey platform. There were several question blocks in the survey (using randomized A and B versions), each focusing on one or two topics thought to be of general interest to the population. See Table 1 for example questions in the Internet survey for question blocks developed.

Each question block appeared on separate screens in an effort to maintain respondent interest, and to allow close monitoring of response quality within each block (e.g., speed in completing a screen) as the respondent worked their way through the questionnaire. Since the focus of this research is methodological, we developed a wide range of questions so that no specific knowledge would be required and the survey focus would be reasonably topical and relevant to a general population. Using conceptually developed blocks of questions allowed us to be able to validate the results within and across content areas and different methodological questions separately.

The question blocks took the form of a matrix of multiple choice questions to which the respondent was asked to respond using five category symmetric Likert scales (e.g., Strongly Disagree (1)–Strongly Agree (5)) or 5-category general rating scales (Very Unhappy–Very Happy; Exemplary–Needs Improvement; Just like me–Not at all like me; Definitely False–Definitely True). Scale extremes were reversed for some scales and later appropriately coded so that "1" represents the negative/disagree and "5" represents the strongest "agreement".

The A and B versions of the survey instrument for each block were delivered to respondents at random, as programmed into the Qualtrics software. Versions were identical, except that version B reversed either the answer scales or the wording of the actual scale items as an

**Table 1**
Examples of questions asked by block.[a]

| |
|---|
| Block 1: I believe that in the next year (Strongly Agree/Strongly Disagree) |
| The economy will be stronger |
| More people will not have jobs |
| Block 2: How happy are you with the following parts of your world (Very Unhappy/Very Happy): |
| The overall quality of your country |
| The social well-being of your community |
| Block 3: How happy are you with the following (Very Unhappy/Very happy) |
| Your personal overall quality of life |
| Your financial well being |
| Block 4: How much do you agree or disagree with the following statements (Strongly Agree/Strongly Disagree) |
| My relationships with friends bring me happiness |
| My work is frustrating |
| Block 5: How much do you disagree or agree with the following statements (Strongly Disagree/Strongly Agree) |
| Barack Obama has my best interest at heart |
| I do not trust Obama's judgment regarding the economy |
| Block 6: In thinking about the use of Facebook and other social media over the last summer and fall, how well do the following statements describe you (Not at all Like me/Just Like me) |
| I feel closer to my friends through Facebook |
| I sometimes feel addicted to Facebook |
| Block 7: Please indicate your degree of agreement with the following items (Definitely False, Definitely True) |
| I am not always an ethical person |
| I have integrity |

[a] All questions are scaled from 1 to 5, with 1 being the least agreement and 5 being the most agreement.

additional methodological check. In addition, each block of questions contained measures of respondent attentiveness and correctness of responses. These measures included unobtrusive recording of the time the respondent took to complete each screen/page, response variance, and frequency of numeric answer selection. Obtrusive measures of data quality included attention questions that had a fixed answer:

1. "If you live in the U.S. select 'Strongly Agree'" (Block 1),
2. "Please answer 'Very Unhappy'" (Block 2),
3. 'How happy are you receiving a very large bill from the IRS" (Block 3),
4. "The Sun rotates around the Earth" (Block 4), and
5. "Obama was the first American President" (Block 5) and
6. "I have never heard of Facebook" (Block 6).

The six attention questions are more than sufficient for assessing data quality for that dimension. The combination of these obtrusive and unobtrusive measures were used to model respondent data quality and investigate whether or not respondents were attentive to the questions within the block and whether or not they provided data from reasonable cognitive focus (resulting in good quality data from each respondent—respondents have read and processed the questions asked effectively). While validity checks as we knew them in face-to-face and telephone surveys are not relevant to Internet panel data, especially using maintained panels, these types of checks on respondent data quality are increasingly relevant to new Internet survey processes and need to be routinely incorporated into Internet survey research.

Since the research objectives are methodological, there was no need to conduct psychometric scale construction measures. Therefore, there is no methodological focus on psychometric reliability and validity assessments for the scales themselves. The purpose of this research was to gather data quality information and identify potential speeding and/or cheating response behaviors for the respondents in the three sample source groups investigated.

## 4. Findings and discussion

### *4.1. Characteristics across the two MTurk samples and regular USA online panel*

Median annual household income differs widely across the three samples, with the MTurk non-USA group having a much lower level household income reported ($20,000–29,999) than did the MTurk USA ($40,000-$49,999) and the USA Regular panel ($50,000–59,999). And, as mentioned earlier, the MTurk sample yielded non-USA and USA respondents so was divided into two samples for this research.

The division of the MTurk sample into USA and non-USA subsamples was done after data collection using IP addresses, as it was only then that we learned that the MTurk crowdsourced sample was highly non-USA, which was not expected. The Regular panel was ordered to be a nationwide USA sample frame, balanced for male/female, so there were no non-USA members in that panel. The researcher has more control over sample characteristics when the sample frame is ordered from a Regular panel (beyond the use of screener questions), as these commercial Internet panels are often maintained for specificity by a firm to allow for finer sample composition targeting at the time of respondent recruiting. This saves researcher time and costs, allowing a more relevant sample targeting for survey participation solicitation (done by the panel).

As shown in Table 2, the vast majority of the MTurk panel sample respondents in our research were from countries other than the USA. This is in contrast to the findings of Buhrmester et al. (2011) who found generally that USA respondents dominate the MTurk respondent base. Possibly, the MTurk crowdsourcing sample demographics are changing over time and participants are becoming more international. The MTurk incentives may also have more value monetarily outside the USA, relatively (e.g., the lower income of the non-USA MTurk sample). This issue remains to be addressed further in future research, but it is clear that our MTurk sample composition varies dramatically from that reported by Buhrmester et al. (2011) and that the conclusions of that research regarding the MTurk sample composition may no longer describe the dynamic Internet survey environment (for MTurk/crowdsourcing).

The sample demographics characteristics shown in Table 2 demonstrate consistent significant differences between the panels in level of education, family structure, ethnicity, mean number of panels belonged to, and average number of surveys completed per week suggesting that inferences about substantive characteristics of the samples may depend on the survey sourcing (where sample is obtained). Male/female distributions are about the same, as would be expected as this was a specified sample characteristic when samples were ordered. Overall, differences across the samples were statistically significant for all demographic variables, but were most pronounced between the MTurk non-USA group and each of the United States groups individually.

It is important to emphasize that the demographic differences evidenced in this research among the sample sources may be more important than appears at the surface. For example, non-USA crowdsourced respondents belong to an average of 2.95 "panels" as contrasted to the 0.82 panels of the MTurk USA group. When the researcher is concerned about "professional survey takers," who are primarily interested in receiving the survey completion incentives, there may be some demographic groups more inclined to "make a living" filling out surveys (e.g., unemployed outside the home, students, and lower income individuals or countries). In addition, for MTurk samples, the non-USA MTurk respondents report answering 10.25 surveys per week, on average. However, the USA MTurk sample reports responding to an average of 16.75 surveys a week, which is a very high number and may signal sample integrity and/or data quality issues.

Again, from a potential subject motivation and Internet data quality perspective, reputable well-maintained commercial panels provide a level of control over data quality for the researcher. Some Internet

**Table 2**
Sample characteristics for each sample source and geographics.

| Characteristic | Sample source and geographics | | | |
|---|---|---|---|---|
| | MTurk USA (n = 161) | MTurk non-USA (n = 675) | Regular USA panel (n = 707) | Difference tests for significance |
| Gender | | | | $\chi^2 = 3.82$, $p < .15$[a] |
| Female | 56.4% | 49.9% | 50.8% | 3.04, $p < .10$[b] |
| Male | 41.6 | 50.1 | 49.2 | 3.72, $p < .10$[c] |
| | 100.0% | 100.0% | 100.0% | .10, $p > .15$[d] |
| Education | | | | $\chi^2 = 49.43$, $p < .001$[a] |
| High school or less | 12.5% | 6.5% | 8.5% | 17.77, $p < .01$[b] |
| Some college | 39.8 | 24.6 | 31.4 | 44.62, $p < .01$[c] |
| College graduate | 31.7 | 38.1 | 34.7 | 17.97, $p < .01$[d] |
| Graduate degree | 16.1 | 30.8 | 25.5 | |
| | 100.0% | 100.0% | 100.0% | |
| Family structure | | | | $\chi^2 = 62.20$, $p < .001$[a] |
| Married couple | 48.4% | 62.7% | 63.2% | 15.09, $p < .01$[b] |
| Female householder | 22.4 | 10.1 | 18.7 | 25.42, $p < .001$[c] |
| Male householder | 20.5 | 17.6 | 14.0 | 45.68, $p < .001$[d] |
| Unrelated sub-families | 1.2 | 4.4 | 0.4 | |
| Unrelated individuals | 7.5 | 5.2 | 3.7 | |
| | 100.0% | 100.0% | 100.0% | |
| Race/ethnic | | | | $\chi^2 = 599.11$, $p < .001$[a] |
| Caucasian | 78.9% | 25.8% | 88.8% | 11.45, $p < .01$[b] |
| Black | 6.2 | 5.8 | 3.3 | 159.10, $p < .001$[c] |
| Hispanic | 7.5 | 1.6 | 3.0 | 563.61, $p < .001$[d] |
| Asia/India/Pacific | 9.9 | 62.1 | 4.7 | |
| Other | 3.1 | 4.7 | 0.2 | |
| | 100.0% | 100.0% | 100.0% | |
| Mean Number of Panels | 0.82 | 2.95 | 1.16 | $F = 39.05$, $p < .01$ |
| Average Number of Surveys Taken Per Week | 16.73 | 10.25 | 3.18 | $F = 62.20$, $p < .001$ |

[a] All panels.
[b] Regular USA to MTurk USA.
[c] MTurk USA to MTurk non-USA.
[d] Regular USA to MTurk non-USA.

panel firms place restrictions on the number of surveys a person may take via that panel per week. The number of surveys taken per week by the Regular panel respondents was, on average, 3.18, which is much lower than either of the two MTurk samples. This may reflect the panel company's controls over respondents' survey taking frequency (via limiting the number of email invitations sent to an individual) and provides a positive data quality signal.

Our research goal was to develop some basic sample descriptors, but the results do suggest that there is more to be researched here regarding sample composition and resultant data quality issues among these samples. Our study does not seek to develop an in-depth explication of demographic and sample characteristics for MTurk and Regular panel members—our focus is data quality issues. Paolacci et al. (2010) present more information on the demographics of MTurk respondents in general, and again, these data may beg for updating to investigate sample compositions and sample source effects. Studies investigating the demographics of MTurk respondents should be split into USA and nonUSA for future analyses, as our research suggests that country differences may yield very different sample subgroups.

Table 3A shows the average responses and standard deviations across the three samples. An important point emerging from Table 3A is that different samples yield different results (possibly due to demographic differences, among other things). There are statistically significant differences in both means and variables across question blocks. Thus, the choice of an Internet sample supplier is critical to data quality and, potentially, to sustentative results and conclusions.

Our purpose was not to provide substantive results for survey topic areas, nor to psychometrically develop scales, and those additional analyses are not needed to see that the exact same questionnaire, administered in exactly the same Internet delivery format at the same time via Qualtrics platform software, may yield different results for samples from different vendors (and, with an overlay effect of country). Internet sample

**Table 3A**
Descriptive statistics for summary results by question block.[a]

| Question block | Number of questions[b,c] | Mean (S.D.) for each geographic and sample | | |
|---|---|---|---|---|
| | | USA Regular panel (n = 707) | USA MTurk (n = 161) | MTurk non-USA (n = 675) |
| 1 Economy, inflation, taxes | 12 | 35.34(4.83) | 36.01(5.33) | 36.78(7.72) |
| 2 Happiness with community and world | 6 | 16.66(4.21) | 16.80(4.41) | 20.28(4.44) |
| 3 Quality of personal life | 14 | 51.13(9.50) | 48.10(9.57) | 50.91(8.94) |
| 4 Personal Concerns and happiness | 11 | 32.54(4.43) | 33.52(4.90) | 32.80(7.05) |
| 5 Evaluation of President Obama | 12 | 39.98(7.61) | 36.80(6.48) | 50.91(8.94) |
| 6 Use of Facebook | 14 | 19.08(8.07) | 22.36(9.00) | 36.61(13.17) |
| 7 Ethicality and honesty | 8 | 24.03(2.08) | 24.01(2.77) | 24.16(4.21) |

[a] All pairwise comparisons of equality of the variances within rows are significant at $p < .01$.
[b] Attention filter questions excluded.
[c] All questions are scaled from 1 to 5.

**Table 3B**
Results of pairwise two sample t-test (with unequal variances) of equality of the mean responses from Table 3A by Internet sample acquisition type, geographical classification, and question block scores.

| | USA Regular panel vs MTurk USA | USA Regular panel vs MTurk non-USA | MTurk USA vs MTurk non-USA |
|---|---|---|---|
| Question Block 1 | −1.46 | −4.14[a] | −1.50 |
| Question Block 2 | −0.37 | −15.54[a] | −8.99[a] |
| Question Block 3 | 3.63[a] | 0.44 | −3.39[a] |
| Question Block 4 | −2.33[a] | −0.82 | 1.53 |
| Question Block 5 | 5.43[a] | −24.42[a] | −22.91[a] |
| Question Block 6 | −4.25[a] | −29.67[a] | −16.35[a] |
| Question Block 7 | 0.08 | −0.67 | −0.55 |

[a] Statistically significant $p < .01$.

source effects on data and data quality are worthy of further research, irrespective of the substantive topic area to be investigated.

As shown in Table 3B, there are statistically significant differences in mean responses across all pairwise comparisons of sample sources and question blocks (content area). This suggests that the inferences that can be drawn from a survey may, again, be heavily influenced by the sample source.

For every question block, except Block 7 (rows in Table 3B), there was some pairwise significant survey source difference. Additionally, for every pairwise source comparison (columns in Table 3B), there were blocks of questions that yielded statistically significant answers depending on what source was used. Thus, Internet survey results may be constrained significantly by the sample source.

### 4.2. Speeding through a survey can signal a lack of attention to questions

A primary indicator of "Speeding" through a survey is the obvious "time to completion". This measure indicates that the respondent does not take the time to thoroughly read a question and, therefore, does not take the time to give a thoughtful answer to a question. As previously mentioned, questions on a particular topic were asked in blocks that ranged in number of questions from 7 to 15. The average time to complete each question block (in seconds) for the three sample groups is shown in Table 4. There were significant differences ($p < .001$) for six of the seven blocks of questions.

The USA Regular panel members took more time to complete their responses than either of the two MTurk samples. This suggests that the MTurk respondents did not read the questions as thoroughly and were, in fact, speeding—potentially yielding lower quality data. There was a significant difference at $p < .03$ for questions regarding use of Facebook and a non-significant ($p < .18$) difference in responses about happiness with community and world. As shown in Table 4, for both of these blocks of questions the USA Regular panel took the most time in responding to the set of questions (albeit it not excessive).

### 4.3. Cheating is detrimental to validity and data quality

Imbedded in the questionnaire was a set of questions that were used as "Attention Filters" designed to measure if respondents are directing appropriate attention to the questions in the Internet survey (so as to actually read them and process for a focused answer)—as opposed to "cheating" by not reading questions. The attention filter questions ranged across multiple cognitive tasks and also included simple directive tasks (Select "Strongly Agree"; Select "Very Unhappy"), complex evaluative questions ("How happy are you with receiving a very large bill from the IRS"), and higher level memory and logic questions ("The Sun rotates around the Earth").

There was a range of difficulty in the attention filters and the more difficult attention filters required closer reading and directed attention to the question. If the respondent took the time to read the question, the response was clear. If a respondent sped through the questionnaire without reading the actual questions, he/she would be likely to incorrectly answer these "attention filter" questions. Thus, it is clear that cheating and speeding are inter-connected (cheating could result in speeding) and both affect data quality, with cheating having the effect of not answering honestly. Someone could speed through a survey and still be reading and answering honestly, although not thoughtfully. Cheaters may never even read the question.

Cheating and attention filter question results are shown in Table 5. Responses to all attention filter questions differed significantly ($p < .001$) across the three sample groups indicating differences in attentiveness to survey instructions (with implications for data quality). The response pattern is similar for the two USA respondent groups, but the non-USA MTurk group deviated greatly. Moreover, the non-USA MTurk sample had the least percent of correct responses for all questions indicating that this group of respondents paid the least amount of attention to the questions (and thus would be expected to furnish the lowest quality of data).

Since the MTurk posted crowdsourcing survey opportunity resulted in a substantially elevated percentage of non-USA respondents (675 of 836 MTurk respondents were non-USA), the lack of attentiveness in the non-USA MTurk sample draws into question the quality of MTurk generated data that has not been split into USA and non-USA subsamples. If the respondents are not paying attention to filter questions enough to answer the clearly stated response directive, how much can you trust this sample's responses on other more substantive questions?

Another form of "cheating" can occur as "straightlining" or "Christmas treeing" or a general pattern of random results that may evidence a lack of

**Table 4**
Average time respondents took to answer by question block in seconds.

| Question block | Number of questions | Geographics and sample source | | | F | p |
|---|---|---|---|---|---|---|
| | | USA panel Regular (n = 707) | USA MTurk (n = 161) | MTurk non-USA (n = 675) | | |
| 1 Economy, inflation, taxes | 13 | 76.65 | 52.04 | 64.70 | 36.10 | <.001 |
| 2 Happiness with community and world | 7 | 32.99 | 21.47 | 27.58 | 1.75 | <.18 |
| 3 Quality of personal life | 15 | 72.33 | 50.52 | 49.32 | 16.26 | <.001 |
| 4 Personal concerns and happiness | 12 | 82.72 | 46.60 | 51.27 | 12.57 | <.001 |
| 5 Evaluation of President Obama | 13 | 122.33 | 74.08 | 94.13 | 10.13 | <.001 |
| 6 Use of Facebook | 15 | 57.70 | 44.06 | 56.82 | 3.69 | <.03 |
| 7 Ethicality and honesty | 8 | 51.10 | 38.66 | 36.23 | 14.49 | <.001 |

**Table 5**
Responses to attention questions as an indicator of speeding.

| Attention question[a] | Correct response[a] | Mean values by sample and geographics[b] | | | F | p |
|---|---|---|---|---|---|---|
| | | USA Regular (n = 707) | USA MTurk (n = 161) | MTurk non-USA (n = 675) | | |
| If you live in the U.S. answer 'Strongly Agree' | 5 | 4.72 (90.1%) | 4.86 (93.2%) | 3.56 (15.6%) | 190.57 | <.001 |
| Please answer 'Very Unhappy' | 1 | 1.30 (86.3%) | 1.18 (90.1%) | 1.78 (63.7%) | 45.53 | <.001 |
| How happy are you receiving a very large bill from the IRS (Very Unhappy, Very Happy) | 1 | 1.59 (84.2%) | 1.43 (90.0%) | 2.66 (44.5%) | 228.38 | <.001 |
| The Sun rotates around the Earth (Strongly Disagree, Strongly Agree) | 1 | 2.10 (70.0%) | 2.02 (71.4%) | 2.47 (55.0%) | 11.50 | <.001 |
| Obama was the first American president (Strongly Disagree, Strongly Agree) | 1 | 1.18 (94.7%) | 1.16 (95.0%) | 1.84 (75.3%) | 96.12 | <.001 |
| I have never heard of Facebook (Not at all like me, A lot like me) | 1 | 1.34 (83.2%) | 1.16 (88.2%) | 1.95 (58.5%) | 73.11 | <.001 |

[a] All questions are scaled from 1 to 5, with 1 representing most disagreement.
[b] In parenthesis is percent giving the correct response.

reading of the survey questions. Table 6 reports one measure: Variance in response patterns within question blocks by sample source.

The measurement of dispersion across sample source types was calculated by computing the mean variance within question blocks and normalizing by the standard deviation of the variances within question blocks in order to calculate a z-score for the particular question block by sample source. This normalized z-statistic can then be compared across sample sources to see if there is more straightlining or random response patterns in one sample source versus another.

The results as shown in Table 6 demonstrate that the USA Regular panel differs significantly from the non-USA MTurk respondents in six of the seven question blocks with respect to this data quality measure. Likewise, in four of the seven question blocks the USA MTurk and the non-USA MTurk samples differed significantly from each other. Again, this shows the importance of sample source for data quality, with the MTurk samples evidencing lower data quality, in general. And, the non-USA MTurk has the worst data quality via this measure.

Another useful measure of data quality is test–retest reliability: What was the consistency in answering the same question presented twice within a block. Block 3 questions were duplicated and the resultant correlations of these repeat questions across respondents are shown in Table 7.

All correlations were statistically different from zero. However, consistent with the previous data quality results, the lowest correlations among duplicate questions were for the non-USA MTurk respondents, ranging from .58 to .67. By contrast, the highest correlations were for the USA Regular panel, ranging from .86 to .92. This shows that the non-USA MTurk respondents were more likely to not be attentive to the questions being asked, possibility resulting in lack of reliability causing bad data quality. They were more likely to answer the same question differently than either of the other two sample sources. The responses are less "trustworthy," as it is impossible to know which of the two responses to believe.

An obvious way of cheating that is very detrimental to sample integrity and date quality is taking the survey more than once (to possibly receive multiple incentives). One conservative way of checking into this is to search for duplicate IP addresses in the final sample. An ancillary analysis of IP address duplicates in our data revealed that approximately 11% of the non-USA MTurk sample contained duplicates, as opposed to 1.86% of the MTurk USA sample and 0% of the USA Regular panel.

These IP address duplicates are strong evidence that there may be members of the sample who include themselves in the survey more than once (thus collecting multiple incentives). Our results suggest that this is possibly a higher risk with MTurk (especially non-USA) than with managed commercially maintained Internet panels. The commercial firms have a high self-interest in maintaining non-duplicate sample integrity.

### 4.4. Data structure

Turning now to the substance of the survey, two blocks of questions of general interest and knowledge were selected for analysis across all

**Table 6**
Variance in response patterns within question block by sample source.

| Question block | Number of questions | Mean variance (z-scores) by sample[a] | | | F | p |
|---|---|---|---|---|---|---|
| | | USA panel Regular (n = 707) | USA MTurk (n = 161) | MTurk non-USA (n = 675) | | |
| 1 Economy, inflation, taxes | 13 | −.1407[3] | −.0681 | .0253[1] | 5.90 | <.004 |
| 2 Happiness with community and world | 7 | −.2098[3] | −.1896[3] | .3203[1,2] | 58.06 | <.001 |
| 3 Quality of personal life | 15 | −.2098[2,3] | .0033[1] | −.0427[1] | 8.88 | <.001 |
| 4 Personal concerns and happiness | 12 | −.0961 | −.0203 | −.0584 | 5.92 | <.001 |
| 5 Evaluation of President Obama | 13 | .1685[2,3] | −.0331[1,3] | −.4393[1,2] | 79.47 | <.001 |
| 6 Use of Facebook | 15 | .4256[3] | .4599[3] | 1.3123[1,2] | 207.06 | <.001 |
| 7 Ethicality and honesty | 8 | .3110[3] | .339[3] | .4880[1,2] | 61.59 | <.001 |

Exponent numbers indicate that this sample is different from the other sample(s), for example:

[1] = Differs from USA Regular panel (sample 1).
[2] = Differs from USA MTurk (sample 2).
[3] = Differs from MTurk non-USA (sample 3).
[2,3] = USA Regular differs from samples 2 and 3.

[a] Follow-up tests used Bonferroni correction and are tested at the .05 level.

**Table 7**
Results of bivariate correlations between duplicate items in Block 3.

| Question | r coefficient[a] by sample source | | |
|---|---|---|---|
| | USA panel Regular (n = 707) | USA MTurk (n = 161) | Non-USA MTurk (n = 675) |
| 1 Quality of life family | .858 | .804 | .579 |
| 2 Personal quality of life | .884 | .828 | .619 |
| 3 Financial well being | .923 | .881 | .667 |
| 4 Personal life satisfaction | .881 | .785 | .674 |
| 5 Safety of neighborhood | .911 | .920 | .638 |
| 6 Personal health and wellness | .888 | .871 | .641 |
| 7 Relationships with others | .876 | .784 | .640 |

[a] All correlations significantly different from zero at $p < .001$.

respondents. The attention filter item in each block was excluded from this analysis. The first block included seven distinct question items regarding how happy respondents were about the quality of their personal life. Table 8 reports a comparison of the summated scores for the three groups of respondents.

As shown in Table 8, there is a statistically significant difference between the three sample sources. However, looking at the overall "substance" results, there is little numerical difference among the three groups in mean summated scores for the two topics analyzed. This is consistent with some previous research (Buhrmester et al., 2011; Gosling, Vazire, Srivastava, & John, 2004).

Our focus here is not on conceptually analyzing the life perceptions of the three sample sources, but rather to note that there are differences in substantive results. The agreement with statements about relationships with friends did not evidence differences in responses across sample sources.

A measure of data structure can be formed by conducting an exploratory factor analysis. This was done for two question blocks, happiness and agreement, as shown in Table 9. Only items from these two question blocks were submitted to factor analysis.

As Table 9 shows, a single factor emerged for the happiness question block for all three samples individually and combined, with an explained variance ranging from 51.77% to 60.21%. The agreement block of survey items resulted in a more complex structure.

Four factors emerged for the agreement block for the USA MTurk sample, for the USA Regular panel, and for the three samples combined. However, a three factor solution resulted for the non-USA MTurk sample.

Thus, the essential dimensionality in the data (underlying data structure) for the non-USA MTurk sample is different from the other two samples on the agreement block. The non-USA MTurk sample also had the lowest variance explained by the factors (58%). This continues a pattern of differences for the non-USA MTurk sample when compared to the MTurk USA sample and the Regular panel data.

Delving deeper into the different data structures for the agreement scale, Table 10 demonstrates that different questions load on different factors across the three samples. Again, this indicates that the data structure can be impacted by the Internet sample source and geographics.

## 5. Conclusions and implications

The objective of this study was to compare MTurk sample integrity and data quality to a regular USA online Internet panel sample with respect to speeding, cheating, underlying data structure, test–retest reliability, and other measures. Three samples are compared: a Regular sample of USA Internet panel members, a USA MTurk sample and a non-USA MTurk sample. In general, the lowest sample integrity and data quality resulted for the non-USA MTurk sample. When an MTurk sample is used, the researcher is advised to specify sample characteristics through carefully developed screener questions, as MTurk is crowdsourcing and, unlike a commercially maintained panel, the researcher cannot pre-specify sample characteristics.

Our results suggest that MTurk samples may be dominated by non-USA respondents, which may result in different sample characteristics, response patterns and data quality. This in turn can impact the substantive results and conclusions drawn from the research. Trading off cost (MTurk has an ease and cost advantage), the research must make an informed choice of an Internet online sample source.

While previous research has focused on comparison of MTurk to Internet users or to other survey techniques in terms of demographics (cf., Buhrmester et al., 2011; Ipeirotis, 2009; Paolacci et al., 2010), we have examined the behavioral aspects of survey taking that can effect survey response quality (as opposed to respondent demographics). Our measures included cheaters identified through duplicate IP addresses (taking the survey more than once) to other measures of speeding and cheating, such as the extent to which the respondent is racing through the survey without thoroughly reading the questions or giving patterned responses (e.g., answering all 1's and missing questions related to filter questions), among other things.

Internet surveys are a relatively new innovation in data collection and are here to stay, as they are the currently a very logical choice for accessing respondents. Yet, this study is the first to examine a series of Internet survey data quality metrics. It is clear from our results that there are differences in data quality and resultant data structure by sample source. The metrics developed here are important ones and need to be developed further in future research. As Internet survey methodology evolves, so should our metrics to measure sample integrity and data quality, if we are to have confidence in our survey results.

**Table 8**
Analysis of variance results for happiness and agreement questions.

| Topic | Number of items | Mean summative scores for each sample group | | | F | p |
|---|---|---|---|---|---|---|
| | | Regular USA (n = 707) | MTurk USA (n = 161) | MTurk non-USA (n = 675) | | |
| Happiness about quality of personal life | 7 | 25.54 | 24.06 | 25.38 | 6.556 | <.001 |
| Agreement with statements about relationships with friends | 11 | 32.54 | 33.52 | 32.80 | 1.943 | < .145 |

**Table 9**
Exploratory factor analysis for happiness and agreement question blocks.

| Sample source by question block | Number of items | Number of factors | Percent variance |
|---|---|---|---|
| Happiness | 7 | | |
| Total sample | | 1 | 54.937 |
| USA Regular panel | | 1 | 60.211 |
| MTurk USA | | 1 | 51.890 |
| MTurk non-USA | | 1 | 51.766 |
| Agreement | 11 | | |
| Total sample | | 4 | 61.913 |
| USA Regular panel | | 4 | 60.588 |
| MTurk USA | | 4 | 59.373 |
| MTurk non-USA | | 3 | 58.053 |

In conclusion, no sample in our data provided error-free data quality, although both USA samples performed better than the MTurk non-USA sample. Identifying respondents who are "suspect" and provide low quality data is an important part of the research process for assuring data quality standards. This study provides further evidence that significant differences in data quality occur among online samples, and at the individual respondent level, and that both are critical considerations for assessing and assuring data quality.

### 5.1. Limitations and directions for future research

No research is without its limitations as there are always trade-offs of time and money, at the least. This study incorporated two sample sources, initially, a respected "regular" consumer USA panel maintained by a commercial firm in that business and an increasingly used crowdsourced online survey sample, Amazon's MTurk. Unexpectedly, our MTurk sample was more non-USA than USA, so in the end, we were able to divide the MTurk sample into these two groups for comparison. This division was not planned, but future research should be conducted into a wider range of online panel samples (not only one) and also investigate (by initial design) USA and non-USA results for data quality measures.

Our research did not focus on substantive issues, yet, our broad scales did show differences in results (happiness, for example). Future research should seek to uncover sample source data quality and integrity differences on substantive issues for which substantive reliability and validity are the focus. It would be interesting to focus on some "known information" (where we know what the answers "should be" for substance—beyond our attention focus here) and see how close each sample base comes to "truth" (as it known to be). The design would be "tricky" but not impossible and would shed light on validity of responses in a very important way. This goes beyond basic data quality.

Likewise, future research should go beyond our research and investigate the motivations that different panel source members have in both obtrusive and unobtrusive ways (direct and indirect questions). These results may have clear linkages to the underlying data quality and sample integrity. Motivation possibilities can range widely: interest, money, keeping frequent flier miles active when there are no "butt in the seat" miles being flown so miles are about to expire (always available opportunity on some airlines), keeping up with new products, something to do to alleviate boredom, etc. The nonfinancial reasons may link to socioeconomic status (e.g., there exist high education and income people who do keep frequent flier miles from expiring via regular survey panel opportunities that can often be accessed via the airline's website).

Internet survey research will continue to be an increasingly important data collection method. There are few barriers to entry to new firms who wish to rent email addresses for use by researchers. It is critical to evaluate the firm from which commercial samples are being drawn and, also, to evaluate crowdsourcing sources.

It can be anticipated that crowdsourcing as a technique for developing samples is also here to stay. Anyone with an Internet connection can use crowdsourced Internet survey respondents (and they do), just as anyone can blog. Crowdsourcing of all types from the individually generated crowdsourced sample to the Amazon MTurk crowdsourced sample has an important role to play in research—it is quick, easy and accesses respondents with lower cost, generally. Future research needs to investigate the less visible costs to data quality and sample integrity and also study how to use these sources and enhance the data quality probabilities.

Crowdsourced samples are useful. We need to better understand their strengths and weaknesses and how to address the weaknesses so as to improve the quality of this new developing research tool. And, although the "regular" commercial panel itself has the advantage (over crowdsourcing) of being a validity check on the sample integrity for the researcher (who is paying for a particular set of characteristics behind the targeted email addresses), there will be data quality issues in any survey.

Thus, we need to understand better all forms of Internet sample sourcing, as this is a newly emerging technological tool unlike any other in the past and is often used in the form of "high tech, low touch" unless a hybrid of personal interview and Internet is used which is very costly. Hybrids are a very infrequently used method but offer great opportunity, albeit it costly, for conducting surveys that need very tight data quality controls. All sample source forms and mixes/combinations need to be studied for data quality and sample integrity. We predict that this marketplace for survey sample sales will continue developing quickly with new entrants emerging (including

**Table 10**
Factor number for agreement scale items.

| Scale item | Sample and geographical source | | |
|---|---|---|---|
| | Regular USA panel | MTurk USA | MTurk non-USA |
| I worry frequently about my financial situation | 2 | 1 | 2 |
| I have positive relationships with my family members | 1 | 2 | 1 |
| My relationships with friends bring me happiness | 1 | 2 | 1 |
| My health is something that I often worry about | 2 | 4 | 2 |
| My spiritual beliefs are a positive guiding force to me | 1 | 4 | 1 |
| I feel my voice is heard in national decisions that affect me | 4 | 4 | 3 |
| I engage in hobbies and pastimes I enjoy | 3 | 3 | 1 |
| My work is not frustrating | 3 | 3 | 3 |
| At this time, I'm generally unhappy with my life | 2 | 1 | 2 |
| Most benefits from by daily efforts will occur in the distant future | 3 | 2 and 3 | 1 |
| I'm pessimistic about the future | 2 | 1 | 2 |

more commercially available help with crowdsourcing) and the researcher needs to be caveat emptor.

## References


Albaum, G., Evangelista, F., & Medina, N. (1998). Role of response behavior theory in survey research: A cross-national study. *Journal of Business Research*, *42*(2), 115–125.

Baker, R., Blumberg, S. J., Brick, M. J., Couper, M. P., Courtright, M., Dennis, J. M., et al. (2010). Research Synthesis: AAPOR report on online panels. AAPOR executive council by a task force. AAPOR Standards Committee. *Public Opinion Quarterly*, *74*(4), 711–781.

Braunsberger, K., Wybenga, H., & Gates, R. (2007). A comparison of reliability between telephone and web-based surveys. *Journal of Business Research*, *60*, 758–764.

Brick, M. J. (2011). The future of survey sampling. *Public Opinion Quarterly*, *75*(5), 872–888.

Brüggen, E., de Ruyter, K., & Wetzels, M. (2005, July). What motivates respondents to participate in online panels? *Paper presented at the World Marketing Congress*. Muenster, Germany: Academy of Marketing Science.

Brüggen, E., Wetzels, M., de Ruyter, K., & Schillewaert, N. (2011). Individual differences in motivation to participate in online panels: The effect on response rate and response quality perceptions. *International Journal of Market Research*, *53*(3), 369–398.

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

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., & Lavrakas, P. J. (Eds.). (2014). *Online panel research: A data quality perspective* (1st ed.). Chichester, West Sussex: John Wiley & Sons, Inc.

Comley, P. (2005). Understanding the online panelist. *Worldwide panel research: Developments and progress* (pp. 409–424). Amsterdam: ESOMAR.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly*, *64*(4), 464–494.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). Are your participants gaming the system? Screening Mechanical Turk workers. *CHI 2010: 1001 Users* (pp. 2399–2402).

Golden, L. L., & Brockett, P. L. (2009). Trials, tribulations and trust: Addressing issues in Internet surveys. *Presented at the Academy of Marketing Science World Congress*. Norway, July: Oslo.

Golden, L. L., & Smith, S. (2010, May). Data quality evidence for Internet survey use in intellectual property law. *Presented at the Academy of Marketing Science 2010 Annual Conference*. Oregon: Portland.

Golden, L. L., Larson, J., & Smith, S. (2011, July). Data quality and sample integrity in Internet research. *Presented at The 15th Biennial World Marketing Conference*. Reims, France: Academy of Marketing Science.

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*, *5993–104*.

Greszki, R., Meyer, M., & Schoen, H. (2014). The impact of speeding on data quality innonprobability and freshly recruited probability-based online panels. In Callegaro (Eds.), *Online panel research: A data quality perspective* (pp. 238–262) (1st. ed.). Chichester, West Sussex: John Wiley & Sons, Inc.

Heerwegh, D., & Loosveldt, G. (2008). Face-to-face versus web surveys in a high-Internet coverage population. *Public Opinion Quarterly*, *72*(5), 836–846.

Hillygus, D. S., Jackson, N., & Young, M. (2014). Professional respondents innonprobability online panels. In Callegaro (Eds.), *Online panel research: A data quality perspective* (pp. 219–237) (1st ed.). Chichester, West Sussex: John Wiley & Sons, Inc.

Howe, J. (2006). Crowdsourcing: A definition. *Crowdsourcing: Tracking the rise of the amateur. 2, . Weblog.*

Ipeirotis, P. (2009). Turker demographics vs. Internet demographics. http://www.behind-the-enemy-lines.com/2009/03/turker-demographics-vs-internet.html Accessed February 25, 2014

Karminska, O., McCutcheon, A. L., & Billet, J. (2010). Satisficing among reluctant respondents in a cross-national context. *Public Opinion Quarterly*, *74*, 956–984.

Kazai, G., Kamps, J., & Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. *CIKM 2012: Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (pp. 2583–2586).

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proc. CHI 2008* (pp. 453–456). ACM Press.

Lee, H., Lindquist, J. D., & Acito, F. (1997). Managers' evaluation of research design and its impact on the use of research: An experimental approach. *Journal of Business Research*, *39*, 231–240.

Menictas, C., Wang, P., & Fine, B. (2011). Assessing flat-lining response style bias in online research. *Australasian Journal of Market & Social Research*, *19*(2), 34–44.

Morea, S. (2014). Market research in the US. *IBIS World Industry Report 54191* April.

Murray, G. R., Rugeley, C. R., Mitchell, D., & Mondak, J. J. (2013). Convenient yet not a convenience sample: Jury pools as experimental subject pools. *Social Science Research*, *42*, 246–253.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). *Running experiments on Amazon Mechanical Turk, judgment and decision making5(5). (pp. 411–419), 411–419.*

Peterson, R. A., & Merunka, D. R. (2014). Convenience samples of college students and research reproducibility. *Journal of Business Research*, *67*, 1035–1041.

Poynter, R. (2010). *The handbook of online and social media research* (1st ed.). Chichester, West Sussex: John Wiley & Sons, Inc.

Poynter, R., & Compley, P. (2003). Beyond online panels. *Proceedings of ESOMAR Technovate Conference*. Amsterdam: ESOMAR.

de Rada, V. D., & Dominquez-Alvarez, J. A. (2013). Response quality of self-administered questionnaires: A comparison between paper and web questionnaires. *Social Science Computer Review*, 1–14. http://dx.doi.org/10.1177/0894439313508516.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowedworkers? Shifting demographics in Mechanical Turk. *CHI 2010: Human factors in computing systems* (pp. 2863–2872).