Mechanical Turk in Consumer Research: Perceptions and Usage in Marketing Academia

Scott A. Wright

Providence College

Joseph K. Goodman

The Ohio State University

forthcoming

in *Handbook of Methods in Consumer Psychology*, eds. Frank R. Kardes, Paul M. Herr, and Norbert Schwarz.

Table of Contents

Consumer psychologists are increasingly relying on crowdsourcing websites like Amazon's Mechanical Turk (MTurk) to conduct research (Goodman & Paolacci, 2017). This transition away from more tradition samples (e.g., undergraduate students) to those obtained by means of the Internet is pervasive in the social sciences: over 15,000 published papers referenced MTurk in the past 10 years (Chandler & Shapiro, 2016) and some journals have seen more than a fourfold increase since 2012 (Goodman & Paolacci, 2017). With its increased usage, the quality of the data obtained using crowdsourcing has received considerable attention. While most concerns have been debunked (Chandler & Shapiro 2016; Goodman, Cryder, & Cheema, 2013, Goodman & Paolacci 2017; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010), some concerns may have merit. For example, a large proportion (18%) of MTurk workers admit engaging in other activities while completing MTurk tasks (Chandler, Mueller, & Paolacci, 2014), and they are more likely to participate in studies where they know the researcher (Necka, Cacioppo, Norman, & Cacioppo, 2016). There are also concerns regarding representativeness as crowdsourced samples tend to be more liberal, better educated, less religious, and younger than the US population (Paolacci & Chandler, 2014). Further, there are questions about how MTurk is viewed by the marketing academic field. For example, do marketing academics perceive that MTurk is viewed negatively outside of consumer psychology or that review teams reject MTurk studies altogether (an outcome experienced by one author, see also Hauser, Paolacci, and Chandler, this issue), perhaps discouraging MTurk research in the first place. Do researchers perceive that MTurk is only used by a select few? That studies are cherry-picked? Or that MTurk studies should not be run on certain days of the week? Unfortunately, it is hard to know if these concerns reflect legitimate facts or anecdotal fables.

In this chapter, we review these concerns and assess marketing academicians' current use and opinions of crowdsourcing tools, particularly MTurk. Further, we assess whether the

use and opinions of MTurk may differ by experience, age, and research paradigm. We addressed these questions by surveying full-time marketing faculty at the top 150 business schools.

Our results show that the use and opinions of MTurk in marketing continue to change in several ways. First, as we might expect, the use of crowdsourcing for research purposes varies widely by research paradigm (e.g., consumer psychology, quantitative modeling, and marketing strategy). Second, and perhaps more surprising, perceptions of data quality also vary by research paradigm and age. Compared to consumer psychologists, marketing faculty that do not describe themselves as consumer psychologists (and are thus less likely/less frequent users of MTurk) are more likely to distrust the validity of crowdsourced data. Similarly, older faculty are more likely to distrust MTurk's validity compared to junior faculty, but this effect seems to be driven by their usage—older faculty are less likely to have used MTurk, which in turn is associated with less trust for the platform. Thus, it seems that researchers and editors who are less familiar with crowdsourced data drive this distrust. Third, our findings identify which techniques researchers use to ensure the quality of their crowdsourced data (e.g., using attention checks, applying filters). Some of these techniques align with recommended practices substantiated in the literature (see Hauser et al., this issue), but we identify multiple areas of discord. For example, the most common practice is to incorporate attention checks; however, attention checks have their own problems (see Hauser et al., this issue) and filtering workers based upon reputation (i.e., > 95% approval rating) is sufficient to obtain high quality data (Peer, Vosgerau, & Acquisti, 2014). In addition, despite research identifying the threat of participant attrition (Zhou & Fishbach, 2016), a large proportion of researchers fail to check attrition rates. Lastly, we find researchers in the past have avoided data collection on specific dates, days of the week (e.g., Saturday), and times of the day (we are not aware of any research that has tested whether this affects data quality). In

sum, despite the extensive research investigating crowdsourced data (e.g, Goodman & Paolacci, 2017; Chandler et al., 2014), there is still skepticism towards MTurk data from marketing academics. Therefore, we should continue to investigate crowdsourcing data practices and issues in marketing and whether perceptions of MTurk are keeping pace with current research.

In the next section, we provide a brief overview of MTurk, before discussing the latest research examining the quality of crowdsourced data. We then examine commonly used practices by researchers to ensure its integrity. Next, we describe the details of our survey, which assess the data quality practices of marketing academics as they pertain to crowdsourcing. We then report our results, which highlight a number of important and unexpected findings of particular interest to consumer psychologists.

Crowdsourcing Overview

Mechanical Turk (MTurk) is an online labor market where requesters post jobs or tasks (referred to as *HITs*) and individuals (referred to as *workers*) choose which of these jobs to complete. MTurk is the most commonly used crowdsourcing platform, but other platforms exist (e.g., Crowdflower, Prolific). One of the key benefits of a crowdsourcing platform is that it facilitates hundreds of small payments to workers for their participation in research. Crowdsourcing platforms equip researchers with a considerable degree of control over the tasks they post and the types of workers they deem eligible to complete these tasks. For instance, on MTurk, researchers can predetermine which characteristics qualify/disqualify workers from participation (e.g., country of residence, worker reputation, participation in previous studies), how much they would like to compensate workers, and whether or not the quality of worker submissions merits any compensation at all. Even if researchers do no

actively reject work for low quality, the mere prospect of work being rejected can motivate workers (Sappington, 1991). Researchers can also post a variety of different tasks. Many are essentially invitations to complete online experiments in exchange for payment; however, as our findings show, researchers commonly collect survey data and pretest stimuli using crowdsourced samples. For a comprehensive review, see Chandler and Shapiro (2016), Goodman et al. (2013), Goodman and Paolacci (2017), and/or Mason and Suri (2012).

Perceptions of MTurk Data: Facts and Fables

A substantial number of researchers have raised concerns with crowdsourced data and appear skeptical, despite consistent evidence demonstrating its high quality, reliability, and validity (Hauser et al., this issue; Holden, Dennie, & Hicks, 2013; Paolacci & Chandler, 2014; Crump, McDonnell, & Gureckis, 2013; Horton, Rand, & Zeckhauser, 2011). Given that researchers cannot directly monitor workers on MTurk, the medium evokes several concerns related to worker attentiveness (Hauser & Schwarz, 2015, 2016; Oppenheimer, Meyvis, & Davidenko, 2009), dishonesty and misrepresentation (Cavanagh, 2014; Chandler & Paolacci, 2017; Sharpe Wessling, Huber, & Netzer, 2017; Shapiro, Chandler, & Mueller, 2013), and overall data quality (Buhrmester, Kwang, & Gosling, 2011; Chandler & Shapiro, 2016). In addition, researchers have raised other concerns related to non-naïveté (Chandler, Paolacci, Peer, Mueller, & Ratliff, 2015), representativeness (Shapiro et al., 2013), and questionable research practices. Though many of these concerns have been investigated empirically (with suggestions to researchers on how to address these issues, Hauser et al., this issue), it appears that many academics—including reviewers and editors—in the field of marketing may continue to hold these beliefs. However, there is currently no systematic study assessing marketing academician's perceptions of MTurk. Before we present our study, we

briefly summarize each of these perceived concerns and how they will be measured in our study.

Attention. Research shows that the attention MTurk workers apportion while completing a task may vary (Chandler, Paolacci, & Mueller, 2013). Workers have admitted to multi-tasking (e.g., watching TV or listening to music) or even walking away during a task. Despite these differences, MTurk workers exhibit attention levels that match, and sometimes exceed, that of traditional samples, such as undergraduate and community samples (Goodman et al., 2013; Hauser & Schwarz, 2016; Paolacci et al., 2010; Ramsey, Thompson, McKenzie, & Rosenbaum, 2016). The greater attention may be due to MTurk's incentive system. MTurk workers want to complete high quality work to ensure that they are paid and that their work is approved, which keeps their approval rating high (e.g., > 95%) making them eligible for more work. Even though most researchers do not reject poor quality work, the mere probability of a given HIT being rejected can motivate an agent (Sappington, 1991). In sum, research suggests attention is not a particular issue for MTurk research, but researchers have expressed concerns. Thus, in our study we will assess whether marketing researchers perceive MTurk workers as less likely to pay attention compared to traditional samples.

Dishonesty and Misrepresentation. Although there is no evidence that MTurk workers are more dishonest than the public, there is evidence that MTurk workers will be dishonest and misrepresent themselves if given strong incentives to do so (Balasubramanian, Bennett, & Pierce, 2017; Chandler & Paolacci, 2017; Sharpe Wessling et al., 2017). Of particular concern is self-screening: when researchers ask workers to participate only if the worker meets certain criteria, such as gender, product ownership, or engaging in a certain behavior (Downes-Le Guin., Mechling, & Baker, 2006). However, a vast majority of workers are honest and do not misrepresent themselves (Chandler & Paolacci, 2017), and it only becomes a problem when the screening criteria is narrow and workers are allowed to self-screen. The

problem can easily be solved by removing the economic incentive to misrepresent, obfuscating eligibility requirements, and/or maintaining participant pools (or "panels", see Hauser et al. this issue for a detailed discussion). Nonetheless, problems around misrepresentation may have led researchers to overgeneralize and believe all MTurk workers are dishonest. Thus, in our study we will examine whether researchers perceive MTurk workers as dishonest, or more likely to lie, than other participants.

Non-naiveté. Non-naiveté is another increasing concern with crowdsourced sampling because research estimates that approximately 10% of workers complete 41% of tasks (Paolacci & Chandler, 2014). Compared to traditional recruitment methods (e.g., university pools), where individuals complete a handful of studies over a short time, crowdsourced workers can complete dozens of tasks a day over a long period. This experience increases exposure to common experimental paradigms (e.g., trolley problem or a mood manipulation), which reduces effect sizes (Chandler et al. 2015), can lead to practice effects (Basso, Bornstein, & Lang, 1999), and likely decreases the effectiveness of attention checks (Peer et al., 2014; Thomas & Clifford, 2017). Nonetheless, there are strategies researchers can use to mitigate problems from non-naiveté (see Hauser et al., this issue).

A similar concern is non-naiveté from sharing information. Through online forums (e.g., MTurk Forum, Hits Worth Turking For, and Turker Nation) workers have posted information about eligibility requirements, attention and memory checks, and experimental conditions (Sharpe Wessling et al., 2017). Though the top two forums, MTurk Forum and Hits Worth Turking, have over 90,000 registered users (the number of active users is unknown), less than 10% of workers report finding MTurk studies through means other than MTurk (Casey, Chandler, Levine, Proctor, & Strolovitch, 2017), and workers post information on an extremely small number of HITs. In sum, researchers may distrust MTurk due to concerns about participant non-naiveté; therefore, in our study we will examine

whether researchers perceive MTurk data as less trustworthy than other similar data collection methods.

Data Quality. Notwithstanding the previously discussed concerns, crowdsourced data has been shown to be comparable, if not better, to data collected using other sampling sources for many cognitive, social, psychological, and decision-making tasks (see Holden, Dennie, & Hicks, 2013; Paolacci & Chandler, 2014; Crump, McDonnell, & Gureckis, 2013; Horton, Rand, & Zeckhauser, 2011). Psychometrically speaking, MTurk data provides alpha and test–retest reliabilities that are comparable to traditional samples (Buhrmester et al., 2011; Shapiro et al., 2013). Crowdsourced data also have high convergent and concurrent validity (see Chandler & Shapiro, 2016). Further, research has consistently replicated behavioral findings in consumer psychology, decision-making, and political science, and often without any difference in effect size (Berinsky et al., 2012; Goodman et al., 2013; Mullinix et al., 2015; Paolacci et al., 2010). Investigating 15 independent replications, political science researchers found that results from MTurk were similar to those from national samples (Coppock, 2016). Also, Kees, Berry, Burton, and Sheehan (2017) found that their MTurk sample (at $0.75 per participant) performed better on manipulation and attention checks and provided more reliable data compared to an equivalent Qualtrics sample (at $3.75 per participant).

Though there is evidence that MTurk data is comparable to data collected using traditional samples, we have little knowledge about whether researchers' perceptions of MTurk mirror this reality. Do researchers, reviewers, and editors trust MTurk data? Or do some still believe that manuscripts should not have all their data from MTurk, or that a lab study is superior? Many of the concerns outlined above are essentially "solvable" (e.g., by prescreening, using instructional manipulation checks, or screening by approval rating; see Hauser et al., this issue). Thus, in our study we will assess whether marketing researchers perceive these as non-issues or as problems that make MTurk data less trustworthy.

Research Practices. MTurk provides several benefits to researchers compared to traditional sampling methods. MTurk is low cost (in payments to participants, but also in staffing costs, accounting costs, etc.), fast, flexible, more representative, and allows researchers to target specific populations (Goodman & Paolacci, 2017). However, these benefits may have unintended consequences, and MTurk may also change how researchers do research. For instance, Pham (2013) warned that researchers, instead of answering questions that are interesting and important, are answering questions that are easily answered on MTurk. Further, the low cost nature of MTurk may exacerbate questionable research practices, such as p-hacking and the file drawer problem because it is easier, faster, and cheaper to run additional studies. Thus, in our study we will assess whether researchers believe MTurk is facilitating a change in research practices.

The next question is whether researchers are following best practices when conducting MTurk studies to address these issues. The deliberate actions taken by researchers, which we refer to as data quality practices, influence data quality. We review the data quality practices of marketing researchers next.

Usage of MTurk: General Usage and Common Data Quality Practices

While it is helpful to know how MTurk is being perceived by marketing academics, it needs to be placed in context of how it is actually being used. Are researchers really using MTurk this frequently? Do they use the best practices of the field to avoid the issues we discussed? There is currently no comprehensive data on who is using MTurk and how it is being used by marketing academics. We first discuss the key metrics of MTurk usage in this section, and in the next section discuss our study.

General Usage. The use of MTurk by consumer psychologists has increased exponentially over the past 10 years (Goodman & Paolacci, 2017). Beyond the number of studies appearing in published articles, we know little about its usage. What percentage of studies are run on MTurk? And how many per month? Are researchers using other platforms? Is MTurk used by non-CB researchers? To address these questions, in our study we assessed the general usage of MTurk and other crowdsourcing websites by all marketing academics.

Compensation. Research has explored the issue of worker compensation and data quality, consistently finding no relationship between payment level and data quality. Even at low compensation rates, data quality remains high (Buhrmester et al., 2011; Mason & Watts, 2010); however, there is a positive association between compensation amount and participation rates (Buhrmester et al., 2011). That is, people are more likely to participate when requesters increase payment amounts.

In terms of payment, researchers and worker websites have suggested paying at least $.10 per minute or the minimum wage in a researcher's location (Goodman & Paolacci, 2017; Prolific.ac). Prolific (formerly called Prolific Academic) endorses "ethical rewards" and asks that researchers pay at least $6.50 USD per hour. Examining what researchers actually paid, Stewart et al. (2015) pooled the MTurk account information across seven behavioral laboratories (located in the US, UK, the Netherlands, and Australia). The dataset included 33,408 workers who completed 114,460 HITs. The median hourly wage reported was $5.54. Although MTurk workers are considered contract workers, it is important to note that these figures are below the U.S. minimum wage of $7.25.

Despite the research on the effects of compensation, we have little information regarding compensation rates by consumer psychologists. MTurk is truly a free market, with no restrictions on participant payments. What we do not know is how marketing academics

compensate workers; thus, in our study we assess the current hourly compensation rate for MTurk studies.

Attention Checks. Though research suggests that workers pay attention at similar or higher rates than traditional samples (Goodman et al., 2013; Paolacci et al., 2010; Ramsey et al., 2016), attention checks have received their own attention in the literature (e.g., instructional manipulation checks or IMCs; Oppenheimer et al., 2009; Hauser & Schwarz, 2015, 2016; Hauser, Paolacci, & Chandler, this issue). It appears that researchers are frequently including attention checks when collecting data on MTurk data. If these checks are prevalent, then some workers may have habituated, exhibiting greater attention to attention checks than traditional subject pool samples (Hauser & Schwarz, 2016). Another issue is that some attention checks can actually induce more systematic, System 2 processing (Hauser & Schwarz, 2015) and alter worker behavior in unintended ways (Berinsky et al. 2012; Rand et al., 2014). The good news is that these attention checks may not be necessary since crowdsourced workers have their own incentives to pay attention and screening based on approval rating (e.g., 95% approval) appears to be just as effective (Peer et al., 2014; see Hauser et al., this issue for more details). Unfortunately, there is no systematic measure of the use of attention checks in marketing research; thus, in our study we will measure the current usage of attention checks.

Attrition Rates. Recent research has highlighted the important issue of participant attrition (i.e., drop outs) among online samples. According to Musch and Reips (2000), the mean dropout rate of a typical Web experiment is around 34%. Zhou and Fishbach (2016) found similarly high dropout rates for MTurk samples, which were statistically higher compared to traditional sampling sources (e.g., lab samples). Though we have not experienced such high dropout rates using MTurk (unless there was an error in the study),

they can become a problem under certain conditions (see Hauser et al., this issue for a discussion of solutions).

More concerning is when attrition rates vary by experimental condition (Zhou & Fishbach, 2016). Dropout rates that vary according to experimental condition jeopardize random assignment, subsequently weakening the internal validity of an experiment and introducing experimental confounds. When researchers do not check for attrition rates, they may reach erroneous conclusions (e.g., that imagining applying eyeliner leads to weight loss; Zhou & Fishbach, 2016). The simple solution is to check for attrition rates, which the authors recommend; however, almost no MTurk studies reported in the literature report attrition rates, suggesting researchers may not be checking or simply not including that information. Thus, in our study we will measure whether researchers report checking attrition rates in their MTurk studies.

Screening Practices. Researchers use various screening and filtering methods to ensure data quality. The most commonly used methods are to pre-screen by participant location (e.g., US-only workers) and worker reputation (e.g., approval rating of 95% or higher, MTurk Master workers). MTurk provides these two screening criteria, which makes them easy and costless to implement. MTurk has approximately 500,000 users (Stewart et al., 2015), with 80% of tasks being completed by less than 10,000 workers (Fort, Adda, & Cohen, 2011), but these individuals are predominately from the US. In the past, Indian workers were as high as 34% of workers (Ipeirotis, 2010), but today it is less than 10% (Goodman & Paolacci, 2014). In terms of data integrity, prior research shows a demonstrable difference between these populations (Litman, Robinson, & Rosenzweig, 2015), with US samples generally providing higher quality data compared to Indian samples (but this may be due to language and not inattention).

Beyond geography, researchers can also easily screen based on reputation, and there is some evidence of a difference in data quality—data quality is generally higher for those that exceed the 95% approval rating benchmark (Peer et al., 2014). Researchers can also pre-screen workers based upon their participation in previous studies by cross listing WorkerIDs against previously collected studies or by using Web-server-based software (e.g., TurkPrime) that maintains a database of previous workers (Goldin & Darlow, 2013). Duplicate responses can threaten data quality through non-naiveté and by violating assumptions of statistical independence (Chandler, Mueller, & Paolacci, 2014). It is also common for researchers to apply post-screens: filtering out workers once data have already been collected. For example, researchers filter out workers based upon self-report measures (e.g., *have you participated in this HIT previously?*) and duplicate IP addresses. In our study we will measure the extent to which researchers are currently using these pre-screening tools.

Ongoing Panels. As previously discussed, self-screening is problematic (Sharpe Wessling et al., 2017; Chandler & Shapiro, 2016). Thus, researchers recommend MTurk researchers adopt a two-step approach when screening participants (Chandler & Paolacci 2017; Sharpe Wessling et al., 2017). The first step is essentially a pre-screening survey to determine who is appropriate for the subsequent focal survey or experiment (to be collected in the second step). The second stage is the focal task(s), where the researcher only invites eligible workers to participate. An alternative to the two-step approach is to use a panel service such as TurkPrime. For an additional fee (depending on the selection criteria), TurkPrime will only allow MTurk workers that meet the researcher's pre-defined selection criteria (e.g., age, gender, income, political affiliation, etc.). We are not aware of any published studies that have used either of these methods, and it is not clear the prevalence of their use. Thus, in our study we assess whether researchers are still allowing participants to self-screen.

Timing. One big advantage to crowdsourced sampling is the ability to collect data at any time. Yet, we have no indication as to whether researchers collect data indiscriminately or if they strategically consider *when* to collect crowdsourced data (i.e., during the day versus night, and on weekdays versus the weekend). Only a handful of studies have explored the temporal aspects of crowdsourced sampling, but they have found that MTurk worker demographics vary over time (Arechar, Kraft-Todd, & Rand, 2017; Casey et al., 2017). For example, workers who complete a task at night (vs. in the morning) are more likely to be single, to have used a smartphone to complete the survey, and tend to be less prolific MTurk workers. Conversely, workers completing tasks in the morning (vs. at night) tend to be male, older, higher in emotional stability, and more conscientious. Thus, there is likely heterogeneity in participant characteristics based on participation timing, which may lead researchers to avoid collecting data at certain times of the day or week. Thus, in our study we will examine whether marketing researchers using MTurk are concerned about timing and avoid collecting data at certain days of the week or times of the day.

Methodology

We conducted a large-scale survey of marketing academics from all sub-fields (consumer behavior, quantitative/modeling, marketing strategy, etc.) at top research institutions to assess the current usage and perceptions of crowdsourcing. We invited marketing faculty from the top 150 research-oriented business schools to participate in an online survey in exchange for inclusion into a drawing for a $200 Amazon gift certificate. To compute business school rankings, we selected the top schools according to the University of Texas at Dallas (UTD) database that tracks publications in the top 24 business journals, which is a subset of the top 45 journals used by the *Financial Times*. We compiled the email

addresses of the marketing faculty associated with each business school identified in the rankings using public repositories (e.g., business school websites and online faculty profiles). We restricted our list to full time faculty only; thus, we excluded adjuncts, visitors, post-docs, and part-time faculty. This yielded 1,851 faculty (12.32 per department), which served as our sampling frame.

All 1,851 marketing faculty where invited, via email, to participate in an anonymous online survey distributed through Qualtrics "on how marketing academics use (or don't use) Amazon's Mechanical Turk (MTurk) for research, and how researchers view data collected on MTurk." To minimize self-selection we purposely mentioned non-use of MTurk to encourage everyone to participate. Nonetheless, participants ultimately self-selected whether to complete the survey; thus, the sample may be more representative of those who use MTurk. We first directed respondents to a confidential survey, which only collected optional emails for follow-up and to be entered into the gift card drawing. We then forwarded participants to an anonymous survey that could not link responses to their identity. We received 320 responses (17% response rate) during the data collection period (from January-February, 2018) and 258 (14%) completed the entire eight minute (median response time) survey.

Measures

Following a brief introduction, respondents answered a series of questions designed to assess their usage of crowdsourcing, worker compensation, data quality practices and perceptions, collection timing, and sample characteristics. Below, we describe each measure in detail.

Measuring Usage of MTurk

General Usage. We presented three questions to assess crowdsourcing usage. We asked respondents whether they use crowdsourcing for research purposes (*Yes/No*), how many studies they run on average per month (*open-ended*), and what percentage of these studies are collected using the following sources: *Amazon Mechanical Turk (MTurk), CrowdFlower (CF), Prolific (formerly known as Prolific Academic), University/Department Participant Pool, Public Places (union, coffee shop, etc.), Field Studies,* and *Other - Please Specify.* We also asked respondents to indicate what percentage of crowdsourced studies are for the following purposes: *Pretesting Stimuli, Data Collection for Experimental Research, Data Collection for Survey Research, Stimuli Design,* and *Other - Please Specify*.

Compensation. To assess compensation as it relates to marketing academics and consumer psychologists, we asked respondents how much they pay workers per minute and how much respondents spend per month (on average) on crowdsourcing.

Data Quality Practices. We included three questions asking respondents what percentage of their crowdsourced studies implement various data quality practices. The first question asked about the use of attention checks and other data verification practices (i.e., trapping questions, reverse wording, and IP address verification). The second set asked questions regarding screening practices (i.e., screening workers according to reputation, location, etc.), and the third set asked about screening practices to avoid non-naiveté (screening out participants that have participated in previous studies, using participant self-screens after a study, etc.). We included a "not sure" option for each data quality practice in question for participants unable to respond or unfamiliar with a particular data quality practice. We also measured whether requesters intentionally *avoid* collecting crowdsourced data at certain times of the day, week, and/or year (e.g., Monday—Friday, afternoons, holidays).

Measuring Perceptions of MTurk Data

Source Quality. To determine the perceived quality of crowdsourced data, relative to other data sources, we asked respondents to rate the quality of data from the various data sources on 7-point scales anchored at *Very Low Quality—Very High Quality*. We asked about the following sources: *Labs - Using Students, Labs - Using Non-students, Labs - Using Graduate Students, Crowdsourcing - MTurk, Crowdsourcing – Non-MTurk, Online Panels (e.g., Qualtrics panel), Field Studies, Panel Data,* and *Scanner Data*.

Distrust. To address potential differences in trust perceptions, we asked respondents to indicate how strongly they agree or disagree with four statements that measured different aspects of trust. The four statements were: "*Papers should not have all their data from Mturk (at least one study should be non-Mturk)*", "*I question the validity of data obtained using mturk*", "*Mturk samples are better than student samples*" (reverse coded), and "*Mturk studies should be re-run using a non-Mturk sample*" ($\alpha = .74$).

Research Practices. To assess whether respondents associate MTurk with problematic research practices, we asked respondents to indicate how strongly they agree or disagree with the following statements: "*Mturk contributes to the 'p-hacking' problem*" and "*Mturk increases the file drawer problem*". Both were measured on 7-point scales anchored at *Strongly Disagree—Strongly Agree* ($\alpha = .81$). We also asked participants whether they believed MTurk was changing research by asking whether "*MTurk has changed my research for the better"* and, moving forward, the extent to which they themselves and the field would use crowdsourcing (both anchored at 1 = *Much Less in the Future* and 5 = *Much More in the Future*).

Attention and Compensation Perceptions. To measure perceptions of attention, we asked participants whether they believed "*MTurk workers pay attention*". To measure

perceptions of compensation, we asked participants whether they believed "*MTurk workers are underpaid*".

Respondent Demographics and Characteristics

We included a series of measures to describe the sample in terms of respondent characteristics: age (*open-ended*), gender *(Male, Female, Other, Prefer not to Answer*), academic rank (*PhD student, Post-Doc or Visiting Professor, Assistant Professor, Associate Professor, Full Professor, Chaired Professor, Other – Please Specify*), and research focus (*Consumer Behavior – Experimental, Consumer Behavior – CCT, Strategy, Modeling – Empirical, Modeling – Analytical, Other – Please Specify*). We also asked respondents to report their involvement with editorial processes by asking if they were currently a journal Associate Editor/Editor, Editorial Review Board (ERB) Member, reviewer, and/or reader.

We also included measures to describe the sample in terms of institutional characteristics. We asked respondents to report school type (*Public or Private*), geographic location (*North America, South America, Europe, Asia, Australia*, and *Africa*), and whether their department offers a PhD program (*Yes/No*). We also asked respondents to report the research focus of their respective institutions according to an 11-point bi-polar scale (1 = *Teaching School*; 6 = *Balanced*; 11 = *Research School*). Given that the sampling frame consisted of the top 150 research business schools, we anticipated the mean response to be well above the scale midpoint.

Results

Descriptive Findings

Tables 1 and 2 provide descriptive data for our sample. As illustrated, the majority of respondents were Assistant (40%) or Associate Professors (29%) focused on consumer psychology with an experimental approach (66%). Respondents were predominately from North American institutions (79%) and were highly research focused: Many work at PhD granting institutions (86%), serve as Editors/AEs (10%), or serve on at least one Editorial Review Board (26%). This was further echoed by our research focus measure (1 = *Teaching School*; 11 = *Research School)*, with a high mean response (M = 9.57; *SD* = 1.61). As noted previously, it is possible that MTurk users may be overrepresented in the sample due to self-selection.

Table 1. Respondent Characteristics.

Characteristics					
	N	%		*n*	%
Marketing Sub-field			**Academic Rank**		
Consumer Behavior (Experimental)	169	66	Assistant Professor	101	40
Modeling (Empirical)	38	15	Associate Professor	75	29
Strategy	24	9	Full Professor	35	14
Modeling (Analytical)	9	4	Chaired Professor	37	145
Consumer Behavior (CCT)	7	3	Other	7	3
Other	10	4	**Editorial Appointments**		
Age (*M* = 43.33)			Editor/Associate Editor	40	10
Under 30	8	4	Editorial Review Board Member	107	26
30-39	99	45	Reviewer	169	41
40-49	57	26	**Gender**		
50-59	26	12	Male	144	56
60-69	22	10	Female	106	41
70+	10	5	Prefer not to answer	7	3

Table 2. Institutional Characteristics.

Characteristics		
	n	%
School Type		
Public	174	68
Private	83	32
Geographic Region		
North America	203	79

Europe	32	12
Asia	17	7
Australia	5	2
PhD Program		
Yes	220	86
No	37	14
Research Focus (*M* = 9.57)*		
< 6	3	1
Balanced School (6)	22	9
(7)	6	2
(8)	17	7
(9)	53	21
(10)	57	22
Research School (11)	99	39

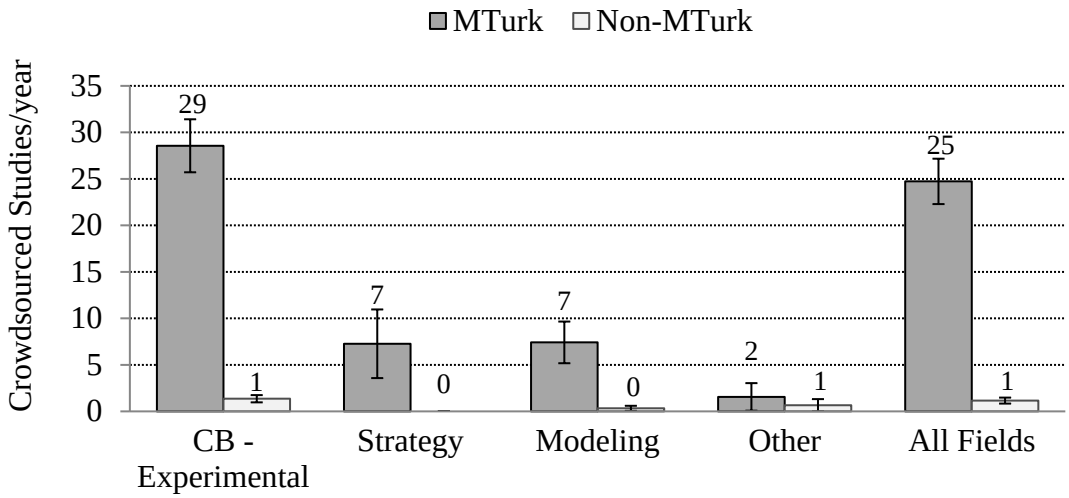
**Measured using an 11-point bi-polar scale anchored at 1 (teaching school) and 11 (research school).*

MTurk Usage and Data Quality Practices

General Usage. To begin, we investigate how and why respondents use crowdsourcing. According to our survey, 77% of respondents have used crowdsourcing for research purposes. Importantly, as illustrated in Figure 1, usage varies according to marketing sub-field. Not surprisingly, those focused on consumer behavior (CB), who employ an experimental approach (i.e., consumer psychologists), collect more studies per year using MTurk ($M = 28.56$, $SE = 1.36$) compared to other research paradigms ($p < .01$).

According to our sample, 56% of all of studies were collected using MTurk, which represents the largest sampling source, followed by University/Department Participant Pools (at 28%) and field studies (at 5%). This translates to approximately 2.06 studies per month using MTurk. Not surprisingly, only 3% of all studies are collected using other, non-MTurk, crowdsourcing platforms (e.g., CrowdFlower and Prolific). Thus, for the vast majority of marketing academics crowdsourcing equals MTurk, and MTurk is their primary sampling source.

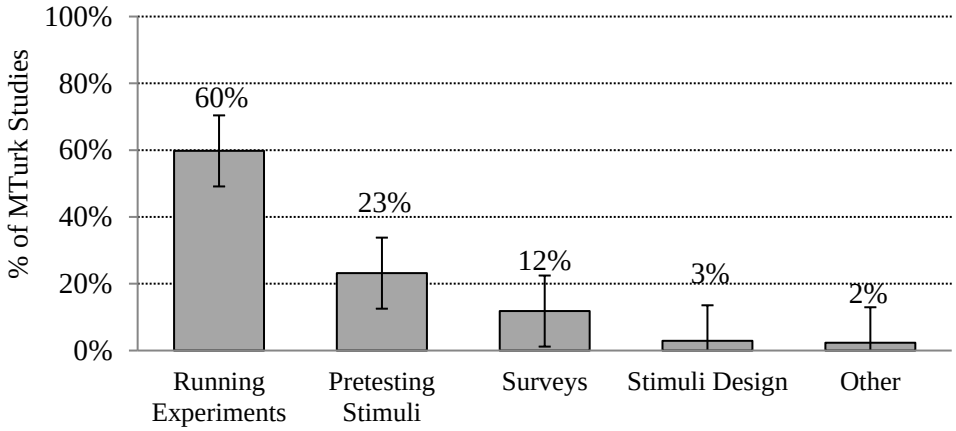
Figure 1. Crowdsourcing Usage by Research Focus.



Notes. CB = Consumer Behavior.

As illustrated in Figure 2, researchers use MTurk for different purposes. According to our sample, the primary purpose is experimental: either to run experiments (at 60%) or to pretest stimuli (at 23%)—presumably to be used in subsequent experiments. However, these uses vary according to research paradigm. Whereas, consumer psychologists predominately collect experiments (at 67%), compared to surveys (at 5%), modelers collect MTurk data about equally for both purposes (at 34% and 36%, respectively).

Figure 2. Usage Purposes.



Compensation. Our results indicate a wide range of payments by researchers. We asked participants "On average, what do you pay crowdsourcing workers per minute?" In retrospect, this was a little unclear. Out of 178 responses, 28 respondents indicated numbers greater than 1, suggesting that they did not understand the question or responded in cents instead of dollars. Another three respondents indicated 0, suggesting they did not understand the question either. To be conservative we examined the remaining 147 respondents, who reported a median response of $.10 per minute (roughly $6 per hour), with the lower quartile still at $.10 per minute and the upper quartile at $.15 per minute ($M$ = $.16, SE = .01). This value is comparable with compensation rates reported in other social sciences. Thus, the results suggest that most marketing researchers report paying $.10 per minute, which is consistent with other fields.

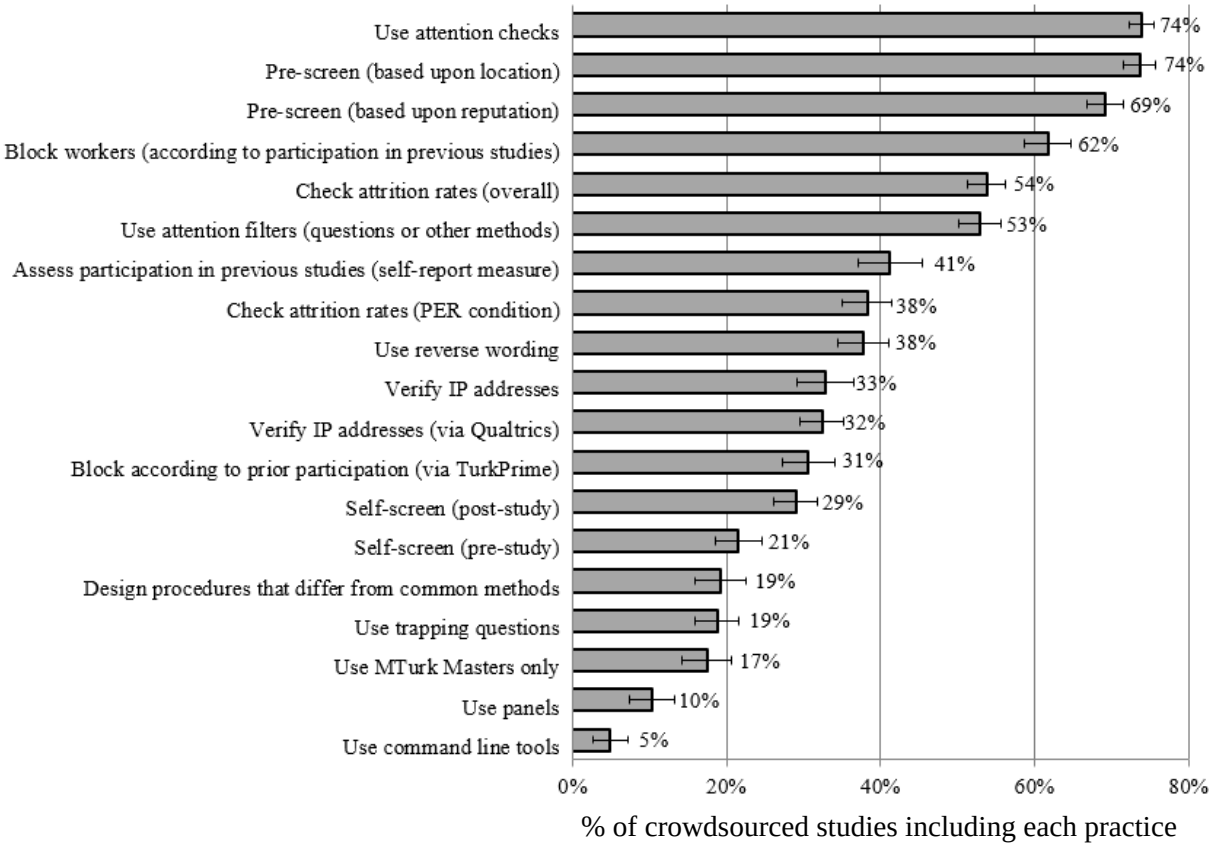
Data Quality Practices

Attention and Screening. We previously discussed several possible practices that marketing researchers may implement to ensure the quality of crowdsourced data. Hauser, et al. (this issue) also suggest several techniques to assess (e.g., measuring response speeds) and increase worker attention (e.g., warnings, trainers, and incremental text display). Our results show that marketing academics implement many of these in their own research (see Figure 3). The top three practices include adding attentions checks (74%), and prescreening workers based upon location (e.g., US-only, 74%) or worker reputation (at 69%).

Attrition. As we discussed previously, it is important to monitor and minimize participant attrition rates while using crowdsourced sampling (Zhou & Fishbach, 2016). According to our sample, respondents reported checking overall attrition rates in approximately half of their crowdsourced studies (54%), with fewer checking the attribution rate per experimental cell (38%).

Non-naiveté. As previously discussed, not all workers are naïve to research methods and may have participated in similar studies. Our results found that some researchers are taking measures to prevent non-naiveté. Researchers reported that in 62% of their studies they blocked workers according to participation in previous studies (with 31% using TurkPrime as a blocking tool), and 41% assess participation in previous studies using self-report measures. These practices are consistent with recommendations made in the literature (see Hauser, Paolacci, & Chandler, this issue). Though respondents reported that 21% of their studies let participants self-screen, they also said that 10% of their studies used panels. However, our survey did not define panels or clearly state what percentage of studies necessitating a pre-screen; thus, we caution interpretation of this measure.

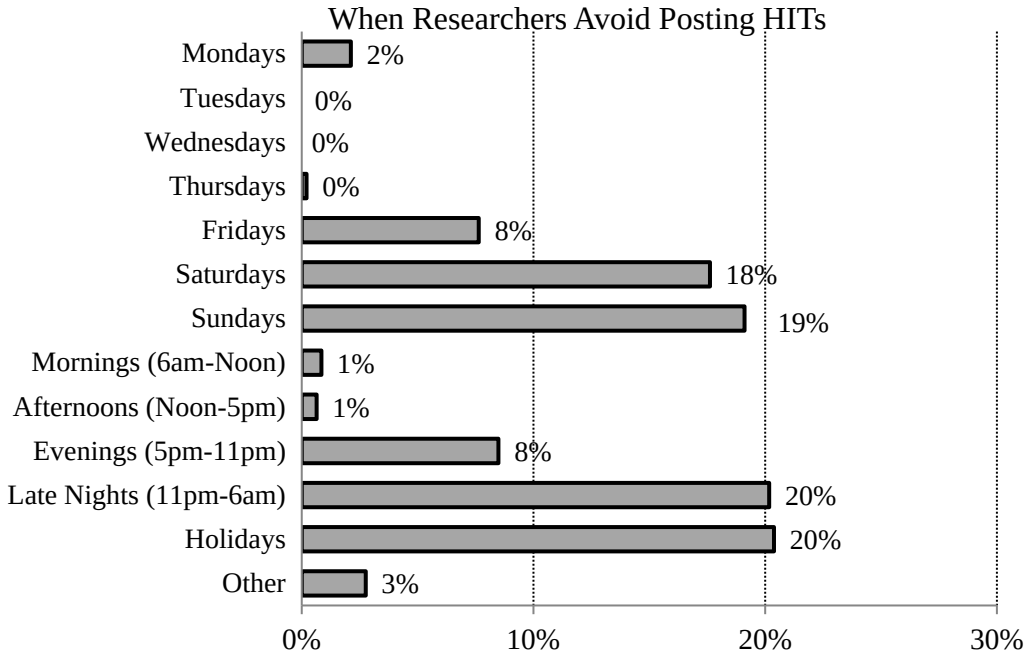
Figure 3. Common Data Quality Practices.



Data Collection Timing

In terms of timing, researchers do consider timing when collecting crowdsourced data (see Figure 4). More specifically, they intentionally avoid posting their crowdsourcing HITs on certain times of the day, days of the week, and dates. The pattern of results is largely consistent with avoiding non-work days and times (i.e., 9-5 am; Monday—Friday; and holidays).

Figure 4. The Timing of Posting HITs.



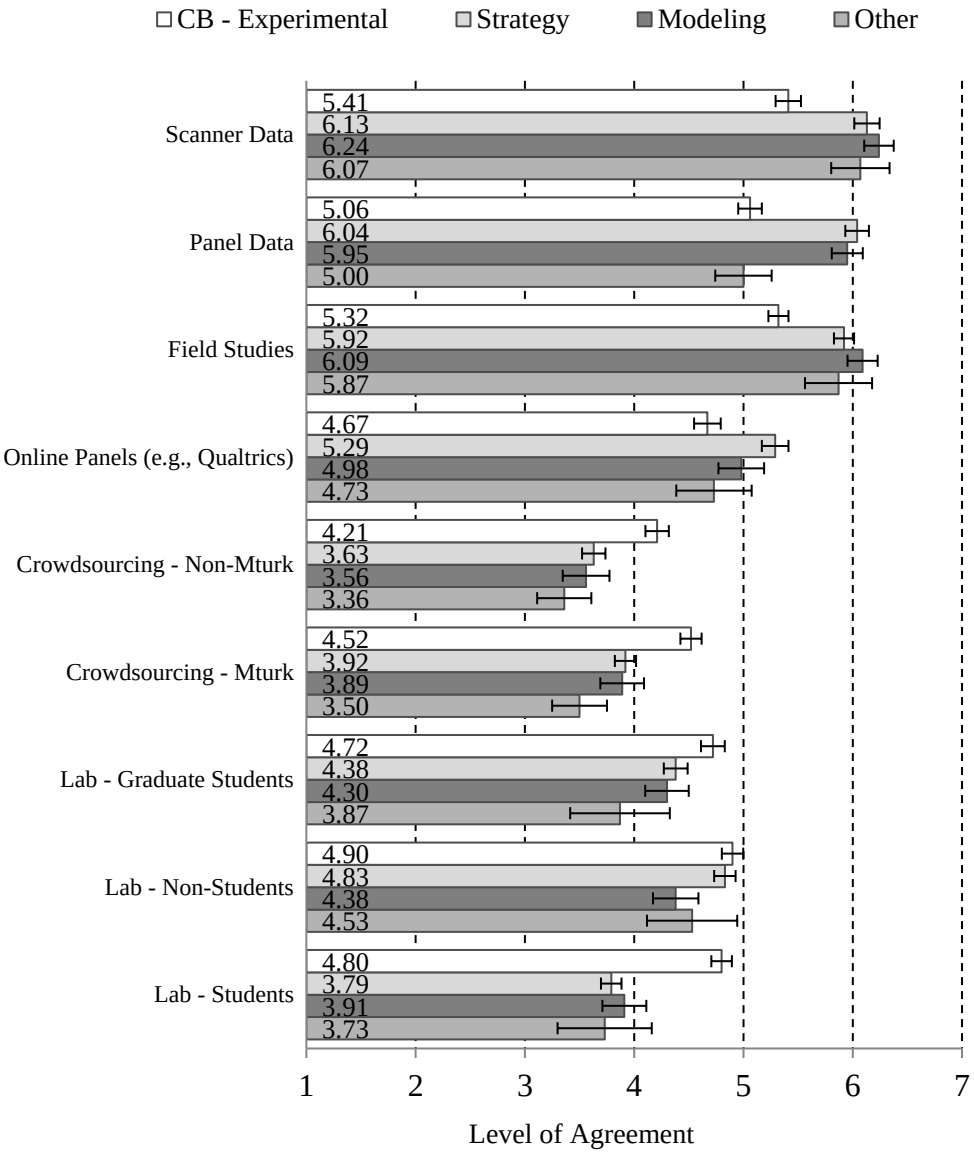
Data Quality Perceptions

General Quality Perceptions. Our results show that perceptions of a data source depends on marketing sub-field, and MTurk is no exception. Marketing academics, across all sub-fields, perceive scanner (M = 5.73, SD = 1.21), panel (M = 5.36, SD = 1.18), and field (M = 5.57, SD = 1.12) data as higher in quality compared to crowdsourced MTurk data (M = 4.29, SD = 1.30; all p's < .05) and crowdsourced non-MTurk data (M = 3.95, SD = 1.26; all p's < .05). However, and perhaps more importantly, these perceptions vary according to research sub-field, with consumer psychologists perceiving MTurk data as being higher in quality (M = 4.52, SD = 1.25) compared to those in quantitative modeling (M = 3.89, SD =

1.35), or other marketing sub-fields (M = 3.76, SD = 1.20, all p's < .004). Consumer psychologists are also less likely to distrust MTurk data (M = 3.58, SD = 1.38) compared to those in other marketing sub-fields ($M_{modeling}$ = 4.43, SD = 1.05; M_{strategy}= 4.36, SD = 1.13, M_{other}= 4.69, SD = 1.34, all p's < .009).

The opposite is true for sampling sources more common to other marketing sub-fields (e.g., quantitative/modeling). Consumer psychologists perceived scanner (M = 5.41, SD = 1.27), panel (M = 5.06, SD = 1.20), and field studies (M = 5.32, SD = 1.11) as lower in quality compared with quantitative modelers ($M_{scanner}$ = 6.24, SD = 0.91; M_{panel} = 5.95, SD = 0.94; M_{field} = 6.09, SD = 0.93, all p's < .02). These results suggest either a familiarity bias and/or a self-serving bias: individual quality perceptions are influenced by what is common to a researcher's paradigm. See Figure 5 for details.

Figure 5. Perceived Data Quality by Marketing Sub-field.



Note. CB = Consumer Behavior.

Quality by Institutional and Individual Characteristics. Next, we examined whether the perceived quality and trust of crowdsourced data varies according to institutional and individual characteristics. We regressed institutional characteristics (public/private, PhD program, geographic region, research focus, see Table 2) on perceived MTurk data quality and trust. This analysis revealed no significant effects on quality ($F(5, 246) = .50$, $p = .78$, *ns*)

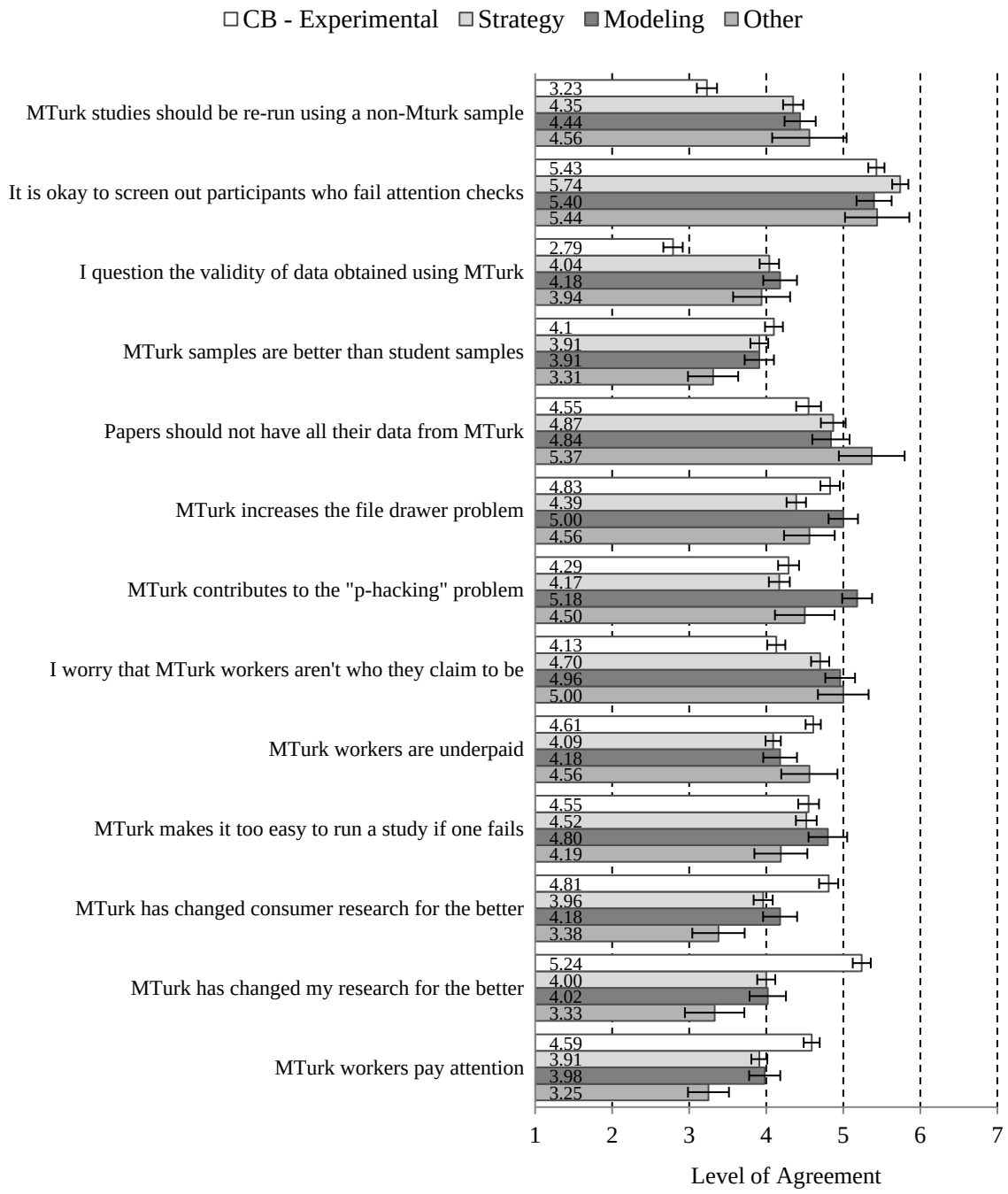
or trust ($F(5, 246) = 1.11$, $p = .35$, *ns*), suggesting the perceived quality and trust of MTurk data did not vary by academic institution type (i.e., PhD program, geographic region, or research focus).

Next, we examined individual characteristics simultaneously (i.e, marketing sub-field, age, rank, editorial appointment, and gender, see Table 1). This analysis revealed significant main effects of sub-field and respondent age for quality and trust. Consumer psychologists perceived MTurk data as higher in quality compared with other marketing sub-fields ($b = .26$, $t (215) = 2.97$, $p = .003$) and perceptions of quality are lower for older respondents ($b = -.15$, $t (215) = -2.21$, $p = .03$). Rank and age are highly correlated ($r = .73$), and removing rank from the model leads to the same conclusion: perceptions of quality are lower for older respondents ($b = -.15$, $t (218) = -2.13$, $p = .04$). The same holds for trust: consumer psychologists reported distrusting MTurk less ($b = -.29$, $t (215) = -3.43$, $p = .001$) compared with researchers in other marketing sub-fields.

Older and higher ranked faculty, along with non-CB scholars all have something in common—they are less likely to have used MTurk (all Wald-χ^2 p's $< .02$). Thus, it is possible that a lack of familiarity drives lower quality and trust perceptions. A simple post-hoc analysis showed that age was associated with a significant decrease in MTurk usage ($b = -.05$, $N=223$, Wald-$\chi^2 = 5.53$, $p < .02$) and trust ($b = -.028$, $t(218) = 14.68$, $p < .001$). When usage (1 = *yes*, 0 = *no*) is entered into the simple model, the effect of age on trust decreases from $b = -.028$ ($t(218) = 14.68$, $p < .001$) to $b = -.017$ ($t(217) = 2.36$, $p = .019$). Age and familiarity did not interact to affect trust (p's $> .1$). Moreover, when usage patterns and feelings of distrust are included in the full regression model, the main effects of research sub-field and age become non-significant (p's $> .8$). Thus, it appears age is a proxy for familiarity, and it suggests that as researchers become more familiar with MTurk they learn to trust it more, not less.

Other Perceptions. We also asked respondents a series of questions assessing other perceptions of MTurk (e.g., whether MTurk workers are underpaid or if it is acceptable for manuscripts to contain MTurk samples exclusively). See Figure 6 for a breakdown of mean responses according to research paradigm. Consistent with our previous findings, researchers in areas outside of consumer psychology, appear to be more skeptical of crowdsourced data. For example, modeling researchers feel that MTurk is more likely to exacerbate problematic research practices (M = 5.09, SD = 1.19) compared with consumer psychologists (M = 4.56, SD = 1.55; p = .03) and researchers in other marketing domains (M = 4.38, SD = 1.23; p =.03).

Figure 6. MTurk Perceptions by Sub-field.



Note. CB = Consumer Behavior.

Worker Compensation Perceptions

Although we found that, on average, respondents felt workers were underpaid (M = 4.48, SD = .08, 95% CI [4.16, 5.12]), the compensation rates reported in our sample did not vary as a function of these perceptions (b = -.05, p = .49).

Concluding Remarks

Crowdsourcing provides researchers with several advantages. It provides opportunities and efficiencies that are unavailable with other sampling sources and empowers researchers with few resources. In light of these advantages, it is of little surprise to see our field, and many others (psychology, political science, computer science, etc.), continue to embrace crowdsourcing. In this chapter our aim was to explore the extent to which the field of marketing academia has embraced crowdsourcing data by measure researcher's perceptions and usage of crowdsource samples, particularly MTurk.

One of our main findings is that data quality perceptions do not align with what has been demonstrated empirically. Despite a wealth of research validating the high quality of crowdsourced data (e.g., Buhrmester et al., 2011; Shapiro et al., 2013; see Goodman & Paolacci 2017 for a review), many marketing academics perceive MTurk data to be of lower quality compared with other data sources (even other online sources). This may be due to variations in usage, as researchers with less familiarity with MTurk (i.e., older respondents and marketing academics outside of consumer psychology) distrust crowdsourced data and perceive it as lower quality compared to researchers more familiar with crowdsourced sampling. This may also reflect an unfamiliarity with the body of research on crowdsourced sampling. Of course, we cannot rule out the alternative explanation, which is that quality perceptions exhibit a familiarity bias and/or a self-serving bias, as individual quality perceptions align with what is common to one's research paradigm.

Our findings also illustrate the myriad of data quality practices researchers implement in their own research (e.g., pre-screens, filters, checks, trapping questions). Many of these practices align with what is currently recommended in the literature, but some are less effective or have negative consequences in terms of enhancing data quality. For example, attention and comprehension checks can introduce unintended effects (e.g., by altering processing styles; Hauser & Schwarz, 2015), and may not provide data quality benefits beyond reputation based pre-screening (Peer et al., 2014). Future research should continue to study the effects of attention checks and their potential to detect bots on crowdsourcing platforms, an issue that has only recently become a problem. Also, researchers may be paying inadequate attention to attrition rates, threatening the validity of crowdsourced data—particularly experimental data. Lastly, researchers appear to collect crowdsourced data at certain times, and avoid others, despite a lack of formal empirical tests to support these behaviors. It would be interesting to explore what naïve theories may be underlying these deliberate actions, and whether there is any merit to these behaviors (we are not aware of any systematic studies).

While there are still mixed perceptions of MTurk, we found more positive views and greater usage among more junior faculty, suggesting that MTurk usage will increase and perceptions will become increasingly more positive. To capture perceptions about the future, we asked respondents whether they intend to use crowdsourcing more or less in the future and how they felt the field would evolve. Responses reflect perceptions of growth, rather than contraction, for individuals (M = 4.21, SD = 1.21) and the field (M = 5.00, SD = 1.31; averages exceeding the scale midpoint, all p's < .01).

In conclusion, shifting methodological approaches often spur a great deal of questions, uncertainties, and misconceptions by researchers—and crowdsourced sampling is no exception. Many of these issues have been resolved in a burgeoning body of research

exploring the quality of crowdsourced data, yet there is evidence that the field has yet to fully acknowledge and embrace these findings. By assessing marketing academicians' current use and opinions of crowdsourcing we hope we have answered questions, reduced uncertainties, and corrected some of these misconceptions.

References

Arechar, A. A., Kraft-Todd, G. T., & Rand, D. G. (2017). Turking overtime: how participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association*, *3*(1), 1-11.

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist*, *13*(3), 283-292.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351-368.

Balasubramanian, P., Bennett, V. & Pierce, L. (2017). The Wages of Dishonesty: The Supply of Cheating Under High-Powered Incentives. *Journal of Economic Behavior & Organization*, 137, 428-444.

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.

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal Differences Among MTurk Workers: Time-Based Sample Variations and Implications for Online Data Collection. *SAGE Open*, *7*(2), 2158244017712774.

Cavanagh, Thomas M. (2014), "Cheating on Online Assessment Tests: Prevalence and Impact on Validity," PhD Thesis, Colorado State University.

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology*, *12*.

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

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

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

Chandler, J., Paolacci, G., & Mueller, P. (2013). Risks and rewards of crowdsourcing marketplaces. In *Handbook of Human Computation* (pp. 377-392). Springer, New York, NY.

Coppock, A. (2016). Generalizing from survey experiments conducted on mechanical Turk: A replication approach. *Polit. Sci. Res. Methods, in press. https://alexandercoppock. files. wordpress. com/2016/02/coppock_generalizability2. pdf.*

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.

Downes-Le Guin, T., Mechling, J., & Baker, R. (2006). Great results from ambiguous sources: Cleaning Internet panel data. In *ESOMAR World Research Conference: Panel Research*.

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

Goldin, G., & Darlow, A. (2013). TurkGate (Version 0.4. 0) [Software].

Goodman, J. K., Cryder, C., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213-324.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, *44*(1), 196-210.

Goodman, J. K., & Paolacci, G. (2014). Questioning the Turk: Conducting High Quality Research with Amazon Mechanical Turk. In *NA - Advances in Consumer Research*. Eds. Cotte, J. & Wood, S., Duluth, MN: Association for Consumer Research (pp 766).

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400-407.

Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *Sage Open*, *5*(2), 2158244015584617.

Hauser, D. J., Paolacci, G., & Chandler, J. (2018). "Common concerns with MTurk as a participant pool: Evidence and solutions," in Frank R. Kardes, Paul M. Herr, and Nobert Schwarz, eds., *Handbook of Research Methods in Consumer Psychology*, Routledge.

Holden, C. J., Dennie, T., & Hicks, A. D. (2013). Assessing the reliability of the M5-120 on Amazon's Mechanical Turk. *Computers in Human Behavior*, *29*(4), 1749-1754.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399-425.

Ipeirotis, P. G. (2010). Demographics of mechanical turk.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: professional panels, student subject pools, and Amazon's mechanical turk. *Journal of Advertising*, *46*(1), 141-155.

Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk. *Behavior Research Methods*, *47*(2), 519-528.

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

Mason, W., & Watts, D. J. (2010). Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter*, *11*(2), 100-108.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, *2*(2), 109-138.

Musch, J., & Reips, U. D. (2000). A brief history of Web experimenting. In *Psychological experiments on the Internet* (pp. 61-87).

Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PloS one*, *11*(6), e0157732.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872.

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.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023-1031.

Pham, M. T. (2013). The seven sins of consumer psychology. *Journal of Consumer Psychology*, *23*(4), 411-423.

Ramsey, S. R., Thompson, K. L., McKenzie, M., & Rosenbaum, A. (2016). Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior*, *58*, 354-360.

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*, *5*, 3677-3689.

Sappington, D. E. M. (1991). Incentives in principal-agent relationships. *Journal of Economic Perspectives*, *5*(2), 45-66.

Shapiro, Danielle N., Jesse Chandler J, and Pam Mueller (2013), "Using Mechanical Turk to Study Clinical Populations," *Clinical Psychological Science*, 1, 213–20

Sharpe Wessling, K., Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research*, *44*(1), 211-230.

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making*, *10*(5), 479.

Thomas, K. A., & Clifford, S. (2017). Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184-197.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493-504.