Computers in Human Behavior 77 (2017) 184–197

Contents lists available at ScienceDirect

Computers in Human Behavior

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

Review

Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments

Kyle A. Thomas [a, *], Scott Clifford [b]

[a] Department of Psychology, Harvard University, United States
[b] Department of Political Science, University of Houston, United States

ARTICLE INFO

Article history:
Received 5 February 2016
Received in revised form
12 August 2017
Accepted 26 August 2017
Available online 28 August 2017

Keywords:
Mechanical Turk
Exclusion
Attention
Manipulation check
IMC

ABSTRACT

Social science researchers increasingly recruit participants through Amazon's Mechanical Turk (MTurk) platform. Yet, the physical isolation of MTurk participants, and perceived lack of experimental control have led to persistent concerns about the quality of the data that can be obtained from MTurk samples. In this paper we focus on two of the most salient concerns—that MTurk participants may not buy into interactive experiments and that they may produce unreliable or invalid data. We review existing research on these topics and present new data to address these concerns. We find that insufficient attention is no more a problem among MTurk samples than among other commonly used convenience or high-quality commercial samples, and that MTurk participants buy into interactive experiments and trust researchers as much as participants in laboratory studies. Furthermore, we find that employing rigorous exclusion methods consistently boosts statistical power without introducing problematic side effects (e.g., substantially biasing the post-exclusion sample), and can thus provide a general solution for dealing with problematic respondents across samples. We conclude with a discussion of best practices and recommendations.

Contents

1. Introduction

The recent introduction of Amazon's Mechanical Turk (MTurk), an easy-to-use online labor market, offers researchers a potentially valuable new tool for recruiting participants. MTurk facilitates rapid recruitment and data collection from a large and diverse pool of participants at costs that are generally substantially lower than in the lab, and offers unique opportunities for cross-cultural research, longitudinal studies, and creating customized panels of participants (Amir, Rand, & Gal, 2012; Berinsky, Huber, & Lenz, 2012;

* Corresponding author. Department of Psychology, Harvard University, William James Hall 964, 33 Kirkland Street, Cambridge, MA 02138, United States.
E-mail address: kylethomas@post.harvard.edu (K.A. Thomas).

http://dx.doi.org/10.1016/j.chb.2017.08.038
0747-5632/

Buhrmester, Kwang, & Gosling, 2011; Chandler, Mueller, & Paolacci, 2014; Goodman, Cryder, & Cheema, 2012; Horton, Rand, & Zeckhauser, 2011; Huff & Tingley, 2015; Leeper, 2016; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010; Rand, 2012; Suri & Watts, 2011).

Additionally, the anonymity and physical separation afforded to participants by the platform may make them feel more comfortable disclosing sensitive or personal information, such as psychological diagnoses (Shapiro, Chandler, & Mueller, 2013), or controversial opinions (Leeper & Thorson, 2015; White, Strezhnev, Lucas, Kruszewska, & Huff, 2016). MTurk may also reduce or eliminate some experimenter expectancy effects, such as reactance (Paolacci et al., 2010; Rand, 2012), or certain kinds of social desirability bias, especially those related to researcher attributes like race or gender (Leeper & Thorson, 2015; White et al., 2016). Recently, researchers have even developed sophisticated software tools—such as MTurkR—to make everyday research through MTurk easier, and unlock new capabilities for complex study designs and cultivating customized participant pools (Huff & Tingley, 2015; Leeper, 2016).

Perhaps unsurprisingly, MTurk has been rapidly adopted as a common research platform across the social sciences, and many of the initial concerns regarding the validity of MTurk research have been addressed. A number of papers have given detailed introductions to MTurk, and explored the demographics of the MTurk population, showing that it provides more representative samples than typical lab studies (see, Berinsky et al., 2012; Casler, Bickel, & Hackett, 2013; Goodman et al., 2012; Huff & Tingley, 2015; Ipeirotis, 2010; Paolacci et al., 2010; Ross, Zaldivar, Irani, & Tomlinson, 2010). Research has also shown that MTurk participants' self-reports of demographic information are reliable (Rand, 2012; Shapiro et al., 2013), and their test-retest reliabilities on a number of psychometric scales are as high or higher than those from the existing literature (Buhrmester et al., 2011; Shapiro et al., 2013).

More importantly, researchers have replicated many well-established results from fields as wide-ranging as decision-making (Berinsky, Margolis, & Sances, 2013; Berinsky et al., 2012; Horton et al., 2011; Paolacci et al., 2010; Peer, Vosgerau, & Acquisti, 2013; Summerville & Chartier, 2013), experimental economics (Amir et al., 2012; Horton et al., 2011; Rand, Greene, & Nowak, 2012; Summerville & Chartier, 2013; Suri & Watts, 2011), social psychology (Horton et al., 2011; Rand et al., 2012; Summerville & Chartier, 2013), cognitive psychology (Casler et al., 2013; Crump, McDonnell, & Gureckis, 2013), clinical psychology (Shapiro et al., 2013), and political science (Berinsky et al., 2012, 2013; Clifford, Jewell, & Waggoner, 2015; Mullinix, Leeper, Druckman, & Freese, 2015).[1] The psychometric properties of multiple well-established personality scales have even been *quantitatively* replicated (Buhrmester et al., 2011; Peer et al., 2013; Rouse, 2015). Research has also shown that pooled judgments from MTurk participants can yield results that approximate those of established experts on a number of assessment tasks, such as annotating online content (Sayeed et al., 2011), and that they can provide evaluations of online interfaces similar to those obtained from the lab (Komarov, Reinecke, & Gajos, 2013).

Even more persuasively, results have recently emerged from large-scale replication projects specifically designed to reproduce well-established findings across a wide multitude of diverse samples, allowing for comparisons of data from MTurk samples and other research populations. For example, researchers working on the Open Science Collaboration project (see Open Science Collaboration, 2012) found little difference between the results of MTurk participants and a wide range of other samples for 13 well-established psychology findings in a large-scale collaborative replication effort (Klein et al., 2014). Mullinix et al. (2015) replicated results from 20 social science experiments through MTurk, with a correlation of $r = .75$ between estimates from MTurk samples and national probability samples (see Coppock, 2016, p. 2). Similarly, Coppock (2016) used MTurk to replicate 12 political science experiments conducted on population-based probability samples and found a correlation of $r = .83$ between estimated treatment effects from the two samples. These latter two sets of findings are especially notable because political science research tends to be one of the most sensitive areas of social science research to sample characteristics (as opposed to say, cognitive psychology research on visual systems).

Taken together, all of the above suggests that MTurk can provide researchers from a wide array of disciplines with data as reliable and valid as data collected in the lab, and given the larger and more representative samples that MTurk can often provide, such results may have greater external validity than data obtained from other convenience samples.

However, MTurk also has potential drawbacks, many of which stem from participants' physical isolation while completing a task. These potential drawbacks include a lack of experimenter control and direct participant oversight, an inability for experimenters to answer questions and resolve potential confusions, and difficulties in creating common knowledge of experimental procedures that involve participant interactions (Chandler et al., 2014; Horton et al., 2011; Paolacci et al., 2010; Rand, 2012). Additionally, many MTurk participants may be familiar with common experimental paradigms, and/or engage in uncontrolled communication with other participants through online forums (Chandler et al., 2014).

As a result, there are recurring concerns about: (1) whether participants buy into interactive experiments, and (2) whether unsupervised participants can provide reliable data. This second concern is two-fold: (2a) participants may try to game the system by ignoring important experimental materials and simply complete the task as quickly as possible to try to maximize earnings; and, (2b) because there is minimal interaction between researchers and participants, even well-intentioned participants may not understand a task, and provide unreliable data because they are unable to ask questions or request clarification before proceeding. These issues are crucial to the internal and external validity of MTurk studies, yet we lack a systematic review of the evidence on these matters.

In this paper, we synthesize the extant research from a diverse range of fields to uncover a robust pattern of results regarding these two challenges to validity, and provide new data to directly address these concerns. We find that the first concern is overblown: MTurk participants seem to buy into well-designed interactive experiments as much as participants in the lab. We also find that exclusion criteria can mitigate the second concern by eliminating or controlling for problematic respondents, which the data show generally reduces statistical noise without introducing significant sampling bias. Moreover, our results suggest that MTurk participants may pay higher levels of attention than even other high-quality Internet and student samples. We conclude with recommendations for best practices regarding interactive experiments and screening for participant attention and comprehension.

[1] It is worth noting that Krupnikov and Levine (2014) reported data suggesting that experiments relying on relatively transparent deception may not replicate well in MTurk populations. However, their conclusions apply to only a narrow set of studies and more comprehensive studies consistently support the generalizability of political science experiments (e.g., Mullinix et al., 2015; Coppock, 2016). Moreover, a reanalysis of the Krupnikov & Levine data has shown that some of the conclusions they drew were perhaps unwarranted (see Coppock, 2016, p. 2).

2. Participant interactions and internal validity

Many social science experiments critically depend upon participant interactions, such as those involving Minimal Groups or Public Goods games. Interactive experiments have traditionally been run in laboratory environments where multiple participants are in the same physical environment, and are thus cognizant of the presence of potential partners for an interaction. One might wonder then whether MTurk participants, who are physically isolated, actually believe they are interacting with real partners. If they do not, this would substantially compromise the internal validity of any study that relied on such beliefs. The available evidence to date, reviewed below, shows that MTurk participants generally buy into experimental interactions as much as participants run in the lab or the field.

A number of experimental economics findings that critically depend on participant interaction have been replicated on MTurk, including many well-established results from the lab and the field in Dictator Games, Ultimatum Games, Public Goods Games, and Trust Games (Amir et al., 2012; Horton et al., 2011; Rand et al., 2012; Summerville & Chartier, 2013). In addition, a number of more complicated experimental economics findings with additional elements, such as priming or manipulating the amount of time allotted to make one's decision, have shown the same results when run on MTurk as in the lab (Horton et al., 2011; Rand et al., 2012; Summerville & Chartier, 2013). Research has also shown that including real stakes in interactive experiments affects behavior on MTurk in the same way that it does in the lab (Amir et al., 2012). A study involving a coordination game—in which a person's best response is to do whatever they think their partner(s) will do—found a correlation of $r = 0.8$ between participants' decisions and what they thought their partner(s) would do, and that participants conditioned their decisions on the information available to their partner(s) (Thomas, DeScioli, Haque, & Pinker, 2014). Finally, results from the lab have even been *quantitatively* replicated in experiments using both a Prisoner's Dilemma (Horton et al., 2011), and a complex networked Public Goods Game (Suri & Watts, 2011).

Recent work directly examined the validity of interactive experiments run through MTurk by replicating a number of classic experiments that involved randomly assigning participants to interact with either a human partner or a computer, or randomly assigning participants to minimal groups (Summerville & Chartier, 2013). In these experiments participants were more likely to reject unfair offers from a human partner than from a computer in an Ultimatum Game, anchored off of other participants' estimates more than off of numbers generated by a computer, incurred real financial losses by giving more to human partners than to computer partners in a Dictator Game, and conditioned their donations on group membership in a Dictator Game with minimal groups. Other research has shown that manipulations that require direct training given by a researcher also seem to be equally effective with MTurk samples, despite the lack of direct personal interactions typically employed in the lab (Casler et al., 2013).

Thus, MTurk participants appear to behave in interactive experiments in much the same way as participants in the lab, despite being in physically different locations. They are even willing to alter their behavior to accommodate their partner(s) in such a way that leads them to incur financial losses in games with real monetary stakes; and, despite the fact that such stakes on MTurk are generally about ten-times less than those used in the laboratory, their behavior parallels that of participants in both lab and field studies (Amir et al., 2012). MTurk participants also report an approximately equivalent level of trust in experimenters' instructions as participants in Harvard's Decision Science Laboratory, which prominently advertises that deception is prohibited (Horton et al., 2011).

In sum, the available evidence suggests that interactive experiments run through MTurk can provide researchers with data that is as valid as data obtained from lab or field experiments: The lack of physical proximity of participants does not appear to reduce the validity of such experiments. However, these conclusions are based on the limited data that is available, and we recommend that researchers include assessments of believability in any interactive experiment (e.g., funnel debriefing to check for suspicion) as a check to ensure internal validity, and report these results to broaden our understanding of the matter.

3. Excluding participants and external validity

The physical distance of MTurk participants also leads to one of the most common concerns with this recruitment method—that the anonymity and lack of direct observation undermines participants' motivation to sufficiently engage with and understand experimental tasks. Of course, concerns with participant motivation and problematic responding are not unique to MTurk—although MTurk participants' physical isolation likely makes them more salient—and many researchers have developed screening methods for catching and removing problematic respondents, both on MTurk and in other samples. Common methods include: using “catch questions” that require participants to select an otherwise unobvious response (often referred to as *instructional manipulation checks* or *IMC*s, see Goodman et al., 2012; Oppenheimer, Meyvis, & Davidenko, 2009); asking questions with an obviously false or highly improbable answer (often referred to as *bogus items*, see Maniaci & Rogge, 2014; Paolacci et al., 2010); using consistency indices that compare responses across either synonymous or antonymous items (see Kurtz & Parrish, 2001; Meade & Craig, 2012); or, giving participants comprehension questions about experimental materials (see Berinsky et al., 2013; Horton et al., 2011; Rand et al., 2012; Thomas et al., 2014).[2] We refer to all of these different kinds of items as *screeners* throughout the rest of the paper (Berinsky et al., 2013). In a recent advance, Maniaci and Rogge (2014) developed and validated sophisticated psychometric scales to identify inattentive, inconsistent, and patterned responding (e.g., choosing the same answer for every item). Their data (as well as data from Berinsky et al., 2013) suggest that such problematic response styles should be modeled as state-like latent variables, and that their scales may provide more precise measurements of these variables than cruder methods of exclusion, thus minimizing false positives and maximizing post-exclusion sample sizes.[3]

In this section, we review existing research and present some new data on important methodological considerations, exclusion rates across samples, and exclusion effects on sample characteristics. Table 1 summarizes this research and will be relied on throughout this section to succinctly reference specific findings. To highlight supporting evidence for various claims, we point readers to where the relevant findings can be found in Table 1 with

[2] Researchers also often clean their data by trimming outliers, but we do not review such approaches here because these methods tend to be relatively weak, and are ineffective for identifying some problematic response patterns (Meade & Craig, 2012).

[3] Throughout the paper we often refer to “exclusion” of participants that fail screeners for expositional clarity and readability rather than “participants that failed the screener(s)” or other similarly clunky terms; but of course, results from participants that fail screeners could still be presented in research where this is a significant concern, as recommended by Berinsky et al. (2013) for political science research in particular due to concerns about representative samples. Where necessary, we explicitly differentiate between exclusion and screener failure, when for example, discussing what researchers might do with data from participants that fail one or more screeners other than excluding them.

Table 1
Screener failure rates across samples, type of screener method used, reported effects of exclusion on results, and reported characterizations of excluded participants.

Source	Screener failure rates			Type of screener	Effect on results	Differences in excluded participants
	MTurk	Lab	Online			
Alter, Oppenheimer, & Zemla, 2010	5-40%			1 IMC		
Ashton-James, Kushlev, & Dunn, 2013	23.2–36.8%			1 IMC		
Ausderan, 2014	29.0%			1 IMC		
Barone, Lyle, & Winterich, 2014	14.7%			1 IMC		
Berinsky et al., 2012	40%		51-54%[a]	1 comprehension item		
Berinsky et al., 2013	9-30%		24-53%[b]	1-4 IMCs	Improved statistical power	Younger; Less political knowledge; Less educated; More male; More minorities; Provided noisier data
Chandler et al., 2014	3-37%[c]			N/A		
Clifford & Jerit, 2014	7.2%	43.0%	45.1%	1 IMC, 2 bogus items	N/A	Students[d]: More minorities; Lower Need to Evaluate; Less political interest; Less political knowledge MTurk: Lower risk aversion; Less political knowledge
Clifford & Jerit, 2015	52%		63%[e]	1-2 IMCs	N/A	Online: None Lab: More black participants; Lower GPA; Less interest in foreign policy
Cryder, Loewenstein, & Scheines, 2013a	4.6%			1 IMC		
Cryder, Loewenstein, & Seltman, 2013b	12.0%			1 IMC	None	N/A
Downs et al., 2010	12-39%			1-2 comprehension items	N/A	Younger; More male; Less professional occupations and students
Gaither, Wilton, & Young, 2013	10.5%			1 IMC and 1 comprehension item		
Gipson, Kahane, & Savulescu, 2013	17.4%			1 IMC		
Goodman et al., 2012 [f]	11-13%	11-12%		1 IMC	Two marginal effects became significant	Lower on Emotional Stability; Variance higher
Goodman & Irmak, 2013	6.5%			1 IMC	None	N/A
Gray, Knobe, Sheskin, Bloom, & Feldman Barrett, 2011	6.7%			1 comprehension item		
Gromet & Darley, 2011	3.6–4.0%			1 IMC		
Halko & Kientz, 2010	20.8%			2 comprehension items		
Hardisty, Frederick, & Weber, 2012	42.6%	22.6%[g]	37.6%[h]	1 IMC	Strengthened results	Provided noisier data
Hardisty, Thompson, Krantz, & Weber, 2013			38.8%[i]	1 IMC	Strengthened results	More variance and outliers
Hauser & Schwarz, 2015??	8%–11%			1 IMC		
Hauser & Schwarz, 2016	4%–5%	61%–74%		1 IMC[j]		
Hawkins & Nosek, 2012			32.9–41.7%[k]	1 IMC and 2–3 comprehension items	None	N/A
Horton et al., 2011	52.3%			4 comprehension items	Converted a qualitative replication into a quantitative replication	N/A
Jones & Paulhus, 2014	2-4%			2 bogus items	None	N/A
Karelaia & Keck, 2013	13-19%			1-2 IMCs and an unspecified number of comprehension items		
Klein et al., 2014	6.2%	8.9%–33.6%	1.2%–46.7%	1 IMC		
Kross et al., 2014	22.6%			1 IMC		
Kteily, Cotterill, Sidanius, Sheehy-Skeffington, & Bergh, 2014	5.3–10.0%		7.0%[l]	1 item for how seriously they took survey	Strengthened results	Provided noisier data
Kugler, Cooper, & Nosek, 2010	10.0%		9.8%[m]	Survey time and unspecified attention checks		
Kurtz & Parrish, 2001		10.6–11%		Inconsistency scale and test-retest inconsistency	None[n]	None
Kushlev, Dunn, & Ashton-James, 2012	40.9%			1 IMC		
Løhre and Teigen, 2014	12.0%			1 IMC and incomplete surveys		
Lombrozo & Rehder, 2012	12.1–24.7%			1 IMC	One marginal effect became significant	N/A

(*continued on next page*)

Table 1 (*continued*)

Source	Screener failure rates			Type of screener	Effect on results	Differences in excluded participants
	MTurk	Lab	Online			
Meier, Moller, Chen, & Riemer-Peltz, 2011	13.1%			1 IMC and 2 comprehension items		
Menatti, Smyth, Teachman, & Nosek, 2011			33%[o]	1 IMC	None	N/A
Meyvis, Goldsmith, & Dhar, 2012		14.4%		1 IMC		
Moran, Rain, Page-Gould, & Mar 2014	21.1%			1 bogus item and response quality		
Oppenheimer et al., 2009		35-46%		1 IMC	Allowed replication of a classic decision-making result; Improved scale reliabilities	Report lower motivation; Lower Need For Cognition; More variance
Paolacci et al., 2010	4.2%	6.5%	5.3%[p]	1 bogus item		
Paulhus & Carey, 2011	7.4%			4 unspecified "validity-check questions"		
Peer et al., 2013	2.6–33.9%			2 IMCs and 1 bogus item	Allowed replication of a classic decision-making result; Improved scale reliabilities	Provided noisier data
Ramsey et al., 2016	50.4%	85.3%	91.5%	Non-intuitive directions[q]		
Rouse, 2015	N/A			2 attention checks	Improved scale reliability	Provided noisier data
Schuldt & Roh, 2014			20.3%	1 IMC	None	N/A
Shapiro et al., 2013	15.1–16.4%			Inconsistent responding and extreme responses	None	More Asian participants; Report more infrequent symptoms
Simmons & Nelson, 2006		38.6%		1 IMC		
Sirota, Kostovicova, & Juanchich, 2013	13.4%			1 IMC		
Stiller, Goodman, & Frank, 2011	14.3–37.2%			2 IMCs and 4 comprehension items		
Sussman & Alter, 2012		18.5%		1 IMC	None	N/A
Thomas et al., 2014	13.0–35.4%			3 comprehension items	Improved statistical power	More minorities; More Extraverted
Weijters, Baumgartner, & Schillewaert, 2013	4.2%			1 IMC	Improved statistical power	Higher rate of inconsistent responses
Yang, Vosgerau, & Lowenstein, 2013	5.5–27.1%			1 IMC and 1 comprehension item	None	N/A

[a] These were high quality panels for online political science research: Polimetrix/YouGov and Survey Sampling International.
[b] This is based on reviews of previous research, and is not a primary source.
[c] This study also included two substantive manipulation checks, which are not included in these calculations. Inclusion of these two screeners leads to very high exclusion rates as the screeners were designed to measure attention as an outcome variable rather than for exclusion purposes.
[d] Results reported here combine students in the online and lab conditions because they were drawn from the same participant pool and provided very similar data.
[e] Subjects were recruited from GfK Knowledge Networks.
[f] The exclusion rates presented here are different from those reported in the paper, because here we compare only US participants (reported to us by C. Cryder, personal communication, May 30, 2014). The MTurk sample included substantially more non-US participants than the other samples, and so limiting the comparison to just US participants allows for a more direct comparison of exclusion rates across samples.
[g] This includes participants run in the lab and through Columbia University's virtual lab participant pool.
[h] Participants were recruited from an unspecified online source.
[i] Subjects were recruited from Colombia University's virtual lab participant pool.
[j] The results summarized here are just for the first two experiments, as their third experiment used an IMC that was so difficult that 74.5% of MTurk participants and 97.8% of subject pool participants failed it, suggesting it is not representative of typical screeners.
[k] Run through Project Implicit (https://implicit.harvard.edu/implicit/). These rates were not reported in the paper, but were communicated to us by Hawkins (C. Hawkins, personal communication, July 17, 2014).
[l] Participants were recruited through Qualtrics Panels.
[m] Participants were recruited from Craigslist Boston or Craigslist New York.
[n] There was one very small moderation effect of their inconsistency measure and a reliability measure, but this finding was the only significant one out of a large number of analyses, not easily interpretable, and the authors dismissed it (see p. 325–326 of their paper for details).
[o] These participants were recruited from Internet discussion boards.
[p] Subjects participated online for course credit.
[q] These rates are not included in the ranges presented in the body of the paper because the screener method was designed to test the limits of attention specifically to compare across sampling methods, and thus are not comparable to other screener methods.

reference to specific sources and/or the following column headings in the table: (1) Screener failure rates, (2) Type of screener, (3) Effect on results, and (4) Differences in excluded participants.

To preview the main takeaways, we find that using screeners to identify problematic participants consistently increases power by eliminating statistical noise, find little consistent evidence that it introduces substantial systematic sampling bias, and discover some general rules of thumb for how best to deal with data from participants that fail screeners. We also find that when similar methods have been employed in the lab or other online environments, screener failure rates tend to be in the same range as those observed on MTurk (see Table 1, Screener failure rates), showing

that while the issue may be more obvious for MTurk studies, it is not unique to MTurk, but is in fact ubiquitous across sampling methods. We conclude the section by discussing optimal exclusion methods—which depend on whether researchers want to assay comprehension or just attention—as well as the importance of assessing correlations between variables of interest and screener failure, and various options for how to balance transparency and comprehensibility when reporting results.

3.1. Exclusion decisions: screener placement and exclusion method

Employing screeners presents researchers with three related decisions: (1) where screeners should be placed in a survey, (2) what kinds of screener to use, and (3) whether exclusion should be done ex ante (preventing participants who fail a screener from completing the study) or ex post (identifying and dealing with participants that failed screeners after data collection is complete). Screeners may be presented to participants before, throughout, or after a survey. Ex post exclusion can be done with any of these options, but typically involves screeners placed at the end or during a survey; whereas, ex ante exclusion requires that screeners be presented either before or during a survey, since by definition, it involves preventing participants from completing all experimental materials. These decisions entail competing costs and benefits, which should be evaluated on a case-by-case basis, but the available research provides some general rules of thumb: (1) Screeners should be as unobtrusive as possible, and ideally presented throughout a study; (2) Comprehension checks tend to be better for screening out problematic participants than attention checks when possible; and, (3) Researchers should assess and report whether screener failure correlates with any variables of interest, and present results stratified by attention or screener failure for transparency when necessary (see Berinsky et al., 2013).

Including attention checks throughout experimental materials probably provides the best assessment of attention, which might wax and wane across a single survey, and is thus most accurately assessed at multiple time points (Berinsky et al., 2013; Maniaci & Rogge, 2014). However, including screeners before or during experimental materials also has the potential to increase self-presentation concerns, facilitate expectancy effects, or lead participants to change their behavior due to an increased focus on otherwise implicit elements of an experiment (Berinsky et al., 2012; Clifford & Jerit, 2015; Rand et al., 2012). For example, Rand et al. (2012) found that giving participants comprehension questions before experimental materials caused them to alter their behavior in ways that were highly relevant to what was being tested. There is also evidence that IMCs can lead to more systematic thinking (Hauser & Schwarz, 2015), though two studies did not find any increases in social desirability concerns or attrition rates (Berinsky et al., 2013; Clifford & Jerit, 2015). However, both of these latter studies found that attempting to increase respondent attention at the outset of a survey by "warning" or "training" participants did lead to such undesirable effects (increased social desirability concerns in Clifford & Jerit, 2015, and higher attrition rates in Berinsky et al., 2013). Furthermore, Clifford and Jerit (2015) found differential effects of the warning manipulations on different groups of interest, implying that obvious screeners presented before experimental materials may also potentially yield false positives or underestimates of effect sizes (depending on the direction of the effect).

Consistently across all of these studies, the salience of a screening procedure to participants seems to be an important factor in whether it affects their behavior. Thus, while presenting screeners throughout experimental materials might provide a more precise metric of attention, if the screeners are too obvious or draw attention to subtle aspects of an experimental design, they may also impact participant behavior, and may even differentially impact behavior in different groups of interest. Clearly more research is needed on this issue, but the available data suggest that screeners presented before important outcome measures have the potential to adversely affect participant behavior.

Presenting screeners after all experimental materials avoids any potential problems with affecting participant behavior during the study, but this approach may miss transient inattention (Berinsky et al., 2013; Maniaci & Rogge, 2014). In contrast, comprehension checks generally assess both comprehension and attention for experimental materials presented throughout an experiment (i.e., participants cannot explicitly comprehend something they don't attend to) so using comprehension checks at the end of an experiment should be able to identify all problematic response patterns, including transient inattention.[4] However, if a researcher is only able to diagnose attention, and not comprehension—an issue we address in more depth below—then the potential benefits of interspersing screeners throughout a survey for a more precise identification of transient inattention must be weighed against the potential costs of affecting participant behavior, and any potential effects on participant behavior should be assessed and reported when possible (this is similar to the recommendation made by Berinsky et al., 2013).

Utilizing attention or comprehension checks placed after experimental materials raises the concern of post-treatment bias by potentially creating post-exclusion imbalance across experimental conditions (see, e.g., Rosenbaum, 1984). Researchers should of course design their experiments to minimize effects of experimental condition on attention or attrition (Zhou & Fishbach, 2016), but such effects may be unavoidable or unobservable. If attention or comprehension can only be assessed post-treatment, then screeners should be identical across experimental conditions and researchers need to be extra vigilant (and transparent) about analyzing and reporting results both pre- and post-exclusion. Additionally, researchers should report exclusion rates and the correlates of exclusion by experimental condition to assess the likelihood of post-treatment bias.

Ex ante exclusion can potentially save researchers time and money, but the incomplete data of excluded participants reduces transparency by making it difficult or impossible to assess any potential effects of exclusion. This is especially problematic because ex ante exclusion requires that screeners are included either before or during a survey, with the concomitant problems outlined above. Ex post exclusion can be done with screeners presented at any point in a survey, and permits assessment and reporting of any observed exclusion effects; however, it also introduces the potential for abuse by adding experimenter degrees of freedom. While the potential problems with ex ante exclusion are difficult to mitigate or even measure, the potential problems with ex post exclusion can be minimized with a disciplined, objective, and transparent usage of attention and/or comprehension checks.

Ex post exclusion allows for screeners to be presented throughout a study, and the potential for data manipulation can be minimized if screeners have objectively right and wrong answers (so-called *gold standard items*, see Chandler et al., 2014) and are used in a clear a priori manner (e.g., excluding everyone that misses any item), as long as researchers report or pre-register all of their methods (a broader issue of scientific integrity that is not specific to

[4] Unlike manipulation checks, which must vary across conditions by definition, comprehension should be assessed in identical fashion across conditions to minimize the potential for post-treatment bias, an issue we address in further detail below.

exclusion). Less stringent criteria for ex post exclusion can also be effective (e.g., excluding participants that miss some number of items determined post hoc, or using items that do not have objectively verifiable answers), but such exclusion criterion should be explicitly justified, and results should be reported both before and after exclusion, to avoid any potential for ad hoc sample manipulation.

Furthermore, including screeners throughout or at the end of an experiment, allows for more sophisticated and transparent reporting of results. As we explain below in section 3.2, screener failure may correlate with independent variables of interest in a particular study. Although these effects generally seem to be small and inconsistent across studies, using ex post exclusion allows researchers to analyze any such potential correlations with screener failure, and report results stratified by screener failure when necessary (as suggested by Berinsky et al., 2013), rather than simply excluding all participants that fail one or more screeners. Especially in research that is highly dependent on sample demographics that might correlate with screener failure (e.g., political science research), stratifying results by screener passage provides an optimal method for balancing transparency and data quality.

Clearly more research is needed on these important issues, but the available evidence provides some guidance on best practices. Screeners should be placed throughout the survey in order to improve estimates of attention. However, screeners given throughout a survey have the potential to alter participant behavior, and so should be as unobtrusive as possible. If a study is particularly sensitive to potentially altered participant behavior (e.g., heightened social desirability might interact with an experimental treatment), then researchers should use fewer screeners placed further from experimental materials. Ex ante exclusion reduces transparency, and would thus seem to necessitate some kind of justification, such as otherwise prohibitive recruitment costs. Thus, when feasible, transparent ex post exclusion with gold standard screeners presented throughout a study seems to provide the optimal balance of costs and benefits, and allows for analysis and presentation of results stratified by screener failure for transparency when applicable.

3.2. Exclusion rates: MTurk vs. other samples

While some scholars have long suspected that MTurk samples provide low quality data, there has been little evidence to support this claim. We review the available evidence below and find that MTurk respondents are often more attentive than respondents from alternative subject pools. However, we also find that MTurk respondents are more aware of formulaic attention checks, suggesting that researchers ought to rely on novel, unobtrusive measures.

Screener failure rates vary widely across MTurk experiments (from 2% to 52%; see Table 1, Screener failure rates-MTurk), depending on the number and difficulty of items used for exclusion. While these rates may seem high to some researchers that are unfamiliar with such methods, they are similar to rates that have been observed in the lab (6%–46%; see Table 1, Screener failure rates-Lab). Comparable rates have also been observed in other commonly used commercial online samples, such as Polimetrix/YouGov, Survey Sampling International, and GfK Knowledge Networks (22%–63%), and various other methods of online recruitment (5%–45%; see Table 1, Screener failure rates-Online). Unsurprisingly, the lowest exclusion rates generally result from a single, obvious attention check, while the highest rates tend to result from multiple, objectively verifiable comprehension checks about experimental materials, or especially arduous or difficult attention checks (for details, compare Screener failure rates and Type of screener columns in Table 1).

Multiple screener items lead to much higher exclusion rates because screeners typically show surprisingly low correlations with each other. Berinsky et al. (2013) found correlations from $r = .38$ to $r = .46$ across four different attention checks (IMCs) in one study session, and cross-wave correlations of $r = .33$ to $r = .39$ for the same screener items presented in another survey two weeks later. They also found that scale scores summing across all four screeners were a better predictor of performance than any single screener (even screeners presented immediately before an experimental task of interest). Maniaci and Rogge (2014) did not report inter-item correlations, but the Cronbach's alpha statistics that they reported for an 11-item scale for inconsistent responding and an 11-item scale for improbable responses ($\alpha = .64$ and $\alpha = .83$, respectively), yield average inter-item correlations of $r = .14$ and $r = .31$, respectively. A reanalysis of a previously published dataset[5] showed correlations between different comprehension checks ranging from $r = .01$ to $r = .58$.

Data from two other previously published studies showed correlations between attention checks ranging from $r = .02$ to $r = .17$[6] and a correlation of $r = .57$[7] between two closely placed IMCs. These surprisingly low and variable inter-item correlations between different screeners lend further support to the view that inattention is best conceptualized as a state-like latent variable that can only be measured with error. Moreover, it highlights the fact that different kinds of problematic responding may produce response patterns that are better captured by different types of screeners (Kurtz & Parrish, 2001; Meade & Craig, 2012). Thus, multiple screeners, and perhaps multiple types of screeners, should be used when possible.

These data suggest that exclusion rates on MTurk are similar to those in the lab or other samples, but there is very little research available that directly compares exclusion rates across samples using the exact same procedure. The few studies that have used the same screeners and methods across different samples provide further support for the claim that MTurk exclusion rates are equivalent to, or possibly even lower than, those observed in other samples. Goodman et al. (2012) observed similar but slightly higher exclusion rates from an MTurk sample than a community sample recruited off the street and a student sample, when they included both native English speakers and ESL participants. However, they did not restrict their MTurk sample to participants in the United States (an option that is easy to implement in MTurk), so their MTurk sample included many more ESL participants than the other two samples, and no significant differences were found across the samples when they were restricted to non-ESL participants, or when the MTurk sample was restricted to just US participants (C. Cryder, personal communication, May 30, 2014). Furthermore, they found that MTurk participants from outside of the US were more likely to be excluded, suggesting that MTurk samples limited to US participants (or other English speaking countries) likely provide higher quality data for survey materials in English, and careful screening may be especially important for international samples.

Paolacci et al. (2010) found no significant differences in exclusion rates between samples recruited from MTurk (4.2%), the lab (6.5%), or other online sources (5.3%). Berinsky et al. (2012) found lower exclusion rates in an MTurk sample (40%) than in samples obtained from high quality online panels (Polimetrix/YouGov—51%; Survey Sampling International—54%), and Berinsky

[5] This data was published in Thomas et al. (2014).
[6] A subset of this data was published in Clifford and Jerit (2014).
[7] Data reported in Clifford and Jerit (2015).

et al. (2013) report greater exclusion rates in Survey Sampling International samples than in MTurk samples across multiple datasets. Similarly, Klein et al. (2014), Ramsey, Thompson, McKenzie, and Rosenbaum (2016), and Hauser and Schwarz (2016) found lower exclusion rates among MTurk participants than for non-MTurk samples (Table 1, Screener failure rates); in fact, this was the main finding of Hauser and Schwarz (2016). Finally, while not reported in their paper, Maniaci and Rogge (2014) did not find higher exclusion rates in their MTurk samples than in their other samples (R. Rogge, personal communication, April 24, 2014).

To add to these previous results, we evaluated how exclusion rates across different samples compare when the same procedure is used. Clifford and Jerit (2014) compared data quality among students who were randomly assigned to take a survey online or in the lab. Here we analyze unreported data from an identical survey conducted on MTurk at the same point in time, allowing us to directly compare attention between MTurk and student samples. Clifford and Jerit (2014) measured attention using an IMC and two bogus items (Table 1, Type of screener), which asked respondents whether they were currently using a computer or electronic device and whether they were taking a survey on politics. The survey also included two comprehension checks about experimental materials. MTurk participants were significantly more likely than the students to pass both bogus items (100% vs. 97%; $p = .004$; 97% vs. 83%; $p < .001$), both comprehension checks (65% vs. 56%; $p = .02$; 56% vs. 42%; $p < .001$), and the IMC (95% vs. 69%; $p < .001$). Overall, the MTurk sample exhibited significantly higher attention and comprehension rates across multiple measures, a finding that holds regardless of whether the comparison is to students online or in the lab.[8]

Our results also provide some evidence that MTurk respondents perform significantly better than students on a common IMC than they do on novel attention checks. Many MTurk participants are familiar with common experimental paradigms (Chandler et al., 2014; Hauser & Schwarz, 2016), and may also be familiar with frequently used IMCs, making them less diagnostic of attention among MTurk participants than they are in other samples. Indeed, a reanalysis of the data for comparable MTurk participants presented in Peer et al. (2013), shows that they were more likely to pass three commonly used screeners (97.4%, from Experiment 1) than they were to pass three novel screeners (77.2%, from Experiment 2), $\chi^2(1, N = 661) = 56.86$, $p < .001$, $\phi = .29$.[9] Peer et al. also found that workers who had completed more MTurk tasks were more likely to pass even novel screeners[10], suggesting that not only do MTurk participants become more familiar with commonly used screeners, but they also become more alert to the screening process in general.

To more directly test how MTurk participants compare to a more naïve sample on common and novel screeners, we further analyzed the data from Clifford and Jerit (2014), using a repeated measures logistic regression with respondent random effects, and included one dummy variable for MTurk vs. student participants, another for type of screener (IMC vs. other), and an interaction term. We found a main effect of sample (Wald $\chi^2(1, N = 686) = 32.02$, $p < .001$), but no effect for type of attention measure (Wald $\chi^2(1, N = 686) = 0.16$, $p = .69$). In line with our hypothesis, we found a significant interaction between sample and measure (Wald $\chi^2(1, N = 686) = 26.11$, $p < .001$), showing that MTurk participants were significantly more likely than the students to pass the common IMC than they were with the other screeners. All of these results suggest that commonly used IMCs are less diagnostic of attention on MTurk than other screeners, presumably due to MTurk participants' familiarity with these items, though we did find higher rates of attention among MTurk participants than among students even with novel screener items.

The similar exclusion rates across samples not only reinforce the idea that inattention is not a bigger problem for MTurk than it is for other samples, but also suggest that the same issues are probably more prevalent in the lab and in other samples than has previously been appreciated. It's possible that the issue has frequently been overlooked in other samples because it is less obvious when an experimenter can directly observe participants or pays a lot of money for a supposedly high-quality sample, both of which may lead to a false confidence in participants' engagement with experimental tasks.

3.3. Exclusion effects: statistical noise vs. sampling bias

Excluding inattentive participants may reduce statistical noise, but it also has the potential to reduce statistical power if exclusion methods are too stringent or introduce sampling bias if screener passage is correlated with respondent characteristics. In this section, we review these issues and contribute new data to these questions. Our findings suggest that exclusion typically reduces statistical noise and only has weak and inconsistent effects on sample composition.

Our review of existing findings is shown in the *Effect on results* column of Table 1.[11] Contrary to some concerns about exclusion, no study reviewed here reported any evidence of exclusion causing pre-exclusion effects to disappear, or producing any new result that was qualitatively different from pre-exclusion patterns (e.g., no effects switched direction or were not at least trending before exclusion). Of the 22 studies that reported comparisons between pre- and post-exclusion results, 10 reported that exclusion had no significant effect on results without any additional details, while 12 reported that exclusion improved results in some way. The improvements in results were as follows[12]: nine studies reported reduced statistical noise (e.g., less variance, fewer outliers, etc.); three studies reported that a trending effect became statistically significant; two studies reported that they were only able to replicate a well-established effect after exclusion; two studies reported that exclusion transformed a qualitative replication of previous results into a quantitative replication; and four studies reported improved psychometric properties of well-established scales. In addition to these studies, we also did not find any qualitative differences in a reanalysis of the Thomas et al. (2014) data, but did find that including data from participants that failed comprehension checks would have consistently weakened the results they presented. This robust pattern across many different kinds of studies suggests that, in general, a disciplined employment of rigorous exclusion methods can effectively reduce statistical noise without introducing any significant systematic sampling bias.

[8] Clifford and Jerit (2014) showed few differences in attention between students in the lab and online. All differences reported above hold whether comparing MTurk to students in the lab or online with one exception—MTurk participants were not significantly more likely to pass one of the comprehension checks than students online (65% vs. 61%, respectively, $p = .31$).

[9] These calculations are based on participants with a 95% approval rating or higher, collapsing across "high" and "low" productivity workers in Experiment 2.

[10] Peer et al. (2013) compare "high productivity" workers (defined as those who had completed more than 500 tasks) with "low productivity" workers (those who had completed less than 100 tasks).

[11] The results from Maniaci and Rogge (2014) are discussed throughout the paper, but are not included in Table 1 because their unique sampling and exclusion methods are not easily comparable to the other studies.

[12] The number of studies reported in this list sum to more than 12 because some studies reported more than one effect of exclusion (see Table 1, Effect on results).

The consistency of these results shows that disciplined exclusion will probably not create any significant problems for most research questions, but exclusion may slightly alter some demographic and personality parameters of a sample (see Table 1, Differences in excluded participants). For example, some studies have found that exclusion may slightly bias samples towards being older, more educated, and more female, and may perhaps lead researchers to underrepresent some minorities (Berinsky et al., 2013; Downs, Holbrook, Sheng, & Cranor, 2010; Maniaci & Rogge, 2014; Ramsey et al., 2016).[13] However, all of these effects were small, Oppenheimer et al. (2009) did not find any differences in age or gender (they did not report results for education or ethnicity), and Shapiro et al. (2013) found instead that Asian participants were more likely to be excluded. The effects of exclusion on personality variables are equally small and mixed. Maniaci and Rogge (2014) found that excluded participants were slightly lower on measures of Agreeableness, Conscientiousness, Openness to Experience, and self-esteem. However, Goodman et al. (2012) did not find any of these effects, but did find that excluded participants were slightly lower on measures of Emotional Stability, and Kurtz and Parrish (2001) did not find any differences for the Big Five traits.

To gain further insight into how excluding participants might bias a sample, we reexamined the effects of exclusion on demographic, personality, and behavioral variables in three previously collected datasets (previous results were published in Clifford & Jerit, 2015 [14]; Clifford & Jerit, 2014; and Thomas et al., 2014). We found no significant differences in age or gender in any of the datasets. While white participants were more likely to pass screeners than non-white participants in the Clifford and Jerit (2014) dataset ($t(687) = 4.60$, $p < .001$, $d = .42$), the dataset from Clifford and Jerit (2015) showed no significant effect for race, although in a related study run in the lab, black participants were less likely to pass an IMC ($\chi^2(1, N = 249) = 5.13$, $p = .02$, $\phi = .14$). There was also a significant effect of self-reported race on number of missed comprehension checks in the Thomas et al. (2014) dataset ($F(4, 780) = 6.26$, $p < .001$, $\eta^2 = .03$), which was due to black participants missing more items than white or Asian participants; Hispanic participants fell somewhere in between, with no significant difference between any of the other three groups, and there were not sufficient sample sizes to reliably examine other ethnic groups. In neither the Clifford and Jerit (2015) nor the Clifford and Jerit (2014) datasets did we find significant differences for education, voter registration, or partisan identification (these were not measured in Thomas et al., 2014); nor were any differences found in income, ideology, church attendance, or voter turnout in the Clifford and Jerit (2015) data (the only dataset that included measures of these variables). We also found no significant differences for four of the Big Five personality traits in the Thomas et al. (2014) dataset, but we did find that participants who were excluded were slightly more Extraverted ($t(799) = 2.03$, $p = .043$, $d = .14$), which is especially notable because other studies reviewed above found various effects for the other four Big Five traits, but not for Extraversion, suggesting that all of these observed effects may be spurious.

Differences in participant motivations may be an important factor in how exclusion affects sample composition (Maniaci & Rogge, 2014), and these effects may differ by sample. The Clifford and Jerit (2014) dataset included both student and MTurk participants, allowing for a comparison of the correlates of attention across samples with different motivations. MTurk participants are motivated at least in part by monetary compensation and the desire to maintain a reputation as a quality worker (Paolacci et al., 2010), both of which depend on the quality of the data they produce. In contrast, students typically have external motivations to participate, but little external motivation to produce quality data. As a result, students may be more dependent on purely internal motivations to produce accurate data (such as a desire to further scientific research), motivations that may be weak or non-existent for many students. Consistent with this hypothesis, we found that risk aversion, which relates to the external outcomes specific to MTurk participants, is positively correlated with attention among MTurk participants ($r(251) = .16$, $p = .01$), but not among students ($r(435) = .04$, $p = .44$). In contrast, we found that the internal motivations of Need to Evaluate and political interest are both positively correlated with attention (in political science research) among students ($r(433) = .10$, $p = .04$; $r(433) = .17$, $p < .001$, respectively), but not among MTurk participants ($r(249) = .03$, $p = .65$; $r(249) = .08$, $p = .23$, respectively). Lastly, a reanalysis of data from another student sample[15] showed that IMC passage was associated with greater attention to foreign policy news ($t(256) = 3.02$, $p = .002$, $d = .40$) and higher GPA ($t(250) = 3.20$, $p = .002$, $d = .40$), two other variables that might be related to the presumed internal motivations of students; however, no statistically significant association was found between attention and GPA in the Clifford and Jerit (2014) data ($r(438) = .07$, $p = .12$) (unfortunately, attention to foreign policy news was not measured in this dataset).

While excluding inattentive participants may reduce the personality and demographic diversity of the sample, ignoring inattention can lead to problems beyond increased noise. For example, political knowledge is frequently measured and used as a moderator of treatment effects in political science experiments (e.g., Arceneaux, 2008; Kam, 2005; Lau & Redlawsk, 2001). Yet, we find that attention is correlated with political knowledge scores ($r(687) = .17$, $p < .001$), which may be because inattentive respondents devote less effort to political knowledge questions, leading to lower scores, or because less knowledgeable respondents are simply less motivated to be attentive.[16] Because both mechanisms could lead to spurious results, such as smaller treatment effects among those scored as less knowledgeable, the best solution would seem to be excluding inattentive participants (or presenting results stratified by screener passage).

Such findings are far from conclusive, but they suggest that the effects of exclusion on sample characteristics may vary based on the motivations of participants within the sample. They also raise the concern that exclusion criteria may result in participants that are more politically engaged than the initial sample (at least in research on political issues), and that this effect may be more pronounced for student samples. More broadly, these findings suggest that attentiveness may be correlated with interest in the topic being studied, not just political knowledge.

Taken together, the contradictory findings across studies and small effect sizes suggest that any systematic effects of exclusion are likely small and inconsistent, and that some of the observed effects may have simply been due to sampling and/or measurement

[13] These findings should be interpreted with caution, because participants that are excluded for lack of attention or comprehension are likely to give unreliable responses to other items as well, which makes any differences between excluded and non-excluded participants difficult to assess (Maniaci & Rogge, 2014; Oppenheimer et al., 2009; Peer et al., 2013).

[14] Because attention was experimentally manipulated in this study, here we only report data from the control condition.

[15] Data reported in Clifford and Jerit (2015).

[16] This data comes from Clifford and Jerit (2014) and includes students in the lab and online, and MTurk participants.

error. However, three effects of exclusion were relatively consistent across studies: Minorities and males were more likely to be excluded, and exclusion rates seem to track variables related to participants' motivations for participating in a study. Yet, the effect sizes of exclusion along even these dimensions were also consistently very small. While the MTurk population may be slightly different than the general population on some demographic and personality factors, it is substantially more representative than other convenience samples, and will likely yield much more representative samples than lab studies, even after exclusion. Outside of MTurk, screeners may differentially exclude respondents with weaker internal motivations to provide high quality data, such as those who are less interested or less opinionated in the topic of the study.

The small potential effects of exclusion on sample characteristics should not cause problems for most social science research, but researchers investigating questions that may be sensitive to small variations in sample parameters (e.g., political science research that depends on extremely representative samples) may need to exercise more caution.[17] In such cases, researchers should present both pre- and post-exclusion results, and allow readers to judge for themselves, as suggested by Berinsky et al. (2013). Additionally, when relying on a national sample, researchers may want to request that the survey vendor create an additional set of survey weights for the post-exclusion sample and present both pre- and post-exclusion results with the appropriate weights.

In summary, using screener items to exclude problematic participants seems to increase statistical power by eliminating statistical noise, without introducing any significant sampling bias that would affect most research questions. These findings should be interpreted with caution because the available data is limited, and inattentive respondents are more likely to give unreliable responses to any measure; however, they also highlight the importance of addressing inattention, particularly when it may be confounded with critical measures of interest. In such cases, we agree with the recommendation of Berinsky et al. (2013), that the best practice is to present results stratified by screener failure, or at least provide such data in an appendix or supplemental materials for full transparency. We conclude this section with the typical call for more research on these issues, and by noting that any study that may be highly sensitive to such parameters should report results both before and after exclusion.

3.4. Optimizing exclusion: attention vs. comprehension

The benefits of using screeners seem to hold for all observed screener methods, exclusion rates, and samples; however, simple approaches to exclusion may lead to larger reductions in sample size than are necessary to achieve these benefits (compare Screener failure rates and Type of screener columns in Table 1). Employing more sophisticated measurements and exclusion criteria may help researchers selectively identify and eliminate only those participants who are truly inattentive, and avoid larger reductions in sample size than are necessary (Kurtz & Parrish, 2001; Maniaci & Rogge, 2014; Meade & Craig, 2012).

Everyone has lapses of attention, misunderstands directions, or skims instructions at least some of the time, and screeners are imperfect measurement tools that entail substantial measurement error (Berinsky et al., 2013; Maniaci & Rogge, 2014). Thus, excluding anyone who misses any screener may avoid increasing experimenter degrees of freedom and post hoc sample manipulation, but this criterion may be stricter than is necessary for some kinds of studies.[18] High exclusion rates may be unavoidable and unproblematic in research where comprehension of complex experimental materials is imperative and an experimenter cannot answer questions or provide clarification (e.g., economic games run on MTurk). However, if a researcher only wishes to weed out inattentive or careless participants, more precise measures can be used to optimize the tradeoff between maximizing sample size and eliminating problematic response patterns.

Some experiments require that participants understand a somewhat complicated task, and it can be difficult or impossible in MTurk studies to address participant confusions or answer participants' questions (Chandler et al., 2014; Horton et al., 2011; Paolacci et al., 2010; Rand, 2012). For example, many experimental economics paradigms can be somewhat difficult to understand in the cursory read they get from many participants, yet internal validity in these experiments critically depends on participants' full comprehension of the task they are engaging in (see Rand, 2012). In such experiments, any uncomprehending participant is essentially not participating in the experiment that the researcher is interested in. Researchers should obviously design experimental materials to be as comprehensible as possible, and if exclusion rates are at the high end of the rates reported here (consult Table 1, Screener failure rates), they should consider trying to make materials more comprehensible before proceeding. However, even with well-designed materials, such experiments may have higher than average exclusion rates, and the available data suggest that the only downside to this will likely be a large reduction in sample size (Table 1, Effect on results). Yet, even with high exclusion rates, recruiting large post-exclusion samples will probably still be substantially quicker, easier, and cheaper through MTurk than in the lab, and the final sample will almost certainly be much more representative than most other convenience samples. Such experiments should employ rigorous comprehension checks, exclude anyone who fails any item, recruit based on projected post-exclusion sample sizes, and simply consider sunk costs from exclusion as part of the cost of running the experiment. If sampling bias is a concern, researchers should report results stratified by screener failure, as recommended by Berinsky et al. (2013).

Outside of experimental economics research, perfect comprehension of all experimental materials and tasks may be impossible to measure (e.g., priming studies), and excluding data from inattentive or careless participants may be the only objective. The research reviewed here shows that excluding all participants who miss any attention check may not catch all problematic respondents if exclusion items are too easy, or may exclude more participants than necessary if exclusion criteria are overly stringent. Fortunately, the more precise measures of problematic responding outlined in Kurtz and Parrish (2001), Maniaci and Rogge (2014), and Meade and Craig (2012) can help researchers optimize the tradeoff between maximizing sample size and excluding problematic respondents. Their methods seem to offer the best available exclusion options for studies in which comprehension cannot be adequately measured, and especially for such studies in which participants are costly or difficult to recruit. While the specific samples and methods they each used were different, all three found exclusion rates of around 10%, suggesting that this may be a reasonable baseline estimate of inattentive or careless response rates that researchers

[17] However, a recent examination of this assumption about political science research finds little evidence of treatment effect heterogeneity among 40 different treatment effects (Coppock, 2016).

[18] Again, here we use "excluding" as short-hand for multiple ways to manage data from participants that fail screeners, and the same analysis applies if results are stratified by screener failure, as suggested by Berinsky et al. (2013).

should generally anticipate. Furthermore, in a comparison with less sophisticated exclusion criteria, Maniaci and Rogge (2014) found that not only did their short scales eliminate fewer participants (less than 10% vs. 26%), but they also led to higher average power gains in a resampling analysis of their data. These more precise methods thus seem ideal for experiments in which screening for attention is the primary objective, and sample size is a major concern (although we note that screener items presented throughout a survey should be as similar as possible to experimental materials in both content and structure to avoid detection and affecting participant behavior).

4. Summary and recommendations

MTurk provides a valuable tool for social science researchers to quickly, easily, and cheaply recruit larger and more representative samples of participants than can be recruited in the lab, and the tools for researchers are getting better all the time. While the physical isolation of MTurk participants presents obvious concerns about both internal and external validity, the available research shows that both of these concerns are manageable, and provides a useful guide for best practices.

4.1. Internal validity and interactive experiments

Interactive experiments run through MTurk seem to be just as internally valid as those run in the lab, as long as experimental designs are credible within the MTurk framework. This latter point is important because the platform has certain constraints that MTurk participants are probably more familiar with than most researchers, such as how participants might plausibly be paired up with potential partners, or how messages could actually be transmitted between partners. Researchers must be aware of these constraints to ensure that the mechanics of interactive experiments are believable, especially if an experiment involves deception as some research suggests MTurk participants may generally be more savvy to deception than lab participants (e.g., Krupnikov & Levine, 2014; Paolacci et al., 2010). The only recommendations we have for running interactive experiments on MTurk are that researchers should familiarize themselves with the mechanics of the platform, avoid deception if possible, and include measures to assess believability or suspicion (e.g., funnel debriefing).

4.2. External validity, screeners, and participant exclusion

In none of the studies reviewed here did ex post exclusion of careless, uncomprehending, or inattentive participants introduce any substantive sampling bias (Table 1, Effect on results), which suggests that rigorous exclusion can help researchers improve external validity by eliminating or controlling for unreliable respondents and statistical noise. The only potential exceptions to this general pattern were that minorities and participants with certain kinds of motivations were consistently more likely to be excluded (Table 1, Differences in excluded participants). However, the effect sizes of these differences were so small that they would be negligible for most research; yet, researchers investigating questions that might be highly sensitive to these variables should exercise especial caution, and report results stratified by screener passage as recommended by Berinsky et al. (2013).

The comparable exclusion rates and effects observed across many different kinds of samples also highlight the fact that these issues are not unique to MTurk (Table 1, Screener failure rates), and suggest that the ability to directly oversee participants in the lab may sometimes give researchers a false confidence in participants' level of engagement with experimental tasks. Thus, just as with interactive experiments, participants' physical isolation does not seem to pose any problems with inattention or comprehension that are unique to MTurk, and the data suggest that assessing participant engagement seems to be just as important in the lab as it is on MTurk. Existing research also provides a guide for best exclusion practices:

1. *All studies should screen for problematic responders.* Every study with human participants is susceptible to problematic responding—whether run through MTurk or elsewhere—and thus, all studies should include measures to weed out such data when feasible or present results stratified by screener passage.
2. *Ex post exclusion is generally better than* ex ante *exclusion.* While ex post exclusion has the potential to introduce additional experimenter degrees of freedom, this can easily be managed by using rigorous and transparent exclusion methods, and/or by presenting results stratified by screener failure (see, Berinsky et al., 2013). In contrast, ex ante exclusion is less transparent and important differences between those that pass or fail cannot be assessed.
3. *Exclusion should be transparent and based on objective criteria.* Exclusion of all participants that miss items with objectively verifiable right and wrong answers (or is done according to pre-registered methods) precludes the potential for data manipulation and adding experimenter degrees of freedom. If more subjective criteria are used for exclusion (e.g., empirically deriving an exclusion threshold), then the methods should be transparent and as objective as possible (e.g., presenting histograms to justify thresholds, presenting results stratified by screener failure, etc.).
4. *Screening methods should be tailored to each individual study.* If comprehension is measurable, then comprehension checks should be used, and these should be identical across conditions to avoid potential confounding. Comprehension checks can also identify other problematic response patterns, and may thus provide a comprehensive metric for identifying problematic data. However, if comprehension cannot feasibly be measured, researchers should employ multiple types of items or scales to precisely measure different kinds of problematic responding. If the research question is likely to be especially sensitive to demographic characteristics of the sample, then researchers should assess and present any correlates with screener failure, and/or present results stratified by screener failure (or at least include this information in supplemental materials).
5. *Multiple screeners should be used.* Problematic response patterns are best characterized as state-like latent constructs, every screener item provides only imperfect measurement of such latent variables, and different kinds of problematic responding yield different diagnostic data patterns. Thus, researchers should utilize multiple items, and, if they do not measure comprehension, multiple types of items to identify different kinds of problematic response patterns.
6. *Screeners should be as similar to study materials as possible.* Using screeners that are similar to other study materials in both structure and content prevents easy detection, minimizes post-exclusion sampling bias, and captures relevant dimensions of attention, comprehension, and/or carelessness. This will also help prevent any undesirable effects of exposure to the screeners on participant behavior.
7. *Scales for identifying problematic response patterns provide the most precise metrics of attention.* In some studies it is not feasible to include an entire scale for measuring problematic

responses, and the research shows that even adding a couple of screeners is useful. However, if a researcher wishes to maximize post-exclusion sample sizes, and has room, the scales and methods presented in Maniaci and Rogge (2014), Meade and Craig (2012), and Kurtz and Parrish (2001) provide the most accurate tools currently available for precisely identifying just the most problematic responders.

8. *Screener items should be novel.* MTurk participants are likely to be familiar with commonly used screeners (e.g., certain formulaic IMCs used by many researchers), and thus, if robust scales are not used, novel screener items will likely provide the best assay of problematic responding. This recommendation has an interesting parallel with Campbell's Law in psychology, which states that when a standardized test becomes established, scores become less indicative of the latent construct they are supposed to measure (e.g., teachers "teach to the test"), and Goodhart's Law in macroeconomics, which states that once a new economic indicator becomes publicized it immediately begins to lose informational value by being directly targeted by policy makers and investors, which decreases its correlation with the latent economic factor to which it was initially related.
9. *Representative samples should be re-weighted after exclusion when possible.* Researchers utilizing representative samples may be reluctant to exclude participants and potentially bias the composition of their sample. However, given that high quality national samples showed levels of inattention comparable to many convenience samples, and that inattention was correlated with potential variables of interest, researchers should be aware that inattention may still compromise key measures or experimental manipulations. When possible, researchers using representative samples should request that the survey vendor also estimate weights after excluding inattentive respondents in order to maintain representativeness. Researchers should then report results both pre- and post-exclusion while using the appropriate weights for each.
10. *The effects of exclusion should be analyzed, and reported when relevant.* This final recommendation is perhaps the most important, because it is critical for implementing all of the other recommendations with scientific integrity. Results should always be analyzed both before and after exclusion, and researchers should also check for, and report, any effects of exclusion on results or sample characteristics. If no effects are found with rigorous, objective, and transparent exclusion criteria, this should be briefly and explicitly stated. However, if researchers use more subjective measures of exclusion (e.g., determining a cutoff threshold post hoc), everything should be reported—even null effects—to ensure transparency.

We conclude by noting that the insights presented here are limited by the availability of reported data on these issues. Further research will be needed to reaffirm the conclusions of this paper, as well as to address some interesting outstanding questions, like whether different kinds of screeners or exclusion methods might have different effects on a sample, or whether any effects of exclusion are different across sampling methods. Continued evaluation of the quality of survey and sampling methods, as well as the quality of data collected from convenience samples such as Mechanical Turk, depends on the full and transparent reporting of results. Thus, our final recommendation is that researchers report methodological details and data that will help improve our growing understanding of these important issues.

References

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99*(3), 436–451. http://dx.doi.org/10.1037/a0020218.

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

Arceneaux, K. (2008). Can partisan cues diminish democratic accountability? *Political Behavior, 30*(2), 139–160. http://dx.doi.org/10.1007/s11109-007-9044-7.

Ashton-James, C. E., Kushlev, K., & Dunn, E. W. (2013). Parents reap what they sow: Child-centrism and parental well-being. *Social Psychological and Personality Science, 4*(6), 635–642. http://dx.doi.org/10.1177/1948550613479804.

Ausderan, J. (2014). How naming and shaming affects human rights perceptions in the shamed country. *Journal of Peace Research, 51*(1), 81–95. http://dx.doi.org/10.1177/0022343313510014.

Barone, M. J., Lyle, K. B., & Winterich, K. P. (2014). When deal depth doesn't matter: How handedness consistency influences consumer response to horizontal versus vertical price comparisons. *Marketing Letters*, 1–11. http://dx.doi.org/10.1007/s11002-013-9276-8.

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*, 351–368. http://dx.doi.org/10.1093/pan/mpr057.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2013). Separating the shirkers from the workers? Making sure respondents pay attention to self-administered surveys. *American Journal of Political Science*. http://dx.doi.org/10.1111/ajps.12081.

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. http://dx.doi.org/10.1177/1745691610393980.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*, 2156–2160. http://dx.doi.org/10.1016/j.chb.2013.05.009.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaivete among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods, 46*, 112–130. http://dx.doi.org/10.3758/s13428-013-0365-7.

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science, 1*(2), 120–131. http://dx.doi.org/10.1017/xps.2014.5.

Clifford, S., & Jerit, J. (2015). Do attempts to improve respondent attention increase social desirability bias? *Public Opinion Quarterly, 79*(3), 790–802. http://dx.doi.org/10.1093/poq/nfv027.

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics, 2*, 1–9. http://dx.doi.org/10.1177/2053168015622072.

Coppock, A. (2016). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. In *Presentation at the 74th midwest political science association annual conference*. Retrieved from https://alexandercoppock.files.wordpress.com/2016/02/coppock_generalizability2.pdf.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One, 8*(3), e57410. http://dx.doi.org/10.1371/journal.pone.0057410.

Cryder, C. E., Loewenstein, G., & Scheines, R. (2013). The donor is in the details. *Organizational Behavior and Human Decision Processes, 120*(1), 15–23. http://dx.doi.org/10.1016/j.obhdp.2012.08.002.

Cryder, C. E., Loewenstein, G., & Seltman, H. (2013). Goal gradient in helping behavior. *Journal of Experimental Social Psychology, 49*(6), 1078–1083. http://dx.doi.org/10.1016/j.jesp.2013.07.003.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010, April). Are your participants gaming the system? Screening Mechanical Turk workers. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 2399–2402). ACM. http://dx.doi.org/10.1145/1753326.1753688.

Gaither, S. E., Wilton, L. S., & Young, D. M. (2013). Perceiving a presidency in black (and white): Four years later. *Analyses of Social Issues and Public Policy, 14*(1), 7–21. http://dx.doi.org/10.1111/asap.12018.

Gipson, J., Kahane, G., & Savulescu, J. (2013). Attitudes of lay people to withdrawal of treatment in brain damaged patients. *Neuroethics, 7*, 1–9. http://dx.doi.org/10.1007/s12152-012-9174-4.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213–224. http://dx.doi.org/10.1002/bdm.1753.

Goodman, J. K., & Irmak, C. (2013). Having versus consuming: Failure to estimate usage frequency makes consumers prefer multi-feature products. *Journal of Marketing Research, 50*(1), 44–54. http://dx.doi.org/10.1509/jmr.10.0396.

Gray, K., Knobe, J., Sheskin, M., Bloom, P., & Feldman Barrett, L. (2011). More than a body: Mind perception and the nature of objectification. *Journal of Personality and Social Psychology, 101*(6), 1207–1220. http://dx.doi.org/10.1037/a0025883.

Gromet, D. M., & Darley, J. M. (2011). Political ideology and reactions to crime victims: Preferences for restorative and punitive responses. *Journal of Empirical Legal Studies, 8*(4), 830–855. http://dx.doi.org/10.1111/j.1740-1461.2011.01242.x.

Halko, S., & Kientz, J. A. (2010). Personality and persuasive technology: An

exploratory study on health-promoting mobile applications. *Persuasive Technology*, 150–161.
Hardisty, D. J., Frederick, S., & Weber, E. U. (2012). Dread looms larger than pleasurable anticipation. In *Behavior decision research in management conference, Boulder, CO*.
Hardisty, D. J., Thompson, K. F., Krantz, D. H., & Weber, E. U. (2013). How to measure time preferences: An experimental comparison of three methods. *Judgment and Decision Making, 8*(3), 236–249.
Hauser, D. J., & Schwarz, N. (2015). It's a trap! Instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open, 5*, 1–6. http://dx.doi.org/10.1177/2158244015584617.
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. http://dx.doi.org/10.3758/s13428-015-0578-z.
Hawkins, C. B., & Nosek, B. A. (2012). Motivated independence? Implicit party identity predicts political judgments among self-proclaimed independents. *Personality and Social Psychology Bulletin, 38*(11), 1437–1452. http://dx.doi.org/10.1177/0146167212452313.
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. http://dx.doi.org/10.1007/s10683-011-9273-9.
Huff, C., & Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research and Politics, 2*(3), 1–7. http://dx.doi.org/10.1177/2053168015604648.
Ipeirotis, P. (2010, March 9). *The new demographics of Mechanical Turk*. Retrieved from http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html.
Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3): A brief measure of dark personality traits. *Assessment, 21*(1), 28–41. http://dx.doi.org/10.1177/1073191113514105.
Kam, C. (2005). Who toes the party line? Cues, values, and individual differences. *Political Behavior, 27*(2), 163–182. http://dx.doi.org/10.1007/s11109-005-1764-y.
Karelaia, N., & Keck, S. (2013). When deviant leaders are punished more than non-leaders: The role of deviance severity. *Journal of Experimental Social Psychology, 49*(5), 783–796. http://dx.doi.org/10.1016/j.jesp.2013.04.003.
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahnik, S., Bernstein, M. J., et al. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. http://dx.doi.org/10.1027/1864-9335/a000178.
Komarov, S., Reinecke, K., & Gajos, K. Z. (2013, April). Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 207–216). ACM.
Kross, E., Bruehlman-Senecal, E., Park, J., Burson, A., Dougherty, A., Shablack, H., et al. (2014). Self-talk as a regulatory mechanism: How you do it matters. *Journal of Personality and Social Psychology, 106*(2), 304–324. http://dx.doi.org/10.1037/a0035173.
Krupnikov, Y., & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science, 1*, 59–80. http://dx.doi.org/10.1017/xps.2014.7.
Kteily, N., Cotterill, S., Sidanius, J., Sheehy-Skeffington, J., & Bergh, R. (2014). 'Not one of us': Predictors and consequences of denying ingroup characteristics to ambiguous targets. *Personality and Social Psychology Bulletin, 40*(10), 1231–1247. http://dx.doi.org/10.1177/0146167214539708.
Kugler, M. B., Cooper, J., & Nosek, B. A. (2010). Group-based dominance and opposition to equality correspond to different psychological motives. *Social Justice Research, 23*(2–3), 117–155. http://dx.doi.org/10.1007/s11211-010-0112-5.
Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO-PI-R. *Journal of Personality Assessment, 76*(2), 315–332. http://dx.doi.org/10.1207/S15327752JPA7602_12.
Kushlev, K., Dunn, E. W., & Ashton-James, C. E. (2012). Does affluence impoverish the experience of parenting? *Journal of Experimental Social Psychology, 48*(6), 1381–1384. http://dx.doi.org/10.1016/j.jesp.2012.06.001.
Lau, R. R., & Redlawsk, D. P. (2001). Advantages and disadvantages of cognitive heuristics in political decision making. *American Journal of Political Science, 45*(4), 951–971. http://dx.doi.org/10.2307/2669.
Leeper, T. J. (2016). Crowdsourced data preprocessing with R and Amazon Mechanical Turk. *The R Journal, 8*(1), 276–288. http://dx.doi.org/10.5281/zenodo.33595.
Leeper, T. J., & Thorson, E. (2015). Minimal sponsorship-induced bias in web survey data. In *Paper presented at the 2015 annual meeting of the midwest political science association, Chicago, IL*.
Løhre, E., & Teigen, K. H. (2014). How fast can you (possibly) do it, or how long will it (certainly) take? Communicating uncertain estimates of performance time. *Acta Psychologica, 148*, 63–73. http://dx.doi.org/10.1016/j.actpsy.2014.01.005.
Lombrozo, T., & Rehder, B. (2012). Functions in biological kind classification. *Cognitive Psychology, 65*(4), 457–485. http://dx.doi.org/10.1016/j.cogpsych.2012.06.002.
Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality, 48*, 61–83. http://dx.doi.org/10.1016/j.jrp.2013.09.008.
Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23. http://dx.doi.org/10.3758/s13428-011-0124-6.
Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods, 17*(3), 437–455. http://dx.doi.org/10.1037/a0028085.
Meier, B. P., Moller, A. C., Chen, J. J., & Riemer-Peltz, M. (2011). Spatial metaphor and real estate: North-south location biases housing preference. *Social Psychological and Personality Science, 2*(5), 547–553. http://dx.doi.org/10.1177/1948550611401042.
Menatti, A. R., Smyth, F. L., Teachman, B. A., & Nosek, B. A. (2011). Reducing stigma toward individuals with mental illnesses: A brief, online intervention. *Stigma Research and Action, 1*(2), 9–21. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3839682/.
Meyvis, T., Goldsmith, K., & Dhar, R. (2012). The importance of the context in brand extension: How pictures and comparisons shift consumers' focus from fit to quality. *Journal of Marketing Research, 49*(2), 206–217. http://dx.doi.org/10.1509/jmr.08.0060.
Moran, J. M., Rain, M., Page-Gould, E., & Mar, R. A. (2014). Do I amuse you? Asymmetric predictors for humor appreciation and humor production. *Journal of Research in Personality, 49*, 8–13. http://dx.doi.org/10.1016/j.jrp.2013.12.002.
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. http://dx.doi.org/10.1017/XPS.2015.19.
Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660. http://dx.doi.org/10.1177/1745691612462588.
Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisfying to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872. http://dx.doi.org/10.1016/j.jesp.2009.03.009.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411–419.
Paulhus, D. L., & Carey, J. M. (2011). The FAD-Plus: Measuring lay beliefs regarding free will and related constructs. *Journal of Personality Assessment, 93*(1), 96–104. http://dx.doi.org/10.1080/00223891.2010.528483.
Peer, E., Vosgerau, J., & Acquisti, A. (2013). Reputation as a sufficient condition for quality on Amazon Mechanical Turk. *Behavior Research Methods*, 1–9. http://dx.doi.org/10.3758/s13428-013-0434-y.
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. http://dx.doi.org/10.1016/j.chb.2015.12.049.
Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172–179. http://dx.doi.org/10.1016/j.jtbi.2011.03.004.
Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427–430. http://dx.doi.org/10.1038/nature11467.
Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society. Series a (General), 147*(5), 656–666. http://dx.doi.org/10.2307/2981697.
Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2010). Who are the Turkers? Worker demographics in Amazon Mechanical Turk. In *Presented at the ACM CHI conference*. Retrieved from http://www.ics.uci.edu/~jwross/pubs/SocialCode-2009-01.pdf.
Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior, 43*, 304–307. http://dx.doi.org/10.1016/j.chb.2014.11.004.
Sayeed, A. B., Rusk, B., Petrov, M., Nguyen, H. C., Meyer, T. J., & Weinberg, A. (2011, June). Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption. In *proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities* (pp. 69–77). Association for Computational Linguistics.
Schuldt, J. P., & Roh, S. (2014). Of accessibility and applicability: How heat-related cues affect belief in "global warming" versus "climate change". *Social Cognition, 32*(3), 217–238. http://dx.doi.org/10.1521/soco.2014.32.3.217.
Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science, 1*(2), 213–220. http://dx.doi.org/10.1177/2167702612469015.
Simmons, J. P., & Nelson, L. D. (2006). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General, 135*(3), 409–428. http://dx.doi.org/10.1037/0096-3445.135.3.409.
Sirota, M., Kostovičová, L., & Juanchich, M. (2013). The effect of iconicity of visual displays on statistical reasoning: Evidence in favor of the null hypothesis. *Psychonomic Bulletin & Review*, 1–8. http://dx.doi.org/10.3758/s13423-013-0555-4.
Stiller, A., Goodman, N. D., & Frank, M. C. (2011). Ad-hoc scalar implicature in adults and children. In *Proceedings of the 33rd annual cognitive science society meeting, Boston, July*.
Summerville, A., & Chartier, C. R. (2013). Pseudo-dyadic "interaction" on Amazon's Mechanical Turk. *Behavior Research Methods, 45*(1), 116–124. http://dx.doi.org/10.3758/s13428-012-0250-9.

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS One, 6*(3), e16836. http://dx.doi.org/10.1371/journal.pone.0016836.

Sussman, A. B., & Alter, A. L. (2012). The exception is the rule: Underestimating and overspending on exceptional expenses. *Journal of Consumer Research, 39*(4), 800–814. http://dx.doi.org/10.1086/665833.

Thomas, K. A., DeScioli, P., Haque, O. S., & Pinker, S. (2014). The psychology of coordination and common knowledge. *Journal of Personality and Social Psychology, 107*(4), 657–676. http://dx.doi.org/10.1037/a0037037.

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods, 18*(3), 320–334. http://dx.doi.org/10.1037/a0032121.

White, A., Strezhnev, A., Lucas, C., Kruszewska, D., & Huff, C. (2016). *Investigator characteristics and respondent behavior in online surveys*. Retrieved from http://christopherlucas.org/files/PDFs/MTurkaudit.pdf.

Yang, Y., Vosgerau, J., & Loewenstein, G. (2013). Framing influences willingness to pay but not willingness to accept. *Journal of Marketing Research, 50*(6), 725–738. http://dx.doi.org/10.1509/jmr.12.0430.

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. http://dx.doi.org/10.1037/pspa0000056.