Article

The MTurkification of Social and Personality Psychology

Personality and Social
Psychology Bulletin
2019, Vol. 45(6) 842–850
© 2018 by the Society for Personality
and Social Psychology, Inc
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146167218798821
pspb.sagepub.com
SAGE

Craig A. Anderson[1], Johnie J. Allen[1], Courtney Plante[1], Adele Quigley-McBride[1], Alison Lovett[1], and Jeffrey N. Rokkum[1]

Abstract

The potential role of brief online studies in changing the types of research and theories likely to evolve is examined in the context of earlier changes in theory and methods in social and personality psychology, changes that favored low-difficulty, high-volume studies. An evolutionary metaphor suggests that the current publication environment of social and personality psychology is a highly competitive one, and that academic survival and reproduction processes (getting a job, tenure/promotion, grants, awards, good graduate students) can result in the extinction of important research domains. Tracking the prevalence of brief online studies, exemplified by studies using Amazon Mechanical Turk, in three top journals (*Journal of Personality and Social Psychology, Personality and Social Psychology Bulletin, Journal of Experimental Social Psychology*) reveals a dramatic increase in their frequency and proportion. Implications, suggestions, and questions concerning this trend for the field and questions for its practitioners are discussed.

Keywords

MTurk, research methods/assessment, philosophy of science, theory

Received April 18, 2018; revision accepted July 27, 2018

We begin with a story of the death of an area of research that once was at the heart of social psychology—the area known as group dynamics. Before the remaining group dynamics researchers madly rush off to send emails to us and to the editors of this journal expressing your partially justified outrage that, in the immortal words from the film *Monty Python's in Search of the Holy Grail*, "I'm not dead yet!" please note that we are on your side, and we acknowledge that your species of scholar is not entirely extinct.

There are multiple reasons for shifts in the research foci and in the research methods of any given scientific field, and we will not clutter this simple article with lengthy discussions of the philosophy or the history or the sociology of science. Instead, we highlight a set of rather mundane and largely social decision-making processes that have a huge and perhaps underappreciated impact on the direction of our (or any other) field of science. Specifically, we refer to the processes by which young scholars (a) get (or fail to get) research-related jobs in academia, (b) get (or fail to get) tenure, (c) get (or fail to get) major grants, and (d) train (or fail to train) the next generation of scholars. This is an evolutionary process, one in which some species (domains of research) thrive and reproduce, whereas others go extinct.

Intellectual Reproduction

We can easily trace our intellectual genealogy back to Leon Festinger, Stanley Schachter, Kurt Lewin, Fritz Heider, and so on. But frankly, most of the giants of early social and personality psychology (SPP) would not get a job in today's academic job market, nor in the job market that has existed since the early 1980s. The reason? They did not have enough top-tier journal articles, comparable with today's *Journal of Personality and Social Psychology* (*JPSP*), *Journal of Experimental Social Psychology* (*JESP*), or *Personality and Social Psychology Bulletin* (*PSPB*). Increasing competition for research-oriented academic jobs has also meant increasing pressure to publish empirical articles in these (and other) top journals. During the same era that the SPP environmental niche first became extremely competitive, the cognitive revolution provided higher power methodologies, making it easier to conduct and get significant results in short-duration

[1] Iowa State University, Ames, USA

Corresponding Author:
Craig A. Anderson, Department of Psychology, Iowa State University, 901 Stange Rd., Ames, IA 50011-1041, USA.
Email: caa@iastate.edu

studies with small sample sizes. Thus, it became possible for a SPP researcher to conduct half a dozen or more studies per year on the same topic. Indeed, one top social cognition scholar of the 1980s was able to publish five *JPSP* articles in the same year, each with multiple small-sample, easy-to-run studies. And, multiple studies quickly became a major factor for publishing an article in our top journals. For example, the average number of studies per article in *JPSP* increased from 1.27 in 1968 to 1.58 in 1978 and 1.78 in 1988 (Reis & Stiller, 1992), and our data indicate that this trend has continued.

Even today, it is impossible to run a half dozen or more group dynamics studies in a given year, especially if one is using real existing groups of people and measuring actual group interactions. It is extremely difficult even if the researcher creates temporary groups who meet only a few times in the lab. The result is that researchers who stake their careers on high-difficulty, low-volume research questions cannot effectively compete for limited journal space in our highest-status journals. We highlight the group dynamics domain merely as an exemplar of many hard-to-study domains that are key to the history and future of SPP as a truly worthwhile enterprise. We could have focused just as easily on interpersonal relations among intimate partners; development and change of interracial attitudes and prejudice; and anger, aggression, and violence between individuals, groups, and nations. Any phenomenon or psychological process that requires face-to-face interactions among people who are experiencing ongoing emotions, arousal, and complex decision processes requires labor-intensive research procedures, and often substantial sample sizes to detect small effects.[1] Such studies could not (and cannot) compete for journal space with easy-to-run big-effect/small-sample phenomena that characterized the early social cognition/social memory movement.

This problem was not entirely unnoticed in the past. One of the reasons (and there were others) for *JPSP* creating separate sections for studies of Interpersonal Relations and Group Processes and for Personality Processes and Individual Differences was to keep at a least a few pages available for subsets of high-difficulty low-volume research domains. But of course, even within those limited pages, relatively easy-to-conduct studies have an advantage, as do their researchers, and as do research domains and research questions that allow the easier-to-conduct types of studies.

The problem that we wish to highlight begins with the simple fact that researchers who do high-difficulty, low-volume research cannot complete as many high-quality studies per year—and therefore cannot publish as many high-status journal articles per year—as those doing relatively low-difficulty, high-volume research. Thus, it is harder for scholars in certain domains to get good jobs out of graduate school, harder to get tenure, harder to get grants, and eventually harder for their graduate students to get decent jobs. This obvious fact has not-so-obvious consequences on the field, a topic to which we return near the end of this article.

Such publication pressures and evolution of science processes are always present. But when the academic environment is relatively uncrowded, when good jobs are relatively plentiful (relative to the number of good candidates), the competitive pressure is considerably lower, and high-difficulty low-volume research domains can thrive and reproduce. In the competitive environment that has existed since at least 1979 there does not appear to be a niche in which high-difficulty low-volume research domains can survive, thrive, and reproduce.

MTurkification

What does this story from the past tell us about the present? In our view, there has been another major change in psychology methodology that has again created an uneven playing field in the quest for territory in our top journals. Like the earlier shift that resulted from the cognitive revolution and the associated availability of affordable computers to run small-scale reaction-time and other related studies, the introduction of easy-to-use and inexpensive online participants has led to another shift that favors quick, easy studies at the expense of high-difficulty low-volume types of studies. As authors, as reviewers, and as a journal editor, it seems to us (a) that Psychology (especially SPP) is in the midst of a major change in how research is conducted, (b) that there are many benefits to the field to using this set of technologies and methods, and (c) that this shift has some heretofore unforeseen consequences for our field. The most commonly used resource underlying this shift is MTurk.

We conducted a simple study of articles published in our top three journals to test whether our impressions were accurate. The main empirical questions of interest in this article are as follows: (a) How dramatic has been the rise in use of online samples (MTurk and other) over the past decade or so? (b) Do online studies in general, and MTurk studies in particular, tend to be of shorter duration than other studies? (c) Has the average duration of studies become shorter, perhaps as a result of MTurk studies? and (d) Has the average number of studies per article increased over time?

Method

A list of all articles published in *JESP, JPSP*, and *PSPB* in the years 2005, 2010, and 2015 was compiled. From this list, 20 articles from each year were randomly selected for *JESP* and *PSPB* and 40 articles from each year were randomly selected for *JPSP*. Twice as many articles were selected for *JPSP* to ensure sufficient representation for each of the three subsections of the journal. This created a database of 240 articles. Block randomization was used to randomly assign one of five coders to each article. Each coder was responsible for 48 articles in total: four articles from *JESP* in each year, four articles from *PSPB* in each year, and eight articles from *JPSP* in each year. All articles were coded once more by a second coder so that interrater reliability could be calculated.

For each article, the following characteristics were coded for every individual study reported in the article: reference, journal, year, month, first page number, study number, initial sample size, actual sample size (after exclusions), source of actual sample size (e.g., reported by the authors or estimated by the coders from the degrees of freedom of reported statistical tests), source of discrepancy between initial and actual sample size, whether the study was online or offline or unclear, whether the study was most likely online or offline (if unclear), source of participants (college, MTurk, other crowdsource,[2] or other), average age, standard deviation of age, range of age, percentage of female participants, percentage of minority participants, actual duration, estimated duration (if actual duration was not reported; <15 min, 15-29 min, 30-59 min, >59 min, or longitudinal), payment, and whether supplemental materials were available. An "online best guess" variable was created by replacing all "unclear" categorizations with whichever category was deemed most likely (online or offline).

Eleven studies were excluded. Five of these were meta-analyses. Three had no participants (one used archival data, one trained a connectionist network, and one described how personality descriptors were collected by the authors for subsequent studies). Two used a very small number of research assistants (seven and 14) to rate words. For one additional study, there was insufficient information in the "Method" section to code any of the variables of interest with reasonable accuracy. After these exclusions, there were 775 studies coded from the set of 240 articles. Table 1 shows the number of studies coded in each journal for each year. A very small amount of missingness (<1% per variable) remained for variables analyzed here because of insufficient information reported in articles. Pairwise deletion was used when necessary, so sample sizes vary slightly throughout the results.

Interrater Reliability

Interrater reliability was calculated for the variables analyzed here. The Cohen's kappa values (unweighted and weighted where appropriate) for these variables are shown in Table 2.

Although most variables had acceptable reliabilities, the estimated duration variable had a fairly low reliability. Examination of coding discrepancies revealed that many of the most extreme disagreements occurred for studies classified as longitudinal. To improve reliability, all studies originally classified as longitudinal by at least one of the two coders were recoded by those two coders using a new coding scheme. In the new coding scheme, studies were classified as longitudinal (or not) and estimated duration was recoded using four categories (<15 min, 15-29 min, 30-59 min, and >59 min). This new duration variable accounted for the time spent across all time points if the study was longitudinal. As Table 2 shows, the reliability for the recoded variable was lower than the original variable—this is due in part to the fact that a smaller number of categories tends to produce lower Kappa values. Under the new coding scheme, the largest discrepancies were identified once again and these disagreements were resolved by the first author. Resolving disagreements substantially increased the reliability, boosting the new variable to moderate agreement. Although this reliability is still somewhat low, we believe it is acceptable given the subjective nature of the variable and the conservative nature of Cohen's kappa as a measure of reliability. All remaining disagreements on the estimated duration variable were resolved by taking the median value (if possible) or by randomly selecting one of the two values.[3] Disagreements for the other variables were resolved by randomly selecting one of the two values.

Table 1. Number of Studies Coded in Each Journal Across the 3 Years.

	Year			
	2005	2010	2015	Total
Journal				
PSPB	38	59	65	162
JESP	52	42	70	164
JPSP	130	146	173	449
Total	220	247	308	775

Note. PSPB = Personality and Social Psychology Bulletin; JESP = Journal of Experimental Social Psychology; JPSP = Journal of Personality and Social Psychology.

Results

The Rise of Online Studies and MTurk

The first question concerns the rise in use of online studies in general and the rise of MTurk studies specifically. Figure 1 displays the absolute and relative frequencies of online and offline studies across the three time points. Figure 2 displays the absolute and relative frequencies of the four sources of participants (MTurk, Other Crowdsource, College, and Other) across the three time points.

Not surprisingly, the absolute and relative frequency of studies conducted online has increased dramatically over time, suggesting that online studies are crowding out offline studies. The absolute and relative frequency of MTurk studies has also increased over time. Although there were no studies that used MTurk in 2005 (it was publicly launched in November of that year) and very few studies that used MTurk in 2010, by 2015 more than one third of all SPP studies in these journals were MTurk studies. The absolute frequencies suggest that MTurk studies have begun to crowd out studies using college participants.[4]

Table 2. Interrater Reliabilities for Primary Analysis Variables.

Variable	Cohen's kappa	Weighted kappa
Online (no, unclear, yes)	0.687	0.789
Online best guess (no, yes)	0.751	—
Participant source (MTurk, College, Other Crowdsource, Other)	0.861	—
Original estimated duration (<15 min, 15-29 min, 30-59 min, >59 min, longitudinal)	0.318	0.587
Recoded estimated duration (<15 min, 15-29 min, 30-59 min, >59 min)	0.281	0.441
Resolved estimated duration (<15 min, 15-29 min, 30-59 min, >59 min)	0.320	0.558

Note. Weighted Kappa was calculated using squared weights; *n*s range from 767 to 773.

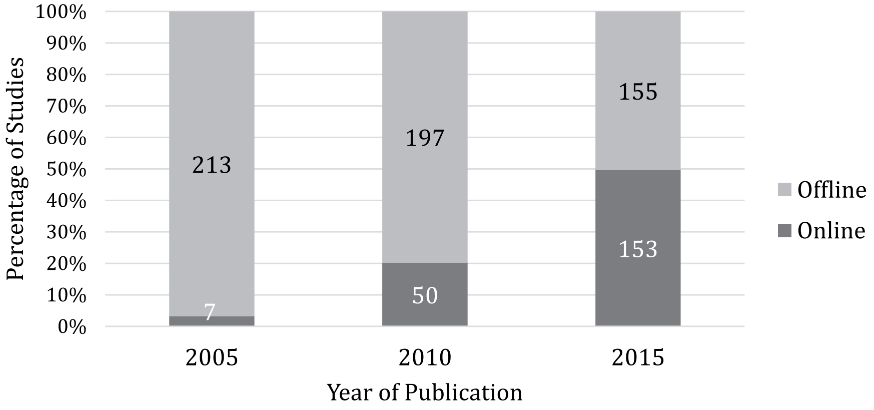

Figure 1. Absolute and relative frequencies of all reported studies that were conducted online and offline in 2005, 2010, and 2015.

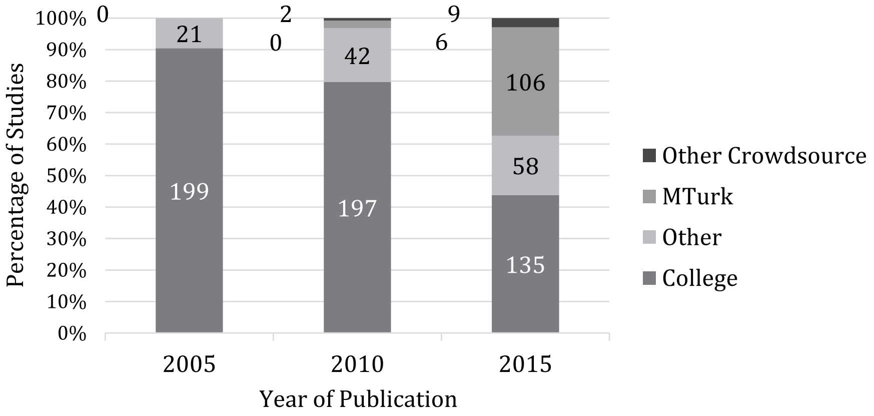

Figure 2. Absolute and relative frequencies for participant source (MTurk, other crowdsource, college, and other) in 2005, 2010, and 2015.

Duration of Studies

We coded studies into four duration categories based on stated length (when the "Method" section gave a duration) or on estimated length, judged from the procedural description (when no duration was given, which was true for the vast majority of studies). Figure 3 displays estimated duration across all years based on participant source. The "Other Crowdsource" category was collapsed into the "Other" category, because only 11 studies fell into the former category.

As is obvious, MTurk studies were the briefest by far, with 60.7% being reported/judged as taking less than 15 min, and another 30.4% taking between 15 and 29 min. The four duration categories were more evenly distributed for college

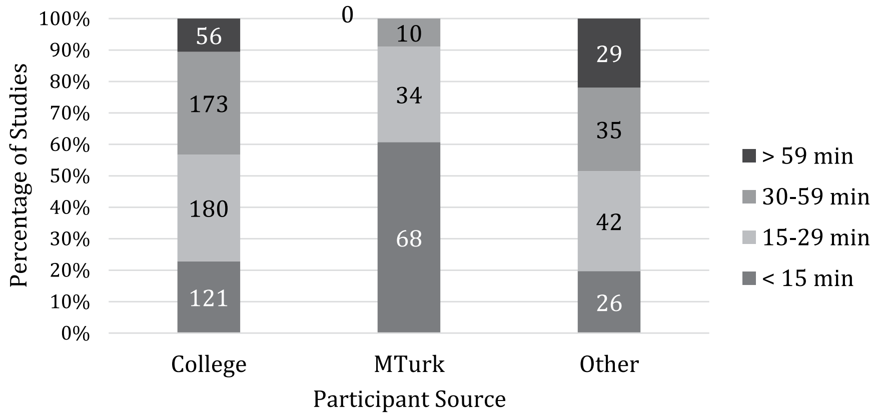

Figure 3. Absolute and relative frequencies of estimated duration as a function of participant source.

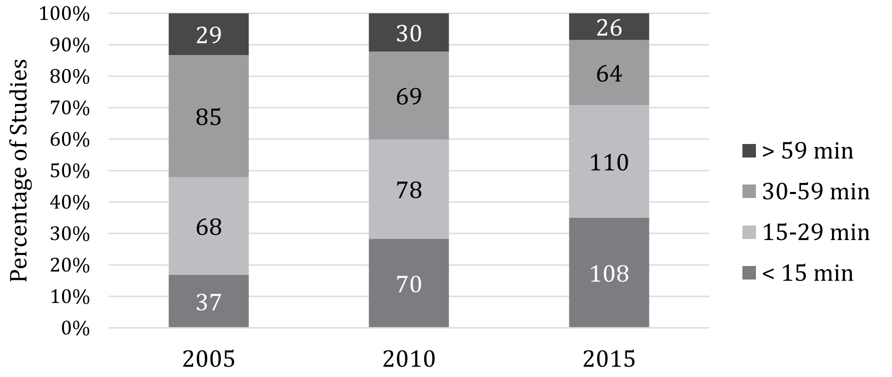

Figure 4. Absolute and relative frequencies of estimated duration of studies as a function of year.

and other samples, although relatively few college samples lasted longer than 59 min.

We also examined the distribution of durations by year, shown in Figure 4. Studies lasting <15 min and studies lasting 15 to 29 min have increased in relative and absolute frequency in recent years. Studies lasting 30 to 59 min have decreased slightly in absolute frequency and moderately in relative frequency while studies lasting >59 min have remained relatively stable in relative and absolute frequency.

However, if we exclude the MTurk studies from the data set, we get a slightly different picture, as shown in Figure 5. Here, the percentages for 2005 remain unchanged because there were no MTurk studies in that year. The percentages are also largely unchanged for 2010 (due to the small number of MTurk studies in that year). But for 2015, after excluding MTurk studies, the percentage of studies lasting less than 15 min decreases by 12.3 and the percentage of studies lasting 30 to 59 or more than 59 min increase by 6.0 and 4.4, respectively. These two figures suggest that much of the decrease in duration of studies over time is attributable to the rise of MTurk studies.

Number of Studies Per Article

Finally, we examined how the average number of studies per article has changed over time for the three journals. These averages are plotted in Figure 6.

As can be seen, the average number of studies per article has increased from 2005 to 2015. Across the three journals, the average number of studies per article was 2.75 (*SD* = 1.68) in 2005, 3.09 (*SD* = 1.73) in 2010, and 3.80 (*SD* = 2.17) in 2015.

Implications and Questions

What We Are Not Saying

We are not saying that MTurk or other crowdsource studies are bad or poor science or harmful to the field. Other articles have addressed many of the strengths and weaknesses of using MTurk in particular, so we did not address these issues here (e.g., Brawley & Pury, 2016; Cheung, Burns, Sinclair, & Sliter, 2016; Fleischer, Mead, & Huang, 2015; Goodman, Cryder, & Cheema, 2013; Harms & DeSimone, 2015; Hauser

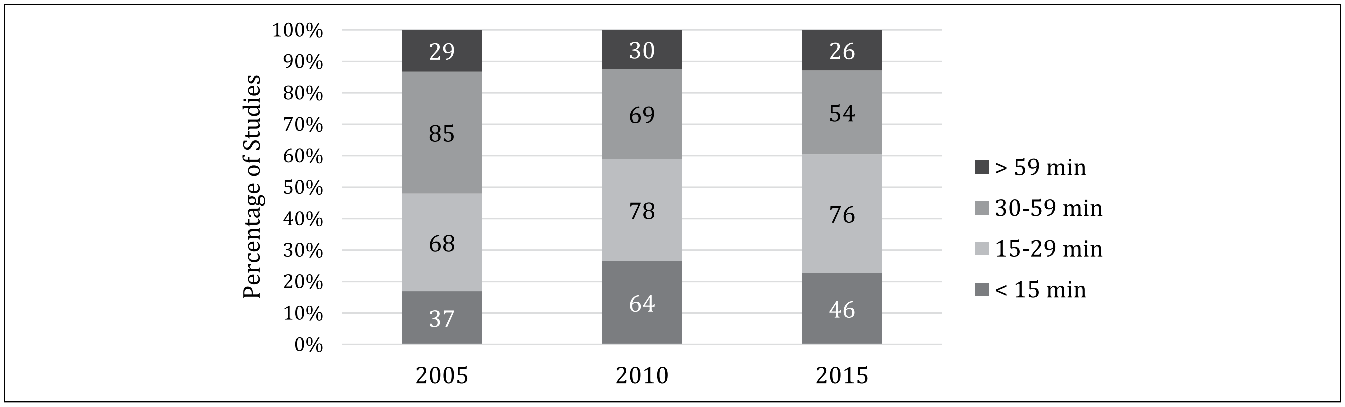

Figure 5. Estimated duration of studies as a function of year (MTurk studies removed).

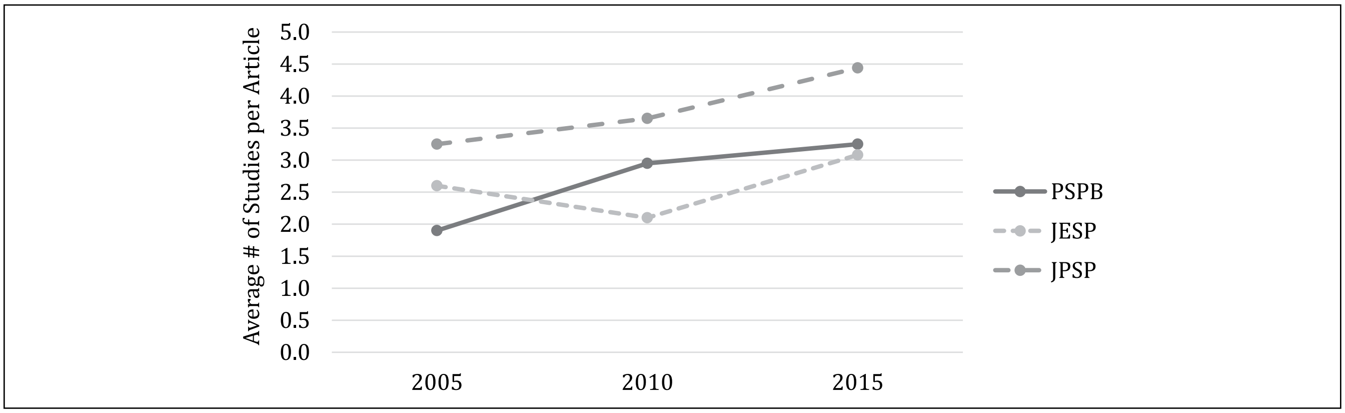

Figure 6. Average number of studies per article as a function of year and journal.
Note. PSPB = Personality and Social Psychology Bulletin; JESP = Journal of Experimental Social Psychology; JPSP = Journal of Personality and Social Psychology.

& Schwarz, 2016; Landers & Behrend, 2015; Schmidt, 2015; Siegel, Navarro, & Thomson, 2015; Smith, Sabat, Martinez, Weaver, & Xu, 2015). Greater use of MTurk is one accessible strategy for increasing samples sizes to increase statistical power and our own research team has used such online tools to good effect in recent years (e.g., Saleem, Anderson, & Barlett, 2015; Saleem, Prot, Anderson, & Lemieux, 2017). The explosive growth of social media and other forms of Internet use in the last decade has created many new domains for basic and applied social and personality psychology, domains that require new theorizing and empirical tests, some of which are undoubtedly most appropriately carried out via MTurk and similar crowdsource tools.

Second, we are not passing judgment on the value of low-difficulty, high-volume research relative to high-difficulty, low-volume research. Each type can be especially useful for specific research questions.

Third, to the extent that there is any blame to be assigned for squeezing out high-difficulty, low-volume research domains, we are not laying the blame on others. As manuscript reviewers and as a journal editor; as a member of numerous hiring, promotion, and tenure committees; as a member of grant review panels, and in numerous other roles, the senior author has almost certainly and inadvertently contributed to the problem, despite efforts to consider the ease or difficulty of the research being judged in these various decision contexts.

What We Are Saying

SPP scholars need to begin discussing two interrelated sets of questions. Should the SPP community be concerned about the subtle ways in which publication pressures and related domain-specific advantages and disadvantages influence which theories, issues, and social problems are allocated territory in our top journals? Should we care that our top journals are increasingly being filled with very brief, easily conducted studies, or that more studies are being required per article, to publish in them? Does our collective ability to improve human society and deal with modern social problems suffer by losing some high-difficulty research domains and researchers? To each question, *our* answer is a resounding "yes," and we believe that most SPP scholars will agree. What are the processes underlying denial of science in such critical domains as climate change, or media violence, or poverty? How can we model the numerous variables known to influence violent crime, racism, war? What kinds of interventions, at what levels of society, help solve these and other "wicked problems?" Brief, online studies may be able to answer a few relevant questions in these domains, but much more-in-depth, time-consuming, difficult-to-conduct studies are also necessary to develop useful theories and workable interventions. If we disproportionately allocate resources (e.g., publications, tenure) to researchers and research domains that eschew difficult-to-conduct studies, then we impair the field's ability to contribute to human welfare. As authors, reviewers, editors, hiring committee members, promotion and tenure committee members, award committee members, and publication board members, SPP scholars need to think about possible solutions at both individual and institutional levels.

What is the proper role of low-difficulty/high-volume MTurk-like studies? Questions concerning strengths and weaknesses have been addressed by numerous scholars, as noted earlier in this article. There are obvious advantages of accessing participants who are not current college students—for example, in terms of testing individual differences of a wider range of ages and life circumstances. The low time and expense aspects also benefit the field not only by allowing more data to be gathered but also by allowing quick access after unusual real-world events. However, some key research questions are not best addressed with a population that routinely does a lot of Internet-based studies; they are not generally a naïve participant population, so studies that require naïve populations (e.g., aggression, conformity, dissonance) may not replicate. Perhaps the biggest problem, though, is that many studies (e.g., interracial attitudes and prejudice, anger/aggression, social neuroscience studies) require labor-intensive face-to-face interaction with real people to get genuine reactions or to use specialized equipment. As a field, SPP scholars need to think carefully about the fit between our research questions, sample characteristics, and the methods best suited to our theory-inspired manipulations and measurements.

In our own work, we have successfully used MTurk samples to test hypotheses about exposure to certain types of media and anti-Muslim attitudes, beliefs, and action tendencies, for both correlational and experimental studies (Saleem et al., 2017). Most of our other studies of media effects, especially experimental studies, have used labor-intensive laboratory designs (e.g., Saleem & Anderson, 2013). The combination of different methods (including both simple and somewhat more complex MTurk studies, and labor-intensive face-to-face studies) helps test generalizability questions and reduces method bias.

Suggestions

Here are a few suggestions gleaned from our own work, from discussions with colleagues, and from reading other papers.

1. Whenever you are evaluating the quality of a manuscript, a job applicant, a grant proposal, a promotion dossier, or other related evaluation targets, be mindful of the difficulty (cost, time) of doing the research.
2. Be aware of the tendency of sheer numbers of studies or publications to overwhelm everything else in subjective judgment contexts.
3. When designing or evaluating research, keep in mind that a prediction that is tested (and supported) in multiple ways (i.e., conceptually replicated) is more convincing than one that is tested repeatedly using only the same methods and population (i.e., directly or closely replicated) (Crandall & Sherman, 2016). As Meehl (1990) said, "Any working scientist is more impressed with 2 replications in each of 6 highly dissimilar experimental contexts than he [*sic*] is with 12 replications of the same experiment" (p. 111). Thus, if a phenomenon of interest works well in brief online studies, showing that it also works well in more difficult-to-run lab or field contexts helps establish its validity across contexts. Finding that it does not work the same way in other contexts or with other populations can also be very informative. Requiring "constraints on generality" statements in publications could help with this by encouraging authors to thoughtfully consider the context of their findings (based on participants, materials/stimuli, procedures, and historical/temporal specificity) and the extent to which results may generalize (Simons, Shoda, & Lindsay, 2017).
4. Similarly, consider the trade-offs associated with emphasizing some scientific goals (e.g., replicability) over others (e.g., external validity)—desirable characteristics for high-quality science are often at odds with one another and no single study can include all desirable characteristics (Finkel, Eastwick, & Reis, 2017). Focusing on certain goals to the exclusion of others as a field is likely to create new problems while solving others. Acknowledging such trade-offs can help us maintain balance.
5. The coding problems that we encountered took us by surprise. As a field, SPP scholars (authors, reviewers and editors) need to do a better job of reporting important methodological details. The availability of almost limitless "space" in supplemental materials should make this one easy.

Coda

The growing dominance in top SPP journals of brief easy-to-conduct studies using Amazon's Mechanical Turk and similar online tools is both enlightening and alarming. We risk extinction of the kinds of difficult-to-conduct studies needed in some of the most important real-world problems that SPP scholars have traditionally tackled; we also risk losing the researchers who rely on them. The field needs to be mindful of the risks and opportunities if we are to become even more useful to the world around us. In this sense, we strongly endorse the *evidentiary value movement* that has been discussed to date largely in the context of the so-called replication crisis (e.g., Finkel, Eastwick, & Reis, 2015). In the present context, this means paying closer attention to the overall value added by a study to our understanding of important theoretical and real-world problems, and less attention to number of studies in a paper, proposal, or career.

We hope that this article sparks thoughtful conversations among SPP scholars, young and old. At a minimum, we hope that the editorial teams at our top journals discuss these issues in the context of their roles as gate-keepers to the territory that determines the evolution of our field.

Authors' Note

Statements in which "we" speculate, ruminate, or refer to memories of the old days (good or bad) are attributable to the senior author, so do not blame the junior authors. We use the royal "we" merely to simplify the reading of this article. Statements based on hard data may be credited to all authors.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. We thank an anonymous reviewer for these suggestions.
2. Other crowdsource was defined as any other "MTurk-like" sample. Examples included the following: users of MyPersonality (a Facebook application), affiliate of Qualtrics, Crowdflower, and academic website.
3. All of the coding difficulties, especially involving duration, suggest that as a field we have not been doing a very good job reporting basic features of our research methods.
4. Inferential statistics seem unnecessary for the results presented in this article. Thus, they are relegated to supplemental materials.

Supplemental Material

Supplementary material is available online with this article.

ORCID iD

Craig A. Anderson https://orcid.org/0000-0001-6353-0023
Johnie J. Allen https://orcid.org/0000-0002-5102-3048

References

Brawley, A. M., & Pury, C. L. S. (2016). Work experiences on MTurk: Job satisfaction, turnover, and information sharing. *Computers in Human Behavior*, *54*, 531-546. doi:10.1016/j.chb.2015.08.031

Cheung, J. H., Burns, D. K., Sinclair, R. R., & Sliter, M. (2017). Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology*, *32*, 347-361. doi:10.1007/s10869-016-9458-5

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93-99. doi:10.1016/j.jesp.2015.10.002

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*, 275-297. doi:10.1037/pspi0000007

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2017). Replicability and other features of a high-quality science: Toward a balanced and empirical approach. *Journal of Personality and Social Psychology*, *113*, 244-253. doi:10.1037/pspi0000075

Fleischer, A., Mead, A. D., & Huang, J. (2015). Inattentive responding in MTurk and other online samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*, 196-202. doi:10.1017/iop.2015.25

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

Harms, P. D., & DeSimone, J. A. (2015). Caution! MTurk workers ahead—Fines doubled. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*, 183-190. doi:10.1017/iop.2015.23

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*, 400-407. doi:10.3758/s13428-015-0578-z

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*, 142-164.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*, 108-141.

Reis, H. T., & Stiller, J. (1992). Publication trends in JPSP: A three-decade review. *Personality and Social Psychology Bulletin*, *18*, 465-472. doi:10.1177/0146167292184011

Saleem, M., & Anderson, C. A. (2013). Arabs as terrorists: Effects of stereotypes within violent contexts on attitudes, perceptions and affect. *Psychology of Violence*, *3*, 84-99. doi:10.1037/a0030038

Saleem, M., Anderson, C. A., & Barlett, C. P. (2015). Assessing helping and hurting behaviors through the tangram help/hurt task. *Personality and Social Psychology Bulletin*, *41*, 1345-1362.

Saleem, M., Prot, S., Anderson, C. A., & Lemieux, A. F. (2017). Exposure to Muslims in media and support for public policies harming Muslims. *Communication Research*, *44*, 841-869.

Schmidt, G. B. (2015). Fifty days an MTurk worker: The social and motivational context for Amazon Mechanical Turk workers. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*, 165-171. doi:10.1017/iop.2015.20

Siegel, J. T., Navarro, M. A., & Thomson, A. L. (2015). The impact of overtly listing eligibility requirements on MTurk: An investigation involving organ donation, recruitment scripts, and feelings of elevation. *Social Science & Medicine*, *142*, 256-260. doi:10.1016/j.socscimed.2015.08.020

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123-1128. doi:10.1177/1745691617708630

Smith, N. A., Sabat, I. E., Martinez, L. R., Weaver, K., & Xu, S. (2015). A convenient solution: Using MTurk to sample from hard-to-reach populations. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *8*, 220-228. doi:10.1017/iop.2015.29