

# Accessibility and generalizability: Are social media effects moderated by age or digital literacy?

Research and Politics  
April-June 2021: 1–16  
© The Author(s) 2021  
Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
DOI: 10.1177/20531680211016968  
[journals.sagepub.com/home/rap](https://journals.sagepub.com/home/rap)  


Kevin Munger<sup>1</sup>, Ishita Gopal<sup>1</sup>, Jonathan Nagler<sup>2</sup>  
and Joshua A. Tucker<sup>2</sup>

## Abstract

An emerging empirical regularity suggests that older people use and respond to social media very differently than younger people. Older people are the fastest-growing population of Internet and social media users in the US, and this heterogeneity will soon become central to online politics. However, many important experiments in this field have been conducted on online samples that do not contain enough older people to be useful to generalize to the current population of Internet users; this issue is more pronounced for studies that are even a few years old. In this paper, we report the results of replicating two experiments involving social media (specifically, Facebook) conducted on one such sample lacking older users (Amazon's Mechanical Turk) using a source of online subjects which does contain sufficient variation in subject age. We add a standard battery of questions designed to explicitly measure digital literacy. We find evidence of significant treatment effect heterogeneity in subject age and digital literacy in the replication of one of the two experiments. This result is an example of limitations to generalizability of research conducted on samples where selection is related to treatment effect heterogeneity; specifically, this result indicates that Mechanical Turk should not be used to recruit subjects when researchers suspect treatment effect heterogeneity in age or digital literacy, as we argue should be the case for research on digital media effects.

## Keywords

Digital literacy, generalizability, online media effects

## Accessibility and generalizability

Social scientists often aim to generate generalizable knowledge—to take an internally valid finding and extrapolate it to a “target population” of interest. In the case of pre-election polling, for example, the target population is “likely voters”; the opinions of non-voters are not immediately useful for predicting how voters are likely to cast their ballot. The formula for determining “likely voting” is essential. Polling firms frequently try to optimize these formulas, and naive estimates of the general population (pooling non-voters and voters) produce severely biased election predictions.

The “target population” for social science research about social media is much more difficult to define. First, it is unclear that each social media user should be given equal weight in the population; there is massive variation in the intensity of social media use, from a few minutes a month to multiple hours per day (Smith and Anderson, 2018). This population can also vary across different types of social

media use; the “target population” of social media posters is not the same as the “target population” of social media consumers.

The second problem is the rapidity and non-transparency of changes in the target population. There are hundreds of pre-election polls conducted before each Presidential Election, and frequently updated models of who is a “likely voter.” In contrast, it takes years from the conception of a research idea to the publication of a given experimental study. What, then, should the “target population” for that study be? The (weighted) population of social media users at the beginning of the study, at the publication

<sup>1</sup>The Pennsylvania State University, USA

<sup>2</sup>New York University, USA

### Corresponding author:

Kevin Munger, Penn State, Pond Lab, University Park, PA 16802-1503, USA.

Email: [kmm7999@psu.edu](mailto:kmm7999@psu.edu)



of that study, or in the future? In practice, collecting this population information is difficult and expensive.

All studies are read after they are conducted. Given the pace of academic knowledge production, this lag time can take years. This suggests a third problem: future target populations contain types of people that past populations do not, and no amount of weighting can address this lack of mutual support. Philosophically, this is always true, but it only poses a threat to generalizable knowledge when the effect under study is heterogeneous in a “type” that lacks support in the past.

There is a growing empirical consensus for just such a heterogeneity in the study of online media effects: older people use and respond to the Internet differently than younger people.<sup>1</sup> One crucial question is the extent to which these heterogeneities are due to age per se, or rather some other quantity that covaries with age.<sup>2</sup> The primary contender for such a quantity is “digital literacy,” a concept best associated with the work of sociologist Eszter Hargittai (Hargittai, 2002, 2005). A recent working paper by Guess and Munger (2020) makes the case for the use of digital literacy in political science, defining it as “information discernment combined with the basic digital skills necessary to attain it.” Although there is substantial heterogeneity in digital literacy at all ages, a nationally representative sample in Guess and Munger (2020) finds a strong negative correlation between age and digital literacy.

With this in mind, Mechanical Turk (MTurk)—a large and growing source of subjects for research in digital media effects—may be precisely the wrong population for answering questions about age-based heterogeneities because of a lack of variation in both age and digital literacy.<sup>3</sup> Only a small number of MTurkers are over 65 years old (Huff and Tingley, 2015), but the situation for digital literacy is worse: 100% of subjects recruited via MTurk are above a certain threshold of digital literacy simply by dint of being on the platform.

Brewer et al. (2016) explain why this is the case. They conduct a survey of older Americans and report serious age-related heterogeneities even within this subgroup: “even among online older adults, those who have tried crowd work are (relatively) younger and more tech savvy than those who have not” (8). In addition, Brewer et al. (2016) perform a qualitative analysis on a small sample of older adults who have not used MTurk, signing them up for the platform and interviewing them about the experience. Although these subjects generally reported being comfortable using computers, they could not complete basic tasks on MTurk:

Many participants were not familiar or comfortable with opening content in new tabs/windows, resulting in questions such as, “How do I get back to the instructions?” . . . after a new tab was opened. Also, participants often forgot the instructions immediately upon opening the new window. (2251)

Thus, even if low digital literacy subjects use MTurk, their poor performance may cause them to receive low ratings and then be excluded from future trials.

In contrast, even though the age skew of the MTurk population poses a problem, it is possible to selectively recruit older MTurkers or reweight the data. However, as Mullinix et al. (2015) argue, reweighting is insufficient when working with joint distributions: “[MTurk] may have similar percentages of older individuals and racial minorities, but may not match the population based sample with respect to older minorities” (123). Analogously, the older people on MTurk are likely to be higher in digital literacy than the older people not on MTurk, and weighting on either dimension in isolation does not permit generalizability to this crucial crosstab.

### Age-related heterogeneity of digital media effects

In the early days of social media, users of these platforms were dramatically more likely to be young, educated, and tech-savvy. Our understanding of the effects of social media usage began with this population, and indeed the overwhelming majority of the research in this area was conducted on subjects drawn from this population.

This fact has obscured the importance of age-related heterogeneities in the experience of using social media, and especially Facebook. In 2018, however, the fastest-growing population of Facebook users was adults over 65 years old (Smith and Anderson, 2018), and these heterogeneities are increasingly showing up in empirical research:

- Using survey data linked with Facebook data, Guess et al. (2019) report that “users over 65 shared nearly 7 times as many articles from fake news domains as the youngest age group” during the 2016 US Presidential election.
- Barberá (2018) finds that people over 65 shared roughly 4.5 as many fake news stories on Twitter as people 18 to 24.
- Grinberg et al. (2019) find that people over 65 were exposed to between 1.5 and 3 times as much fake news on Twitter as the youngest people; the slope depends heavily on partisanship.
- The most famous digital voting experiment, conducted in concert with Facebook in 2010, finds similar results (Bond et al., 2012, 2017): “The [Facebook GOTV experiment] effect size for those 50 years of age and older versus that of those ages 18 to 24 is nearly 4 times as large for self-reported voting and nearly 8 times as large for information seeking.” In absolute terms, the treatment caused a 1.5 percentage point increase in self-reported voting among the youngest group and a 6 percentage point increase in the oldest group.

By most standards of effect heterogeneity, these differences are massive. The size and consistency of this empirical regularity suggests that heterogeneous effects are a first-order concern for research on the topic of social media effects—average treatment effects may be misleading. This evidence all refers to Facebook. Although Facebook is not representative of all social media in terms of who uses it (Hargittai, 2015) or how they use it (Hargittai et al., 2018), our theory of digital literacy as the key conceptual moderator driving these effects on Facebook predicts that analogous (but not identical) heterogeneities will exist on other social media platforms for which there is sufficient variation in the age or digital literacy of users.

However, all of this evidence is still suggestive; the replications undertaken in this study demonstrate the existence and extent of the issue for survey experiments in which the treatment effects are moderated by age or digital literacy. Digital literacy is far from the only relevant moderator to covary from age; working within the tradition of online credibility assessment, Liao and Fu (2014) find that older adults are generally more trusting of online content and less swayed by the kind of credibility cues that operate in the proposed replications. We cannot fully test each of these alternative pathways in the current paper, so we have focused on digital literacy.

We note that this is the procedure suggested in a recent large-scale demonstration of a tendency towards treatment effect homogeneity. Coppock (2018) presents compelling evidence that a wide variety of survey experimental findings generated from nationally representative samples can be replicated using MTurk samples. None of these studies, however, investigate a treatment effect which is theorized to be heterogeneous in age or digital literacy. As Coppock argues:

Future disagreements about whether a convenience sample can serve as a useful database from which to draw general inferences should be adjudicated on the basis of rival theories concerning treatment effect heterogeneity (or large, well-powered empirical demonstrations of such theories). (2018: 624)

This position is an echo of Druckman and Kam's (2011) defense of the use of student samples: that the burden of proof should be on critics of student samples to demonstrate why they are inappropriate for a given study.

We argue that the evidence (both qualitative and experimental) summarized above makes a plausible case for digital literacy as a significant moderator of online treatment effects, and thus for the inadequacy of convenience samples like MTurk that do not ensure sufficient variation on this dimension.

Literacy here is a useful metaphor. We have good reason to expect that a written framing experiment will have different effects on people who can and cannot read.

The scope of "literacy" is easily defined: the ability to understand written text. "Digital literacy" is a more diffuse concept, but the experiments we choose to replicate relate

to a central process in online media: the ability to navigate social media feeds and come away with true or useful inferences. Digital literacy may moderate a host of other processes relate to information on the Internet, but we aim to establish that this is the case in the narrow but important examples below.

## Replicating results from MTurk

In an influential paper, Messing and Westwood (2014) demonstrate that social endorsements in the form of Facebook "likes" are more powerful than partisan source cues in determining what news stories subjects opt to consume. However, the study was far from representative of the US population in terms of age:

The 18 to 34 demographic was slightly overrepresented: with 25% of respondents in the 18 to 24 age group, 35% in the 25 to 34 group, 31% in the 35 to 54 group, and 9% in the 55 and older group.

In 2010, 33% of American adults were 55 years and older. As a result, we believe the sample employed in this study does not allow the authors to make claims about how low digital-literacy individuals respond to social endorsements.

We hypothesize that the older and less digitally literate adults excluded from the original study (which was conducted on MTurk) but included in the Facebook sample will not respond as strongly to the Facebook social cues.<sup>4</sup> Depending on the magnitude of this heterogeneity, the main effects may no longer be significant on this population.

More recently, Anspach and Carlson (2018) examine the role of social commentary on source credibility and information acquisition. They begin with a standard news article "preview post" (how Facebook displays a preview of an article shared by users), and manipulate it by adding slanted commentary by the user sharing the post. They find that this manipulation causes subjects to recall the misinformation shared by the user's comment rather than the correct information shared in the news preview itself.

In this case, we hypothesize that the use of the MTurk sample *underestimates* the extent of effect heterogeneity (for which there was no evidence in the original study). The older, less digitally literate people not found on MTurk should be more likely to be taken in by misinformation.

In addition to these direct replications, we have included the shortened 10-item survey battery developed by Hargittai and Hsieh (2012) to measure digital literacy. This battery (validated with behavioral measures) is the standard survey operationalization of the concept. Measuring digital literacy allows us to understand the extent to which that—or age—is more responsible for the hypothesized effect heterogeneities. Although the evidence in the literature is inconclusive, our hypothesis here is that age will have a large and significant effect even after controlling for digital literacy, which will also have a large and significant effect.

Finally, to narrow the focus of the theoretical concern with specific online samples, we replicated an established effect that is not theorized to be moderated by age or digital literacy. Coppock (2018) replicates a whole host of survey experiments on MTurk; we selected the oldest finding so replicated, Haider-Markel and Joslyn's (2001) framing experiment on support for gun control. We hypothesized that the results of our replication will mirror the original finding and Coppock's (2018) recent replication, and that there will be no treatment effect heterogeneity in age or digital literacy.

## Questions and stimuli

We now describe the experimental stimuli and dependent variable measurements used in the three original experiments, as well as our measurement strategy for digital literacy.

### *Anspach and Carlson*

Our replication consisted of one randomized image stimulus, followed by five questions. The stimulus was one of three: (A), the conservative commentary condition; (B) the liberal commentary condition; and (C) the no commentary condition.<sup>5</sup> The measurement questions were as follows:

- On the last page, there was an article about President Trump's approval ratings at the 6-month mark of his presidency. We'd like to ask you a few questions about that article.
- According to the article, what percentage of Americans approve of President Trump's performance? [15%, 23%, 36%, 49%, I don't know]
- According to the article, what is the biggest criticism of the recent Washington Post-ABC News poll? [It surveyed more Democrats than Republicans, It surveyed more Republicans than Democrats, It was inaccurate in its prediction of the 2016 Presidential election, The poll was funded by liberals, I don't know]
- As best as you can tell, how trustworthy is the person who wrote/posted the article?
- As best as you can tell, how trustworthy is the news outlet that published the article?
- As best as you can tell, how trustworthy is the poll cited within the article? [Very trustworthy, Somewhat trustworthy, Not sure, Somewhat untrustworthy, Very untrustworthy]

Figure 1 displays our operationalization of the stimuli; these are almost identical to the stimuli used in the original.

### *Messing and Westwood*

Our replication consisted of four choice sets, each of which contained a visual stimulus and four possible headlines. Respondents were randomly assigned to one of three

conditions—source cues only (A), social cues only (B), and both cues (C). They were subsequently instructed to choose one of the four headlines. This process was repeated four times.

A major challenge (and open question, for replicators of these kind of experiments) is how to change this setup to more closely mirror the contemporary Facebook feed. Anspach and Carlson's (2018) images are much more recent, and very different from the ones used in Messing and Westwood (2014). What is the correct way to "replicate" the latter study? To use an artificial/outdated framework, or to update the framework? If the latter "replication" were to produce different results, how can we be confident that this was truly due to the different sample and not some artifact of the change?

Our plan was to do the latter, and thus to make both of the proposed replications look similar (they do, of course, involve different modifications and tasks). The original stimuli can be seen in the 2014 paper, and Figure 2 updates the framework to match Facebook's current design scheme.<sup>6</sup>

### *Haider-Markel and Joslyn*

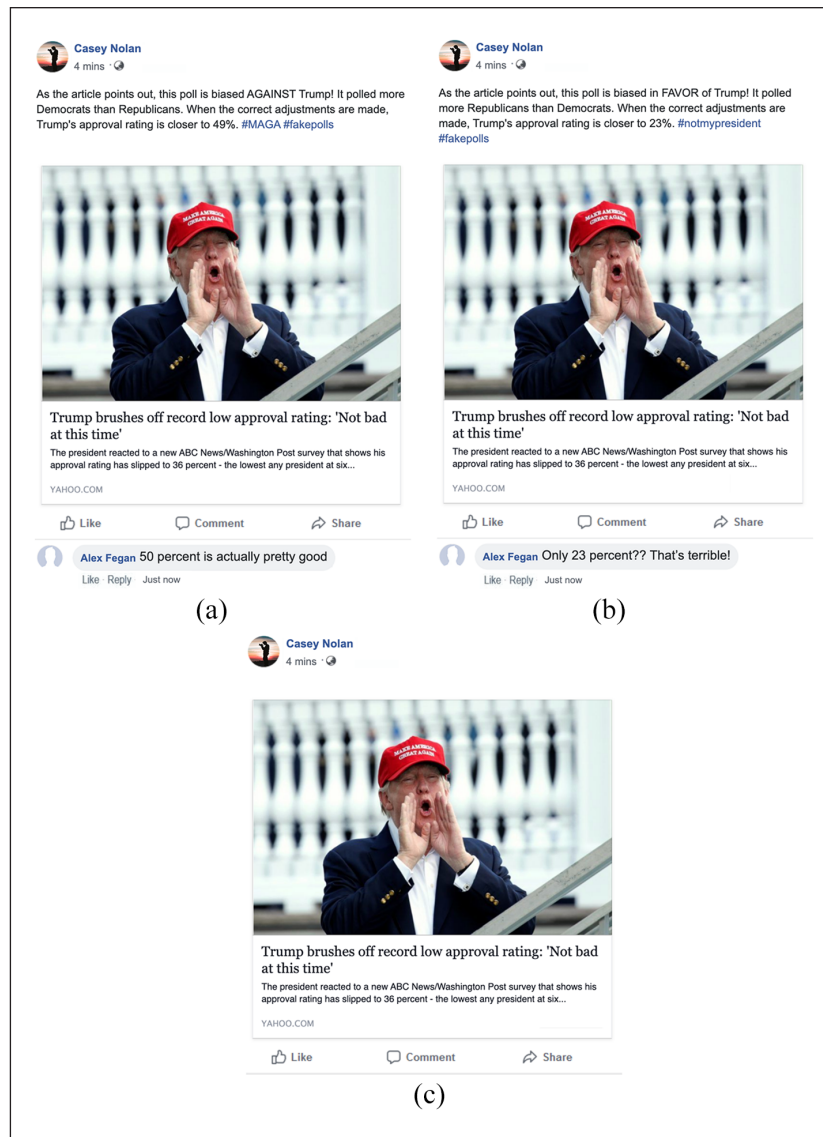
Our replication consisted of a single framing experiment, randomizing whether subjects receive one of two frames on the topic of regulating concealed handguns. After receiving the stimulus, subjects answered a 7-point question about their support for concealed handgun laws:

- Concealed handgun laws have recently received national attention. Some people have argued that law-abiding citizens have the **right** to protect themselves. What do you think about concealed handgun laws?
- Concealed handgun laws have recently received national attention. Some people have argued that laws allowing citizens to carry concealed handguns **threaten public safety** because they would allow anyone to carry a gun almost anywhere, even onto *school grounds*. What do you think about concealed handgun laws?

### *Digital literacy*

Hargittai (2005) developed an extensive battery of survey questions to measure digital literacy, later updated and shortened in Hargittai and Hsieh (2012). We used the 10-item version from their latest paper in our analysis. In the questionnaire, respondents were presented with a list of Internet-related terms and were subsequently asked to rate their familiarity with each of them. Specifically, they were asked: "How familiar are you with the following computer and Internet-related items? Please choose a number between 1 and 5 where 1 represents 'no understanding' and 5 represents 'full understanding' of the item."





**Figure 1.** Experimental images for the replication of Anspach and Carlson's (2018) study.

The terms used in our experiment were specifically designed to measure web-use skills of people who have low levels of Internet experience, implying that our threshold for being digitally literate was quite low. The list we presented respondents with included the following terms: advanced search; favorites; preference setting; PDF; spyware; Wiki; JPG; blog; malware; and phishing. Half the items on this list consist of highly understood terms and overall the distribution covers terms that encompass general use and those related to security and privacy. The idea was to present items that would be relevant to a wide variety of users.

## Implementation

Following the results in Guess and Munger (2020), we conducted these replications on a sample recruited from Lucid.

Of all the sources of experimental subjects studied in the paper (other than a small sample of individuals recruited from an introductory computer skills course), Lucid offers the highest number of both older and low digital-literacy respondents and thus the greatest diversity in terms of both variables. Lucid is a recently developed platform for recruiting nationally representative samples, with specific functionality designed for social science researchers (Coppock and McClellan, 2019).

The overall cost per participant is exactly US\$1 for Lucid's standard 10-minute Qualtrics survey instrument. Attrition rates from this sample tend to be moderate. The usefulness of attention checks on this population is not yet established; there is some concern that people who are lower in digital literacy may misunderstand the attention check, and removing these subjects would cause



**Figure 2.** Experimental images for the replication of Messing and Westwood's (2014) study.

more of the problem (sample selection bias) that our design aims to solve (Berinsky et al., 2019). Therefore, our approach was to include an attention check but to present the results for both the full population (everyone who completed the survey) and only those people who pass the attention check.

As discussed below, we needed a large sample sizes to detect the treatment effect heterogeneity at the heart of our argument. The details of the power calculation are discussed in Appendix D. In order to be able to reliably detect these heterogenous treatment effects, we calculated that we needed to recruit 9000 subjects (Blair et al., 2019).

After asking standard demographic, political knowledge, and media consumption questions (specifically, all of the questions used to produce control variables in the replicated experiments), we added the digital literacy scale discussed above.

Subjects then completed each of the three experiments in randomized order. There is no reason to expect spillovers

across the conditions, but we were well powered to detect them if they existed.

Our main result, following Coppock (2018), plots the Average Treatment Effects (ATEs) for the three experiments in our Lucid sample against the previous ATEs recorded from MTurk samples. We also explicitly test for treatment effect heterogeneity in both age and digital literacy.

## Results

Our registered report called for a sample of 9000 recruited from Lucid. However, in the weeks before we planned to implement the study, a working paper titled "Evidence of rising rates of inattentiveness on Lucid in 2020" was circulated online (Aronow et al., 2020). The paper presented compelling evidence that an increasing number of respondents were failing extremely basic attention checks and recommended including several of these attention checks at the very beginning of any surveys using Lucid.

We decided to follow this advice. The survey architecture of Lucid allows the requesting researcher to reject subjects from the beginning of the survey who do not meet the requirements; this is the option we took, to avoid paying fraudulent responders. However, the process by which Lucid actually finds subjects is opaque; although we paid for 9000 respondents, we ended up with a total of 24,749 respondents accessing the survey.

As attention checks, we asked respondents the following two questions at the beginning of the survey:

- For our research, careful attention to survey questions is critical! To show that you are paying attention please select “I have a question.”
- We want to get your opinion on a number of political topics of interest to the US population today. There are not any right or wrong answers to these questions. But to ensure that you are reading each question carefully, we need you to select the sixth option out of the following eight options.

After eliminating those who had failed these initial attention checks and those who did continue the survey far enough to provide answers to essential questions, we ended up with 11,783 subjects, although we paid for only 9000.

We also included a mid-survey attention check, which was passed by the 10,579 of the 11,783 subjects who passed the first round. Our results are based on this sample.

We also include Appendix C the results on the full sample of 11,783. Results are not substantively different. This number of subjects is greater than the 9000 we had registered. We have opted to include the subjects we got “for free”; Appendix B shows all results restricted to the first 9000 subjects who passed the first attention checks (regardless of their performance on the mid-survey attention checks), and the results are again not substantively different.

### *Anspach and Carlson experiment*

In the Anspach and Carlson experiment, we first calculated the average responses to the outcome measurement questions across the three different treatment conditions. This included the outcome measures that we do not believe to be moderated by digital literacy. Table 1 replicates Table 3 from the original paper, plotting the average treatment effects of the two treatment conditions on several measures of trust.<sup>7</sup> As expected, our results replicate the main effect that each of the political commentary conditions reduce trust in the news outlet, author/poster, and the cited poll.

We include interaction terms for digital literacy and age not present in the original paper. We did not expect to find treatment effect heterogeneity in this test. This expectation was borne out in the case of age, although we do find a significant main effect of age, unlike the original experiment.

We do, however, find evidence of treatment effect heterogeneity by subject digital literacy. The interaction term of  $-0.071$  on the “Preview + con. commentary \* Digital Literacy” coefficient in the second column of Table 2 indicates that among subjects exposed to the conservative commentary condition, each unit increase on the normalized 5-point Digital Literacy scale caused a decrease in the reported Trustworthiness of News Outlets of  $7/100$ s of a point on the 4-point Trustworthiness scale. In fact, the addition of this interaction term renders the main effect insignificant; there are no treatment effects of the commentary conditions among subjects with the lowest level of digital literacy. The reduction in trustworthiness caused by these conditions only occurs among more digitally literate subjects. The original study, using subjects recruited from MTurk, would not have detected this heterogeneity, as it would not have had subjects on the low end.

Next, the first panel in Table 2 replicates Table 4 from the original study, serving as a “manipulation check” that the subjects observed the different treatment conditions. The original analysis is a simple table, so we replicated that analysis and further added interaction terms generated by dividing the samples into quartiles based on digital literacy and age. Here, we expect to find heterogeneity by age and/or digital literacy. This expectation was borne out; there is a significant difference in the percentage of subjects in the highest and lowest quartile of age reporting the manipulated quantity in one of the two treatment conditions; for digital literacy, it is both of the treatment conditions.

The main test is presented in Figure 3, replicating Figure 2 in the original. The original figure reports the percentage of subjects in each treatment condition who selected either the correct number from the poll reported in the stimulus or one of the incorrect numbers reported in the “comment misinformation” treatment conditions. The middle panel of Figure 3 shows an identical analysis with essentially identical results.

The bottom panels report the treatment effects on this “correctness” dependent variable at different levels of digital literacy and age. The Appendix Table 4 provides the details of these regressions. We find significantly positive main effects for both age and digital literacy on correctly answering.

As expected, we find that digital literacy significantly moderates the treatment effect of each treatment condition and that age does so for one of two (the conservative commentary condition). Figure 3 demonstrates the magnitude of these heterogeneities; the reduction in correct answers caused by the conservative condition is 19 percentage points for subjects at the low end of the digital literacy scale but 34 percentage points for subjects at the high end.

However, these treatment effect heterogeneities are in the *opposite direction* from what we expected. The left panel demonstrates that the disinformation effect of the conservative commentary condition was larger on the youngest subjects and smaller on the oldest subjects. In the right panel, only the informational condition saw increased rates of

**Table 1.** Treatment effects on perceived trustworthiness of News Outlet, Author of post and Cited poll (Replicated Table 3 from Anspach and Carlson (2018)). (a) With age in the interaction.

	News Outlet	Author/poster	Cited poll	
Preview + con. commentary	-0.219*** (0.026)	-0.222*** (0.073)	-0.226*** (0.071)	-0.306*** (0.067)
Preview + lib. commentary	-0.261*** (0.026)	-0.286*** (0.072)	-0.293*** (0.070)	-0.277*** (0.065)
Age	-0.008*** (0.001)	-0.005*** (0.001)	-0.003*** (0.001)	-0.004*** (0.001)
Male			0.233*** (0.020)	0.164*** (0.019)
White			-0.026 (0.024)	0.006 (0.023)
Party			-0.264*** (0.012)	-0.159*** (0.011)
Preview + con. commentary		-0.001	-0.001	0.001
* Age		(0.002)	(0.001)	(0.001)
Preview + lib. commentary		-0.002	-0.002	0.00004
* Age		(0.002)	(0.001)	(0.001)
Constant	3.123*** (0.019)	3.284*** (0.052)	3.088*** (0.053)	3.113*** (0.049)
Observations	10,759	10,759	10,759	10,759
R <sup>2</sup>	0.011	0.029	0.079	0.045
Adjusted R <sup>2</sup>	0.010	0.028	0.078	0.044
Residual Std. Error	1.104 (df = 10756)	1.079 (df = 10753)	1.051 (df = 10750)	0.993 (df = 10750)
F Statistic	57.389*** (df = 2; 10756)	63.841*** (df = 5; 10756)	115.095*** (df = 8; 10750)	48.779*** (df = 5; 10753)
			83.008*** (df = 2; 10756)	63.466*** (df = 8; 10750)

Note: \*p0.1; \*\*p0.05; \*\*\*p0.01.



(b) With digital literacy in the interaction.

	New Outlet	Author/Poster	Cited poll	
Preview + con. commentary	-0.219*** (0.026)	-0.266*** (0.026)	-0.113 (0.103)	-0.120 (0.100)
Preview + lib. commentary	-0.261*** (0.026)	-0.370*** (0.026)	-0.059 (0.104)	-0.058 (0.101)
Digital Literacy	0.135*** (0.019)	0.106*** (0.019)	0.106*** (0.019)	0.073*** (0.019)
Male		0.170*** (0.021)	0.232*** (0.021)	0.161*** (0.019)
White		-0.047*** (0.024)	-0.065*** (0.024)	-0.042* (0.022)
Party		-0.315*** (0.012)	-0.266*** (0.012)	-0.160*** (0.011)
Preview + con. commentary				
*Digital Literacy		-0.071*** (0.027)	-0.041 (0.027)	-0.038 (0.026)
Preview + lib. commentary		-0.075*** (0.027)	-0.083*** (0.027)	-0.082*** (0.026)
*Digital Literacy				
Constant	3.123*** (0.019)	3.060*** (0.018)	2.664*** (0.074)	2.699*** (0.074)
Observations	10,759	10,759	10,759	10,759
R	0.011	0.020	0.024	0.077
Adjusted R	0.010	0.020	0.024	0.076
Residual Std. Error	1.104 (df = 10756)	1.083 (df = 10756)	1.081 (df = 10753)	1.052 (df = 10750)
F Statistic	57.389*** (df = 2; 10756)	111.049*** (df = 2; 10750)	53.418*** (df = 5; 10753)	83.008*** (df = 8; 10750)
			41.432*** (df = 5; 10753)	60.530*** (df = 8; 10750)

Note: \* p0.1; \*\* p0.05; \*\*\* p0.01.

**Table 2.** Manipulation check.

(a) Replicated Table 4 from Anspach and Carlson (2018).

Perceived Flaw	Article Preview(%)	"Oversampled Repu" comment (%)	"Oversampled Dem" comment (%)
I don't know	59	12	10
It surveyed more Democrats than Republicans	16	15	77
It surveyed more Republicans than Democrats	9	66	7
It was inaccurate in its prediction of the 2016 Presidential election	11	5	4
The poll was funded by liberals	5	2	3

(b) Manipulation check by age.

Perceived Flaws	Article Preview(%)				"Oversampled Rep" comment (%)				"Oversampled Dem" comment (%)			
	18-31	31-45	45-61	61-99	18-31	31-45	45-61	61-99	18-31	31-45	45-61	61-99
I don't know	13	13	17	16	4	3	3	3	3	3	3	2
It surveyed more Democrats than Republicans	5	6	3	2	4	4	3	3	17	19	20	21
It surveyed more Republicans than Democrats	3	4	2	0	17	16	17	16	2	3	1	1
It was inaccurate in its prediction of the 2016 Presidential election	2	3	3	3	1	1	1	1	1	1	1	1
The poll was funded by liberals	1	1	1	1	0	1	1	0	1	1	0	1

(c) Manipulation check by digital literacy.

Perceived Flaws	Article Preview (%)				"Oversampled Rep" comment (%)				"Oversampled Dem" comment (%)			
	1-3.1	3.1-3.9	3.9-4.5	4.5-5	1-3.1	3.1-3.9	3.9-4.5	4.5-5	1-3.1	3.1-3.9	3.9-4.5	4.5-5
I don't know	15	15	12	16	4	3	2	3	4	2	2	2
It surveyed more Democrats than Republicans	3	4	4	6	3	3	4	4	17	18	17	24
It surveyed more Republicans than Democrats	2	2	3	2	14	17	16	20	1	2	2	2
It was inaccurate in its prediction of the 2016 Presidential election	3	3	2	3	1	2	1	1	1	1	1	1
The poll was funded by liberals	1	1	1	2	1	1	0	1	1	1	1	1

correct responses among higher digital literacy subjects; in both of the commentary conditions, the subjects' digital literacy was unrelated to their chance of correctly responding.

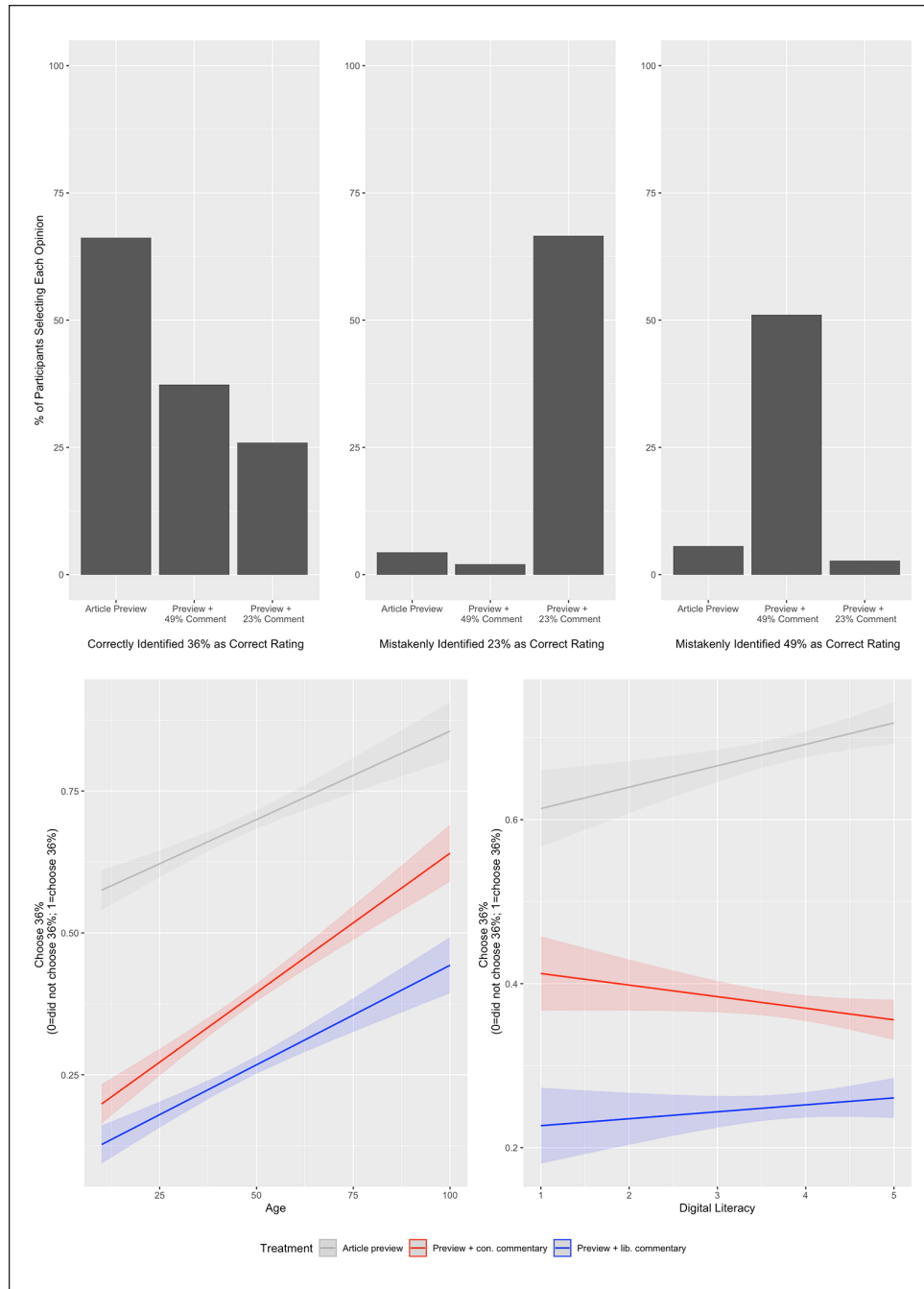
Overall, we found the expected treatment effect heterogeneity in both age and digital literacy on the disinformatinal effect of partisan commentary. However, this heterogeneity was in the opposite direction; older and less digitally literate people were *less* affected by the partisan commentary. This heterogeneity would not have been detectable without older or less digitally literate subjects in the sample.

### Haider-Markel and Joslyn experiment

The next results replicated the Haider-Markel and Joslyn priming experiment. The analysis here was

straightforward; the dependent variable was a 7-point scale measuring support for concealed handgun laws. Table 3 replicates Table 1 in the original, which uses an ordered logit to calculate the treatment effect of the gun safety frame. We added interaction terms for age and digital literacy. We also ran the analysis with ordinary least squares (OLS). Here, we did not expect to find treatment effect heterogeneity in either age or digital literacy.

We replicate the original finding, with a remarkably similar point estimate of the main treatment effect. As expected, we find that this treatment effect is *not* significantly moderated by either age or digital literacy, despite the large sample size.



**Figure 3.** Figure 2 from Anspach and Carlson (2018), replicated. Treatment effect heterogeneity: correctly reporting poll results.

### Messing and Westwood

The implementation of the Messing and Westwood replication was marred by a design flaw on the part of the authors. Focusing on the conceptual details of the project, we neglected a crucial practical detail: we did not randomize the order of headlines in the choice task, and headline order dominated any treatment effects of social and source cues.

We had, *ex ante*, considered our effort a “replication” in the sense defined by Nosek and Errington (2020): not an exact instantiation of the original protocol (because that is impossible) but a “conceptual” replication in the sense that researchers might reasonably consider each protocol to be testing the same theory. The results are presented below; however, we do not consider these results to have been produced by a procedure that can accurately be described as a replication.

**Table 3.** Table 1 from Haider-Markel and Joslyn (2001), replicated.  
(a) Regression with Age.

	Determinants for opposition to concealed handgun laws			
	Ordered Logit	Ordered Logit	OLS	OLS
Gun Rights Frame	-0.671*** (0.035)	-0.735*** (0.097)	-0.765*** (0.039)	-0.869*** (0.110)
Age	0.010*** (0.001)	0.009*** (0.001)	0.011*** (0.001)	0.010*** (0.002)
Party	-0.887*** (0.022)	-0.887*** (0.022)	-1.001*** (0.023)	-1.001*** (0.023)
Female	0.167*** (0.035)	0.167*** (0.035)	0.173*** (0.039)	0.173*** (0.039)
Gun Rights Frame * Age		0.001 (0.002)		0.002 (0.002)
1  2	-0.877*** (0.056)	-0.909*** (0.071)		
2  3	-0.373*** (0.055)	-0.404*** (0.071)		
3  4	0.153*** (0.055)	0.121* (0.071)		
4  5	0.913*** (0.055)	0.882*** (0.071)		
5  6	1.453*** (0.056)	1.422*** (0.072)		
6  7	1.961*** (0.058)	1.929*** (0.073)		
Constant			3.419*** (0.061)	3.470*** (0.079)
Observations	10,759	10,759	10,759	10,759
R			0.185	0.185
Adjusted R			0.184	0.184
Residual Std. Error			2.003 (df = 10754)	2.003 (df = 10753)
F Statistic			609.200*** (df = 4; 10754)	487.569*** (df = 5; 10753)

Note: \*p0.1; \*\*p0.05; \*\*\*p0.01.

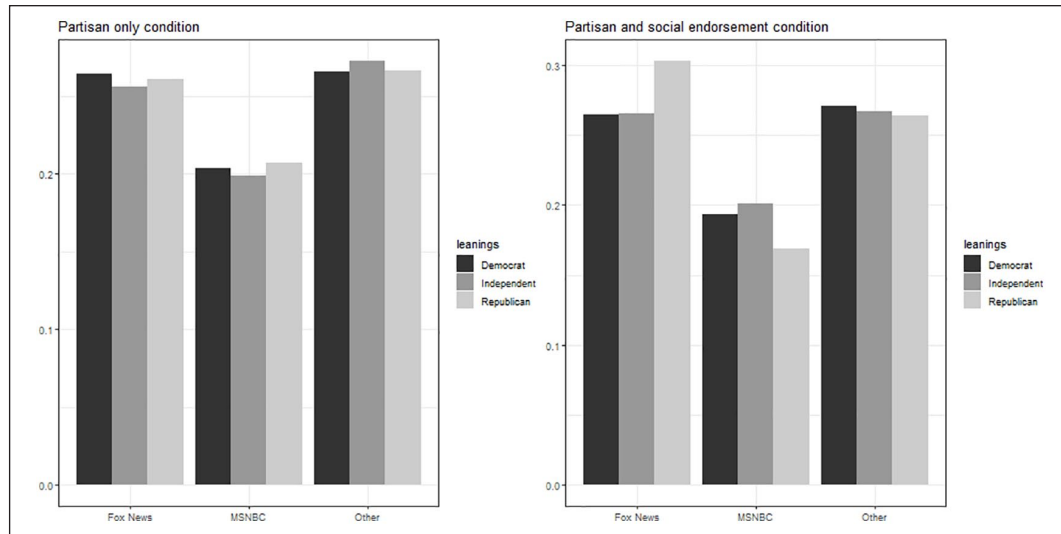
(b) Regression with Digital Literacy.

	Determinants for opposition to concealed handgun laws			
	Ordered Logit	Ordered Logit	OLS	OLS
Gun Rights Frame	-0.666*** (0.035)	-0.651*** (0.139)	-0.764*** (0.039)	-0.811*** (0.157)
Digital Literacy	-0.041** (0.018)	-0.039 (0.026)	-0.039* (0.021)	-0.045 (0.029)
Party	-0.861*** (0.022)	-0.861*** (0.022)	-0.980*** (0.023)	-0.980*** (0.023)
Female	0.179*** (0.035)	0.179*** (0.035)	0.186*** (0.040)	0.186*** (0.040)
Gun Rights Frame * Digital Lit		-0.004 (0.036)		0.012 (0.040)
1  2	-1.472*** (0.080)	-1.464*** (0.105)		
2  3	-0.968*** (0.079)	-0.960*** (0.104)		
3  4	-0.443*** (0.079)	-0.436*** (0.104)		
4  5	0.314*** (0.079)	0.321*** (0.104)		
5  6	0.849*** (0.079)	0.856*** (0.104)		
6  7	1.351*** (0.080)	1.358*** (0.105)		
Constant			4.067*** (0.087)	4.090*** (0.116)
Observations	10,759	10,759	10,759	10,759
R			0.178	0.178
Adjusted R			0.177	0.177
Residual Std. Error			2.012 (df = 10754)	2.012 (df = 10753)
F Statistic			581.003*** (df = 4; 10754)	464.782*** (df = 5; 10753)

Note: \*p0.1; \*\*p0.05; \*\*\*p0.01.

Nevertheless, we provide details of our analysis. We first coded the articles selected across the three different treatment conditions, averaging across the four choice sets. This allowed us to conduct a baseline replication of the original findings. Figure 4 replicates Figure 2 from the original paper, plotting the average rate of stories from different sources selected by partisans, comparing the “partisan only” condition with the “partisan plus social endorsement” condition.

We do *not* replicate the original findings; indeed our results are entirely inconsistent with them. Next, we checked to see if age or digital literacy significantly moderated the treatment effects. Using the template from the original Figure 3, Figure 5 calculates the average treatment effect of strong social endorsements: the likelihood that respondents select the story with over 10,000 endorsements. Messing and Westwood’s analysis looks for heterogeneities depending on



**Figure 4.** Figure 2 from Messing and Westwood (2014), replicated.

whether there is agreement, disagreement, or neither, between the respondent's partisanship and the partisan leaning of the stimulus; because they found little evidence for effect heterogeneity, we did not conduct this analysis. Instead, the rug plots in Figure 5 report how this treatment effect varies across different levels of age and digital literacy.

We do find evidence of significant heterogeneity. Again, however, our results are unexpected. Older people are *less* likely to select the headline with High social endorsements in the top left panel (source and endorsement condition) but older people are *more* likely to select the High-endorsement headline in the top right panel (endorsement-only condition). These findings are not internally coherent, but instead reflect the fact that the Social and Endorsement conditions were, through our oversight, colinear with the order of the headlines.

Digital literacy is only significant in the bottom left panel, but, again, the *direction* of the effect is flipped in the bottom right panel.

## Conclusion

Generalizability (or external validity) is one of the most pressing topics in social science. While our methods have become more credibly able to produce locally valid knowledge, the transportation of this knowledge to novel contexts has not yet been treated as rigorously. Influential recent work demonstrates that the results of a class of survey experiments can be replicated in different contexts, and that the results from a common convenience sample (MTurk) are generalizable.

The larger question is the scope of this result—that is, how generalizable is evidence of generalizability? The replications conducted here aim to establish scope conditions for the generalizability of survey experiments by

demonstrating that there are subgroups of the population (older, less digitally literate) and classes of survey experiments (those mediated by knowledge of digital media) for which MTurk cannot produce generalizable results.

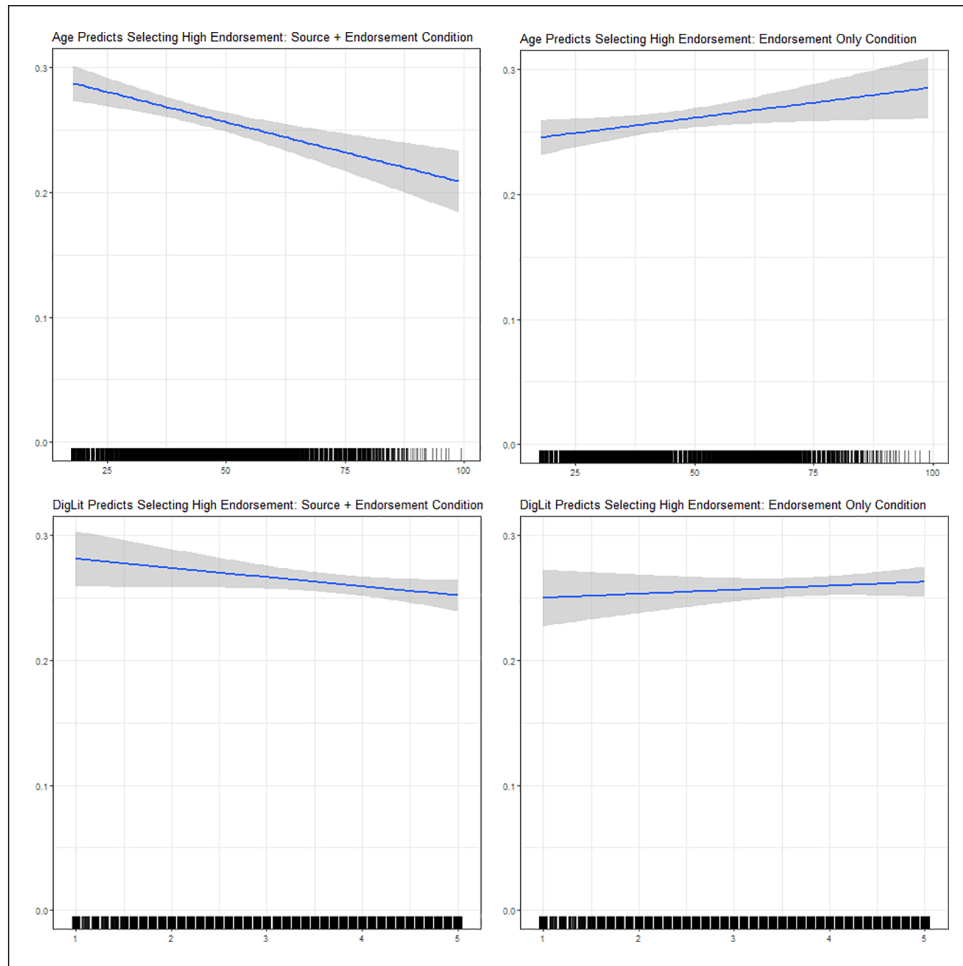
No individual study is supposed to “generalize” on its own, of course; the goal is to conduct many similar studies, aggregate their results, and then apply the knowledge gained to a novel context. For this to work, however, there must be overlap between the initial studies and the target context in the relevant covariates—that is, in the covariates for which there is treatment effect heterogeneity.

We show in this paper that a base of knowledge about the effect of digital media generated through experiments conducted on MTurk will not be able to generalize to the population of Internet users because MTurk does not contain enough respondents who are older or lower in digital literacy (Hargittai and Shaw, 2020). Due to the evolving novelty of the political experiences afforded by digital communication technology, we advocate that particular attention be paid to the composition of samples used to research these topics.

In our replication of Anspach and Carlson (2018) on a much larger sample, we find evidence that the main effects of social commentary on Facebook posts they study are significantly moderated by the age and digital literacy of respondents.

This critique does not—and indeed cannot, from the evidence presented in a single article like this one—apply to all social science research using MTurk samples. Previous research on the generalizability of studies from MTurk has necessarily covered only a subset of possible research questions, and it thus does not apply to all social science research using MTurk samples, either. We leave as an open question whether a given class of research designs should be “generalizable from MTurk samples until proven otherwise,” or





**Figure 5.** Figure 3 from Messing and Westwood (2014), replicated.

whether the burden of proof is in the opposite direction. In either case, we argue that any research on the topic of generalizability should include evidence of scope conditions: examples of research designs that are not generalizable from a given context, as well as examples that are.

We begin amassing evidence of these scope conditions through our “placebo” replication of Haider-Markel and Joslyn (2001). Here, we had no theoretical reason to expect that the treatment effect of priming one aspect of the gun control debate would be heterogeneous in subject age or digital literacy; our well-powered replication finds no evidence of these treatment effect heterogeneities.

In the course of conducting replications of research on digital media effects, there are unavoidable issues that are as much philosophical as methodological. The Facebook interface changes in important ways over time, so “exact replication” is impossible. However, following Nosek and Errington’s (2020) definition of replication, this is the context when replication is most important: many potentially important parameters were never tested in the past because the current state of the world did not previously exist.

By showing subjects a stimulus that reflects the current Facebook design, we test the theory from Anspach and Carlson (2018) in the current parameter space. Our results agreed with theirs; this means that our expectation about the parameter space in which the theory holds should be expanded. If we had found the opposite, results that differed from theirs, we would conclude the opposite: that our expectation about the parameter space in which the theory holds should be reduced.

The wrinkle introduced by studying rapidly changing and technologically mediated parameters like the interface of a given social media platform or the user population of MTurk/Lucid is that there is no guarantee that the parameter space will remain “similar enough” within the course of the process of knowledge accumulation across studies that these studies can in fact be said to continue to constitute “replications.” Facebook might, for example, adopt a “dislike” button like YouTube has, giving additional avenues for negative commentary. Booming economic conditions might cause all but a few US citizens to abandon being recruited by MTurk. Indeed, as happened in the process of

conducting the current study, the population of respondents recruited from Lucid changed dramatically, possibly due to the COVID-19 pandemic (Aronow et al., 2020).

These are challenging issues. As social science practice continues to improve (correctly emphasizing the need for replication, which remains an essential tool) and as more of human behavior takes place in online contexts that can change more rapidly and fundamentally than the physical world, however, these issues will only grow more pressing.


### Declaration of conflicting interests

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

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by the John S. and James L. Knight Foundation through a grant to Stanford University's Project on Democracy and the Internet.

### ORCID iDs

Kevin Munger  <https://orcid.org/0000-0002-4399-5250>

Joshua A. Tucker  <https://orcid.org/0000-0003-1321-8650>

### Supplemental materials

The supplementary files are available at: <http://journals.sagepub.com/doi/suppl/10.1177/20531680211016968>.

### Notes

- See, for example, Barberá (2018), Bond et al. (2012, 2017), Grinberg et al. (2019), and Guess et al. (2019), which are summarized in the following section.
- Alternatively, we might say we're looking for the mechanism by which age moderates treatment effects. If we could somehow specify all of the mechanisms and there were a residual effect, that would be the effect of age per se. This distinction is interesting but too complicated to resolve here.
- A recent paper by Hargittai and Shaw (2020) provides evidence for the non-representativeness of MTurk in terms of online experiences and digital literacy.
- Messing and Westwood (2014) do not control for age in their analysis. Our suspicion is that age played a significant role even in this restricted sample, but we cannot be sure.
- The original experiment also included a condition in which the respondents were asked to read the entire article, but was found to have similar effects to (C). To minimize the risk of attrition, we removed this condition.
- Figure 2 presents the stimuli actually used in the experiment, but we note that these are not identical to the stimuli in the registered report. The latter included images associated with each article rather than images associated with the source in panels (a) and (c). We noticed this discrepancy after the registered report was accepted and opted for this slight modification to better match the original experimental setup.
- Note that Anspach and Carlson include measures of "Need for Affect" and "Need for Cognition"; in light of the fact that they find no effect heterogeneity along these measures and due to space constraints, we did not include these measures in our analysis.

### Carnegie Corporation of New York Grant

This publication was made possible (in part) by a grant from the Carnegie Corporation of New York. The statements made and views expressed are solely the responsibility of the author.

### References

- Anspach NM and Carlson TN (2020) What to believe? Social media commentary and belief in misinformation. *Political Behavior* 42: 697–718.
- Aronow PM, Kalla J, Orr L, et al. (2020) Evidence of rising rates of inattentiveness on Lucid in 2020. *Investment Weekly News Preprint*: 402.
- Barberá P (2018) Explaining the spread of misinformation on social media: Evidence from the 2016 US Presidential Election. In Symposium: Fake News and the Politics of Misinformation. American Political Science Association. Available at: <http://pablobarbera.com/static/barbera-CP-note.pdf>
- Berinsky AJ, Margolis MF, Sances MW, et al. (2019) Using screeners to measure respondent attention on self-administered surveys: Which items and how many? *Political Science Research and Methods* 9(2): 430–437.
- Blair G, Cooper J, Coppock A, et al. (2019) Declaring and diagnosing research designs. *American Political Science Review* 113(3): 838–859.
- Bond RM, Fariss CJ, Jones JJ, et al. (2012) A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415): 295.
- Bond RM, Settle JE, Fariss CJ, et al. (2017) Social endorsement cues and political participation. *Political Communication* 34(2): 261–281.
- Brewer R, Morris MR and Piper AM (2016) "Why would anybody do this?" Understanding older adults' motivations and challenges in crowd work. In: Proceedings of the 2016 CHI Conference on human factors in computing systems, San Jose, USA, 7–12 May 2016. Association for Computing Machinery, pp.2246–2257.
- Coppock A (2018) Generalizing from survey experiments conducted on mechanical Turk: A replication approach. *Political Science Research and Methods* 7(3): 613–628.
- Coppock A and McClellan OA (2019) Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics* 6(1): 1–14.
- Druckman JN and Kam CD (2011) Students as experimental participants. In: Druckman JN, Green DP, Kuklinski JH, et al (eds) *Cambridge Handbook of Experimental Political Science*. New York: Cambridge University Press, pp.41–57.
- Grinberg N, Joseph K, Friedland L, et al. (2019) Fake news on Twitter during the 2016 US presidential election. *Science* 363(6425): 374–378.
- Guess A and Munger K (2020) To see what's inside of one's screen: Digital literacy and online political behavior. Unpublished manuscript.

- Guess A, Nagler J and Tucker J (2019) Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5(1).
- Haider-Markel DP and Joslyn MR (2001) "Gun policy, opinion, tragedy, and blame attribution: The conditional influence of issue frames." *Journal of Politics* 63 (2): 520–543.
- Hargittai E (2002) Second-level digital divide: Differences in people's online skills. *First Monday* 7(4).
- Hargittai E (2005) Survey measures of web-oriented digital literacy. *Social Science Computer Review* 23(3): 371–379.
- Hargittai E (2015) Is bigger always better? Potential biases of big data derived from social network sites. *ANNALS of the American Academy of Political and Social Science* 659(1): 63–76.
- Hargittai E and Hsieh YP (2012) Succinct survey measures of web-use skills. *Social Science Computer Review* 30(1): 95–107.
- Hargittai E and Shaw A (2020) Comparing internet experiences and prosociality in amazon Mechanical Turk and population-based survey samples. *Socius* 6: 1–11.
- Hargittai E, Fuchslin T and Schäfer MS (2018) How do young adults engage with science and research on social media? Some preliminary findings and an agenda for future research. *Social Media + Society* 4(3): 2056305118797720.
- Huff C and Tingley D (2015) "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research and Politics* 2(1): 1–12.
- Liao QV and Fu W-T (2014) Age differences in credibility judgments of online health information. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21(1): 1–23.
- Messing S and Westwood SJ (2014) Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication Research* 41(8): 1042–1063.
- Mullinix KJ, Leeper TJ, Druckman JN, et al. (2015) The generalizability of survey experiments. *Journal of Experimental Political Science* 2(2): 109–138.
- Nosek BA and Errington TM (2020) What is replication? *PLoS biology* 18(3): e3000691.
- Smith A and Anderson M (2018) *Social media use in 2018*. Pew. Available at: <https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/> (accessed December 2020).