

Mechanical Turk and the “Don’t Know” Option

Adam R. Brown, *Brigham Young University*

Jeremy C. Pope, *Brigham Young University*

ABSTRACT

Luskin and Bullock’s (2011) randomized experiment on live-interview respondents found no evidence that American National Election Studies and Time-Sharing Experiments for the Social Sciences respondents hide knowledge behind the “don’t know” (DK) option. We successfully replicated their finding using two online platforms, the Cooperative Congressional Election Study and Google Surveys. However, we obtained different results on Amazon’s Mechanical Turk (MTurk). We attribute this difference to MTurkers’ experience with attention checks and other quality-control mechanisms, which condition them to avoid errors. This conditioning leads MTurkers to hide knowledge behind DK in ways not observed on other platforms. Researchers conducting political knowledge experiments or piloting surveys on MTurk should be aware of these differences.

A long line of research demonstrates the public’s political ignorance (Achen and Bartels 2016; Berelson, Lazarsfeld, and McPhee 1954; Clifford and Jerit 2016; Converse 1964; Delli Carpini and Keeter 1996; Luskin and Bullock 2011), yet several studies suggest that the public may be more knowledgeable than we think (Krosnick et al. 2002; Mondak 2000; 2001; Mondak and Davis 2001; Nie, Verba, and Petrocik 1979; Popkin 1991; Prior and Lupia 2008). In part, this debate hinges on the presence or absence in surveys of a “don’t know” (DK) option, with some scholars contending that respondents hide some of their knowledge behind DK responses.


Luskin and Bullock (2011) presented an experiment showing that including or excluding DK makes little practical difference to estimates of knowledge. They conducted their experiment on two platforms; we replicated it on three more. On one of these platforms—Amazon’s MTurk—respondents appear to hide significant political knowledge behind DK. We attribute this difference to MTurkers’ experience with attention checks and other quality-control mechanisms, which condition them to scrupulously avoid errors. This conditioning, unique to MTurk, suggests a need for more caution among researchers who pilot questions or administer experiments on this platform, at least for certain types of questions.

THEORY

Mondak and colleagues argued that respondents conceal knowledge behind DK responses and once advocated, in some cases, rescoring DK responses as correct or partially correct to produce higher, more accurate estimates of public knowledge (Mondak 2000; 2001; Mondak and Anderson 2004; Mondak and Davis 2001). They argued that neutral options (for opinion questions) and the DK option (for knowledge questions) lead respondents to “satisfice,” avoiding the cognitive work of answering a question even if they are capable of a better response. Concurring, Krosnick et al. (2002, 373) asked “whether offering a no-opinion option attracts only respondents who would otherwise have offered meaningless responses, or whether offering a no-opinion option also attracts respondents who truly have opinions and would otherwise have reported them.”

Other researchers counter that “discouraging DKs reveals precious little hidden knowledge” (Luskin and Bullock 2011, 554; see also Tourangeau, Maitland, and Yan 2016). “When people who initially select a DK alternative are subsequently asked to provide a ‘best guess,’ they fare statistically no better than chance” (Sturgis, Allum, and Smith 2008, 90). Luskin and Bullock (2011) drew their conclusions from a randomized experiment, largely replicated here. In what we label the *encourage guessing* condition, Luskin and Bullock prefaced a battery of political knowledge questions with this statement: “If you aren’t sure of the answer to any of these questions, we’d be grateful if you could just give your best guess.” In the *encourage DK* condition, they included this preface: “Many people have trouble answering questions like these. So, if you can’t think of the answer, don’t worry about it. Just mark that you don’t know and move on to the next one.”¹

Adam R. Brown is associate professor of political science at Brigham Young University. He can be reached at brown@byu.edu.

Jeremy C. Pope  is professor of political science at Brigham Young University. He can be reached at jpope@byu.edu.

© The Author(s), 2021. Published by Cambridge University Press on behalf of the

American Political Science Association
Published online by Cambridge University Press

All respondents then answered the same battery, with DK always an option. This subtle treatment indeed influenced respondents, who chose DK less often and answered correctly more often in the *encourage guessing* condition. However, this decrease in DK

Over time, these mechanisms condition MTurkers to pay closer attention to detail than users on other platforms (Hauser and Schwarz 2016). MTurkers thus are not only politically and demographically distinct from users on other platforms; they also are conditioned to

This conditioning, unique to MTurk, suggests a need for more caution among researchers who pilot questions or administer experiments on this platform.

responses did not yield a greater increase in correct answers than could be attributed to guessing. They therefore concluded that including a DK option was harmless—that is, respondents were not hiding significant knowledge behind the neutral option.

Luskin and Bullock (2011) administered their experiment on two live-interviewer platforms: the American National Election Studies (ANES) and Time-Sharing Experiments for the Social Sciences (TESS). Much public opinion research now takes place online, where the absence of a live interlocutor fundamentally changes social incentives. We therefore replicated their experiment on three online platforms: the Cooperative Congressional Election Study (CCES), Google Surveys (GS), and MTurk. All platforms have their unique features, of course. The CCES presents a lengthy political survey, with some respondents compensated through points and rewards in the YouGov system. In contrast, GS presents brief pop-up surveys to Internet users attempting to load unrelated websites; these respondents answer the questions simply to make the survey go away so they can view their desired content. Replicating Luskin and Bullock's (2011) experiment on these diverse platforms provides an important check on their results.

More to the point, however, we also replicated their experiment on MTurk, the most distinctive platform. MTurk makes no attempt to recruit a representative user base, resulting in well-known demographic peculiarities. Nevertheless, scholars have replicated several published experiments on MTurk, reassuring researchers of the platform's reliability (Ansolabehere and Schaffner 2014; Berinsky, Huber, and Lenz 2012). Huff and Tingley (2015, 8) therefore concluded that MTurkers are "not all that different from respondents on other survey platforms," particularly when analyzed by demographic subgroup.

Nevertheless, MTurkers face unique incentives that may affect their behavior in subtle ways, especially concerning neutral response options like DK. Most MTurk jobs are not survey

avoid errors, making neutral response options more attractive. Survey weights and subgroup analysis might correct MTurk's demographic skew, but they cannot account for these conditioned behaviors. Unique among respondent pools, MTurk "is a population that learns" (Hauser and Schwarz 2016). Because MTurkers are conditioned to avoid errors, they may calculate that it is better to hide behind the DK option when they have any uncertainty about their response, making them distinct from respondents recruited through other platforms.

DESIGN

To test this hypothesis, we replicated Luskin and Bullock's (2011) experiment across three platforms by administering a battery of political knowledge questions to CCES, MTurk, and GS respondents, always with a DK option (Jones 2021).² (Methodological details specific to each platform are footnoted.³) We randomly assigned respondents to Luskin and Bullock's two conditions. Before viewing the knowledge battery, half of the respondents were encouraged to guess if they did not know an answer; the other half was encouraged to mark "DK." Our battery included the following five items:

- To the best of your knowledge, does your state have its own constitution?
- Is the US federal budget deficit—the amount by which the government's spending exceeds the amount of money it collects—now bigger, about the same, or smaller than it was during most of the 1990s?
- For how many years is a US Senator elected—that is, how many years are there in one full term of office for a US Senator?
- On which of the following does the US federal government currently spend the least? (Options: foreign aid, Medicare, national defense, Social Security.)
- Who nominates judges to the Supreme Court? (Options: the President, the House of Representatives, the Senate, the Supreme Court.)

Nevertheless, MTurkers face unique incentives that may affect their behavior in subtle ways, especially concerning neutral response options like DK.

research but rather so-called "human intelligence tasks" such as transcribing text from images and completing other simple work. Job providers can accept or reject a user's work, and users' resulting ratings affect their ability to receive future assignments. Similarly, market research and social science surveys posted to MTurk regularly include attention checks or similar quality-control devices (Peer, Vosgerau, and Acquisti 2014). Poor performance on one task directly affects an MTurk user's ability to earn money in the future.

RESULTS

Among CCES respondents, 69% correctly answered that their state had a constitution, 63% answered that the deficit had grown, 49% answered that US Senators serve for six years, 28% identified foreign aid as the federal government's smallest expenditure, and 73% said that the President nominates Supreme Court judges. The mean CCES respondent answered 2.8 of five items correctly—the same as GS respondents but lower than the 3.1 mean among MTurkers. As shown in table 1, these baseline platform differences

Table 1
Ordinary Least Squares Estimates of Total Items Correct and DK Responses

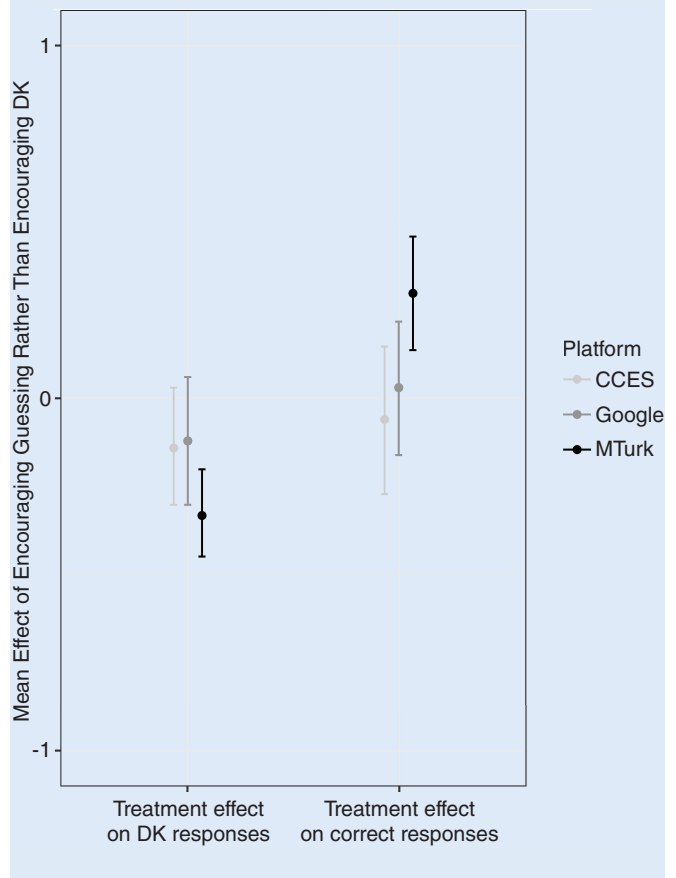
	Correct	Don't Know
CCES	-0.36* (0.067)	0.18* (0.057)
Google Surveys	-0.26* (0.065)	0.10 (0.055)
Male	0.70* (0.051)	-0.36* (0.043)
Democrat	0.28* (0.067)	-0.25* (0.057)
Republican	0.37* (0.073)	-0.34* (0.062)
Four-Year College Degree	0.60* (0.052)	-0.33* (0.044)
Age 25–34	0.13 (0.086)	-0.17* (0.073)
Age 35–44	0.36* (0.095)	-0.32* (0.080)
Age 45–54	0.57* (0.099)	-0.35* (0.084)
Age 55–64	0.68* (0.099)	-0.32* (0.084)
Age 65+	0.89* (0.11)	-0.48* (0.092)
Constant	1.9* (0.098)	1.6* (0.083)
N	2,521	2,521
R ²	0.17	0.08

Notes: * $p \leq 0.05$ (two-tailed). OLS coefficients shown with standard errors in parentheses. MTurk is the omitted platform; females, independents, and respondents ages 18–24 are the omitted categories. Rounding is to two significant digits.

persist even after controlling for demographic differences in ordinary least squares (OLS) regression.⁴ CCES and GS respondents gave fewer correct answers (first column) and more DK responses (second column) than MTurk respondents. Perhaps MTurkers’ frequent participation in social science research makes them a more knowledgeable group overall.

Because participants were assigned randomly into conditions, we did not include demographic controls when estimating treatment effects.⁵ Figure 1 (left panel) summarizes the average treatment effect of *encourage guessing* (as opposed to *encourage DK*) on correct and DK responses.⁶ On all platforms, *encourage guessing* reduced DK responses relative to *encourage DK*, although the effect was significant only for MTurk respondents (-0.33, $p < 0.01$). The reduction was -0.14 ($p = 0.11$) for CCES and -0.12 ($p = 0.21$) for GS respondents. Random guessing alone would yield an approximate 32% accuracy rate.⁷ If respondents on all platforms converted their reduced DK responses to truly random guesses, we would expect meaningless increases in correct responses of 0.11 (MTurk), 0.045 (CCES), and 0.038 (GS).⁸ For GS, that is almost exactly what we

Figure 1
Mean Effect of Encouraging Guessing Rather Than Encouraging DK



found: an insignificant 0.030 ($p = 0.75$) increase in correct responses under the *encourage guessing* condition. Curiously, CCES respondents appear to have provided (insignificantly) fewer correct responses under *encourage guessing* (-0.062, $p = 0.56$).⁹ To summarize, CCES and GS respondents saw even smaller effects of the *encourage guessing* treatment than Luskin and Bullock (2011) originally reported, but the overall pattern clearly supports their general contention that including or omitting the DK option does not change estimates of knowledge in the sample. When we observed a marginal increase in correct responses (on GS), it was not greater than can be attributed to random guessing, again affirming Luskin and Bullock’s (2011) general argument.

Among MTurk respondents, however, a substantially different picture emerges. On this platform only, *encourage guessing* raised average scores by 0.30 relative to *encourage DK* ($p < 0.01$)—far greater than the 0.11 increase we expected from random guessing and almost the exact amount by which *encourage guessing* reduced DK responses (see figure 1, right panel). On its face, this result implies that nearly every MTurk respondent induced to guess rather than mark DK ultimately marked a correct answer instead—although we note that the 95% confidence interval around our +0.30 estimate extends as low as 0.14. At least some MTurk respondents clearly responded to the *encourage guessing* treatment, giving more correct responses than could be obtained by chance alone.

We found no evidence that MTurk respondents were more likely to search for answers online in one condition than in another, behavior that could produce these results spuriously. In both of our experimental conditions, the 25th, 50th, and 75th

that the presence or absence of a DK option makes little difference.¹² Nevertheless, the muted but slightly different response pattern reveals the importance of context and the need to be cautious in how our claims generalize.

If MTurk users react differently than users on other platforms to neutral options, then researchers should be aware of their unique properties and characteristics when designing surveys and survey experiments.

percentiles of elapsed time were identical—55, 72, and 101 seconds, respectively—meaning that respondents did not spend more time answering our knowledge battery in one condition when compared to the other.¹⁰ We conclude that at least some MTurkers hide knowledge behind the DK option, in clear contrast to the other platforms.¹¹

CONCLUSION

In an experiment administered using live interviewers, Luskin and Bullock (2011) found no evidence that ANES or TESS respondents hide knowledge behind the DK option. Their treatment successfully induced people to guess rather than choose DK, but this guessing did not reveal concealed knowledge. We arrived at a similar conclusion using two online platforms, CCES and GS. However, MTurk respondents behave differently. Perhaps the attention checks, accuracy bonuses, and other quality-control devices frequently employed on MTurk condition its users to select neutral options unless they are certain of their response. In any event, MTurkers appear to hide some knowledge behind the DK option—even though our experiment did not use any of the attention checks or accuracy bonuses common to MTurk surveys. It therefore might make sense to omit the DK option when using MTurk.

On MTurk only, inducing respondents to guess not only reduces DK responses but also increases correct responses in a nearly one-to-one relationship. Our results do not call into question MTurk's general utility as a research platform, but they do suggest caution concerning studies of political knowledge specifically and the use of neutral response options generally. If MTurk users react differently than users on other platforms to neutral options, then researchers should be aware of their unique properties and characteristics when designing surveys and survey experiments. We do not dispute the general conclusion of Luskin and Bullock (2011) on most platforms, but the choice of MTurk as a research platform complicates decisions about when to use the DK option and likely points to the need to consider how different experimental manipulations may vary across platforms.

For instance, we note that our CCES and GS respondents took less notice of our manipulation generally than Luskin and Bullock's (2011) ANES and TESS respondents. Their respondents significantly decreased their DK responses under *encourage guessing* and significantly increased their correct responses—albeit not sufficiently to rule out random guessing. By contrast, we observed smaller decreases in DK responses among CCES and GS respondents than Luskin and Bullock (2011) reported and no measurable increase in correct responses. Unlike our MTurk results, this pattern supports Luskin and Bullock's (2011) broader conclusion

SUPPLEMENTARY MATERIALS

To view supplementary material for this article, please visit <http://dx.doi.org/10.1017/S1049096520001651>.

DATA AVAILABILITY STATEMENT

Replication materials can be found on Dataverse at <https://doi.org/10.7910/DVN/KAPQWT>. ■

NOTES

1. We omit Luskin and Bullock's (2011) third neutral condition, which contained no preface.
2. Although we preserved Luskin and Bullock's (2011) treatment prompts, we adapted the subsequent knowledge battery to contain items relevant to our other projects. We also tried varying whether the interface allowed respondents to skip items (among CCES respondents only) but found so little variation in item nonresponse across experimental conditions (i.e., a difference of 0.0009 items skipped out of five; $p=0.98$) that we did not include it here. In both conditions, two thirds of respondents answered all five items and 94% answered at least four.
3. (1) MTurk: Participants received \$0.40 for participating, no matter their performance; we did not financially incentivize scores. The middle 50% of respondents took between 55 and 101 seconds to complete the survey, which also included a brief demographic battery, implying a fair hourly wage of \$14–\$26 per hour. (2) GS: The strict format imposed by GS on researchers required us to include the treatment prompt as the first of 10 questions, with respondents acknowledging "I understand" to proceed or marking "I prefer not to participate" to exit. Requiring explicit acknowledgment of the treatment prompt might be expected to increase treatment effects in this pool; as it happens, effects were weakest among these subjects. Respondents then answered the five-item knowledge battery, followed by four demographic questions. (3) CCES: Respondents answered the knowledge battery as part of a module that appeared in the 2016 survey, after they had already completed the survey's common content. (4) Although weights are available for CCES and GS respondents, we present unweighted results according to Franco et al. (2017). See the online supplemental appendix for weighted results.
4. Models in this table pooled respondents assigned to both conditions. Additional models, including negative binomial regression, are in the online supplemental appendix.
5. From an abundance of caution, we tested whether age, sex, party affiliation, and college education predicted group assignment within each platform. There were no significant relationships. We also estimated treatment effects using models that control for these demographic variables, with no meaningful differences from the effects reported here. See the online supplemental appendix.
6. The online supplemental appendix contains several figures and tables expanding on the results from figure 1. Figure A1 plots the mean number of DK responses by condition and platform; figure A2 plots the mean number of correct responses. Using OLS regression to interact our treatment and platform variables, table A3 shows that our treatment has a significantly different effect on the number of correct responses when comparing MTurk to either CCES ($p<0.01$ two-tailed) or GS ($p=0.04$)—results that persist (CCES $p<0.01$, GS $p=0.07$) with demographic controls. Other models in table A3 show similar differences when predicting the number of DK responses: MTurk versus CCES ($p=0.06$ with controls, $p=0.09$ without) and versus GS ($p=0.04$ with controls, $p=0.06$ without). Table A5 yields similar results using negative binomial rather than OLS regression. Tables A4 and A6 replicate tables A3 and A5, respectively, but with the addition of survey weights for our CCES and GS respondents. Adding these weights attenuates the differences between MTurk and CCES when predicting DK responses. Because we are presenting results from a randomized experiment, we followed advice from Franco et al. (2017) in favoring unweighted over weighted analysis.
7. A respondent guessing randomly would have a 50% chance of answering the first item correctly, 33% for the second, 25% for the fourth, and 25% for the fifth item. As for the third item, the query about Senator term length was open-ended; nevertheless,

- nearly all respondents (99%) answered 2, 4, 6, or 8 years; conservatively, therefore, we estimated a 25% chance of correctly guessing. Thus, random guessing would produce an average score of 1.58/5 (32%). If the 0.35 MTurk reduction in DK responses led only to random guessing, we would expect correct responses to rise by only 0.11 (0.35*32%). The observed increase of 0.30 differs significantly.
8. For these predictions, we multiplied each platform’s reduction in DK rates by 32%.
 9. When CCES weights were applied (see the online supplemental appendix), *encourage guessing* was associated with a -0.35 ($p < 0.01$) change in DK responses and a +0.10 ($p = 0.48$) change in correct responses. Because $0.35 \times 32\% = 0.11$, this +0.10 is almost exactly what we would predict as a result of random guessing.
 10. The means differed slightly across conditions: 97 seconds in *encourage guessing* versus 95 seconds in *encourage DK*. Comparison of means tests returned an insignificant result using either raw ($p = 0.75$ two-tailed) or logged ($p = 0.84$) times. We cannot evaluate whether MTurk’s conditioning leads those respondents to search online at higher rates (in both conditions) than respondents on other platforms.
 11. Following Pietryka and MacIntosh (2013), we investigated whether women and men behaved differently with respect to the treatments. Although gender affects responses generally (see table 1), it did not interact with our treatment on any platform; p -values were consistently above 0.3.
 12. More precisely, this pattern supports their conclusion that the presence or absence of a DK option makes little difference for identifying respondents who could answer correctly by way of full knowledge or educated guesses, but also that the DK option matters for the total number of correct answers recorded. We thank an anonymous reviewer for this clarification.

REFERENCES

Achen, Christopher H., and Larry M. Bartels. 2016. *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, NJ: Princeton University Press.

Ansolabehere, Stephen, and Brian Schaffner. 2014. “Does Survey Mode Still Matter? Findings from a 2010 Multimode Comparison.” *Political Analysis* 22:285–303.

Berelson, Bernard R., Paul F. Lazarsfeld, and William N. McPhee. 1954. *Voting: A Study of Opinion Formation in a Presidential Campaign*. Chicago: University of Chicago Press.

Berinsky, Adam J., Gregory A. Huber, and Gabriel S. Lenz. 2012. “Evaluating Online Labor Markets for Experimental Research: Amazon’s Mechanical Turk.” *Political Analysis* 20:351–68.

Clifford, Scott, and Jennifer Jerit. 2016. “Cheating on Political Knowledge Questions in Online Surveys: An Assessment of the Problem and Solutions.” *Public Opinion Quarterly* 80:858–87.

Converse, Philip. 1964. “The Nature of Belief Systems in Mass Publics.” In *Ideology and Discontent*, ed. David E. Apter, 206–61. New York: The Free Press of Glencoe.

Delli Carpini, Michael X., and Scott Keeter. 1996. *What Americans Know about Politics and Why It Matters*. New Haven, CT: Yale University Press.

Franco, Annie, Neil Malhotra, Gabor Simonovitz, and L. J. Zigerell. 2017. “Developing Standards for Post-Hoc Weighting in Population-Based Survey Experiments.” *Journal of Experimental Political Science* 4:161–72.

Hauser, David J., and Norbert Schwarz. 2016. “Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks Than Do Subject Pool Participants.” *Behavior Research Methods* 48 (1): 400–407.

Huff, Connor, and Dustin Tingley. 2015. “Who Are These People? Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents.” *Research and Politics* 2:1–15.

Krosnick, Jon A., Allyson L. Holbrook, Matthew K. Berent, Richard T. Carson, W. Michael Hanemann, Raymond J. Kopp, Robert Cameron Mitchell, Stanley Presser, Paul A. Ruud, V. Kerry Smith, Wendy R. Moody, Melanie C. Green, and Michael Conaway. 2002. “The Impact of ‘No Opinion’ Response Options on Data Quality.” *Public Opinion Quarterly* 66:371–403.

Luskin, Robert C., and John G. Bullock. 2011. “‘Don’t Know’ Means ‘Don’t Know’: DK Responses and the Public’s Level of Political Knowledge.” *Journal of Politics* 73 (2): 547–57.

Mondak, Jeffery J. 2000. “Reconsidering the Measurement of Political Knowledge.” *Political Analysis* 8:57–82.

Mondak, Jeffery J. 2001. “Developing Valid Knowledge Scales.” *American Journal of Political Science* 45:224–38.

Mondak, Jeffery J., and Mary R. Anderson. 2004. “The Knowledge Gap: A Reexamination of Gender-Based Differences in Political Knowledge.” *Journal of Politics* 66:492–512.

Mondak, Jeffery J., and Belinda Creel Davis. 2001. “Asked and Answered: Knowledge Levels When We Will Not Take ‘Don’t Know’ for an Answer.” *Political Behavior* 23:199–224.

Nie, Norman, Sidney Verba, and John R. Petrocik. 1979. *The Changing American Voter*. Cambridge, MA: Harvard University Press.

Peer, Eyal, Joachim Vosgerau, and Alessandro Acquisti. 2014. “Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk.” *Behavior Research Methods* 46 (4): 1023–31.

Pietryka, Matthew T., and Randall C. MacIntosh. 2013. “An Analysis of ANES Items and Their Use in the Construction of Political Knowledge Scales.” *Political Analysis* 21 (4): 407–29.

Pope, Jeremy. 2021. “Replication Data for: Mechanical Turk and the ‘Don’t Know’ Option.” Harvard Dataverse. doi: 10.7910/DVN/KAPQWT.

Popkin, Samuel L. 1991. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. Chicago: University of Chicago Press.

Prior, Markus, and Arthur Lupia. 2008. “Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills.” *American Journal of Political Science* 52 (1): 169–83.

Sturgis, Patrick, Nick Allum, and Patten Smith. 2008. “An Experiment on the Measurement of Political Knowledge in Surveys.” *Public Opinion Quarterly* 72: 90–102.

Tourangeau, Roger, Aaron Maitland, and H. Yanna Yan. 2016. “Assessing the Scientific Knowledge of the General Public: The Effects of Question Format and Encouraging or Discouraging Don’t Know Responses.” *Public Opinion Quarterly* 80:741–60.