Mechanical Turk and the "Don't Know" Option

Adam R. Brown and Jeremy C. Pope

Abstract. Luskin and Bullock's (2011) randomized experiment on live-interview respondents found no evidence that ANES or TESS respondents hide knowledge behind the "don't know" option (DK). We successfully replicate their finding using two online platforms, the Cooperative Congressional Election Study (CCES) and Google Surveys (GS). However, we obtain 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 on MTurk or piloting surveys on MTurk should take note of these differences.

A long line of research demonstrates the public's political ignorance (Berelson, Lazarsfeld and McPhee 1954; Converse 1964; Delli Carpini and Keeter 1996; Luskin and Bullock 2011; Achen and Bartels 2016; Clifford and Jerit 2016), yet several studies suggest that the public may be more knowledgeable than we think (Nie Verba, and Petrocik 1979; Popkin 1991; Mondak 1999, 2001; Mondak and Davis 2001; Krosnick et al. 2008; Prior and Lupia 2008). In part, this debate hinges on the presence or absence in surveys of a "don't know" option (DK), 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 replicate it on three more. On one of these platforms—Amazon's Mechanical Turk (MTurk)—respondents do 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 avoid errors scrupulously. This conditioning, unique to MTurk, suggests a need for more caution among researchers piloting questions or administering experiments on this platform, at least for certain types of questions.

Theory

Mondak and colleagues have 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 1991, 2001; Mondak and Davis 2001; Mondak and Anderson 2004). They argue 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) ask "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."

Others 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 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] All respondents then answered the same battery, with DK always an option. This subtle treatment indeed influenced respondents, who chose DK less and answered correctly more in the *encourage guessing* condition. However, this decrease in DK 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—respondents were not hiding significant knowledge behind the neutral option.

[1] We omit Luskin and Bullock's third, neutral condition, which contained no preface at all.

Luskin and Bullock 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 changes social incentives fundamentally. We therefore replicate 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, Google Surveys (GS) presents brief popup 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 experiment on these diverse platforms provides an important check on their results.

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

Still, MTurkers face unique incentives that may affect their behavior in subtle ways, especially when it comes to neutral response options like DK. Most MTurk jobs are not survey research, but so-called "human intelligence tasks" like transcribing text from images or 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.

Over time, these mechanisms condition MTurkers to pay closer attention to detail than users on other platforms (Hauser and Schwarz 2016). MTurkers are thus not only politically and demographically distinct from users on other platforms; they are also conditioned to 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.[2] Methodological details specific to each platform are footnoted.[3] We randomly assigned respondents to Luskin and Bullock's two conditions. Prior to

[2] Though we preserve Luskin and Bullock's treatment prompts, we adapt the subsequent knowledge battery to contain items relevant to our other projects. We also toyed with varying whether the interface allowed respondents to skip items (among CCES respondents only) but found so little variation in item non-response across experimental conditions (a difference of 0.0009 items skipped out of 5, p=0.98) that we do not detail it here. In both conditions, two-thirds of respondents answered all 5 items, while 94% answered at least 4.

[3] (1) MTurk. Participants received $0.40 for participating, no matter their performance; we did not incentivize scores financially. The middle 50% of respondents spent between 55 and 101 seconds completing the survey, which also included a brief demographic battery, implying a fair hourly wage of $14-26 per hour. (2) Google Surveys. The strict format imposed by Google Surveys 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.

viewing the knowledge battery, half were encouraged to guess if they did not know an answer, while half were encouraged to mark DK. Our battery included these 5 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 United States 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.)

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 6 years, 28% identified foreign aid as the federal government's smallest expenditure, and 73% said that the President nominates judges. The mean CCES respondent answered 2.8 of 5 items correctly—the same as GS respondents, but lower than the 3.1 mean among MTurkers. As shown in Table 1,

Requiring explicit acknowledgement of the treatment prompt might be expected to increase treatment effects in this pool; as it happens, effects are actually weakest among these subjects. Respondents then answered the 5-item knowledge battery, followed by 4 demographic questions. (3) CCES. Respondents answered the knowledge battery as part of a module that appeared in the 2016 survey, after respondents had already completed the survey's common content. (4) Though weights are available for CCES and GS respondents, we present unweighted results per Franco et al. (2017). See the supplement for weighted results.

these baseline platform differences persist even after controlling for demographic differences in OLS regression.[4] CCES and GS respondents give 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.

[Table 1]

Because participants were assigned randomly into conditions, we do 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* reduces DK responses relative to *encourage DK*, though the effect is significant only for MTurk (-0.33, $p<0.01$) respondents. The reduction is -0.14 ($p=0.11$) for CCES and -0.12 ($p=0.21$) for GS. Random guessing alone would yield approximately a 32% accuracy rate.[7] If respondents on all platforms converted their reduced DK responses into

[4] Models in this table pool respondents assigned to both conditions. Additional models, including negative binomial regression, appear in a supplemental appendix.

[5] Out of abundance of caution, we tested whether age, sex, party affiliation, and college education predict group assignment within each platform. There were no significant relationships. We also estimated treatment effects using models that do control for these demographic variables, with no meaningful differences from the effects reported here. See the supplemental appendix.

[6] The 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 ordinary least squares 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 vs CCES ($p=0.06$ with controls, $p=0.09$ without) and vs GS ($p=0.04$ with controls, $p=0.06$ without). Table A5 yields similar results using negative binomial rather than ordinary least squares 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 follow advice from Franco et al. (2017) in favoring unweighted over weighed 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. As for the third, its query about Senate term lengths was open-ended; nevertheless, nearly all respondents (99%) gave an answer of 2, 4, 6, or 8 years; conservatively, then, we estimate 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.

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 find: 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 sum up, CCES and GS respondents see even smaller effects of the *encourage guessing* treatment than Luskin and Bullock originally reported, but the overall pattern clearly supports Luskin and Bullock's general contention that including or omitting the DK option makes does not change estimates of knowledge in the sample. When we do observe a marginal increase in correct responses (on GS), it is not greater than can be attributed to random guessing, again affirming their general argument.

[Figure 1 here]

Among MTurk respondents, however, a substantially different picture emerges. On this platform only, *encourage guessing* raises average scores by 0.30 relative to *encourage DK* (p<0.01)—far greater than the 0.11 increase we would expect 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 wound up marking a correct answer instead—though it we note that the 95% confidence interval around our +0.30 estimate extends as low as 0.14. At least some MTurk respondents clearly do respond to the *encourage guessing* treatment, giving more correct responses than could be obtained by chance alone.

[8] For these predictions, we multiply each platform's reduction in DK rates by 32%.
[9] When CCES weights are applied (see the supplement), *encourage guessing* is associated with a -0.35 (p<0.01) change in DK responses and a +0.10 (p=0.48) change in correct responses. Because 0.35×32%=0.11, this +0.10 is almost exactly what we would predict as a result of random guessing.

We find 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 our experimental conditions the 25th, 50th, and 75th percentiles of elapsed time were identical—55, 72, and 101 seconds respectively—meaning that respondents did not spend longer 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 arrive 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 employed frequently 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 might therefore make sense to omit the DK option when using MTurk.

[10] The means differed slightly across conditions: 97 seconds in *encourage guessing* versus 95 seconds in *encourage DK*. Comparison of means tests return 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] We investigated whether women and men behaved differently with respect to the treatments, following Pietryka and MacIntosh (2013). Though gender affects responses generally (see Table 1), it did not interact with our treatment on any platform, with p-values consistently above 0.3.

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 when it comes to studies of political knowledge specifically and the use of neutral response options generally. If MTurk users react differently to neutral options than users on other platforms, then researchers should be aware of their unique properties and characteristics when designing surveys or 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 probably 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 did Luskin and Bullock's ANES and TESS respondents. Luskin and Bullock's respondents significantly decreased their DK responses under *encourage guessing* and significantly increased their correct responses—albeit not by enough to rule out random guessing. By contrast, we observed smaller decreases in DK responses among CCES and GS respondents than Luskin and Bullock reported, and no measurable increase in correct responses. Unlike our MTurk results, this pattern supports Luskin and Bullock's broader conclusion that the presence or absence of a DK option makes little difference.[12] Still, the muted, but slightly different, response pattern reveals the importance of context and the need to be cautious in how our claims generalize.

[12] More precisely, this pattern supports their conclusion that the presence or absence of a DK option makes little difference for the identification of respondents who could answer correctly by way of full knowledge or educated guesses, but also that the DK option does matter 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 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*. 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 – 368.

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 – 887.

Converse, Philip. 1964. "The Nature of Belief Systems in Mass Publics," in *Ideology and Discontent*, David E. Apter, ed.

Delli Carpini, Michael X. and Scott Keeter. 1996. *What Americans Know about Politics and Why It Matters*. 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 – 557.

Mondak, Jeffery J. 1999. "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 – 429.

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.

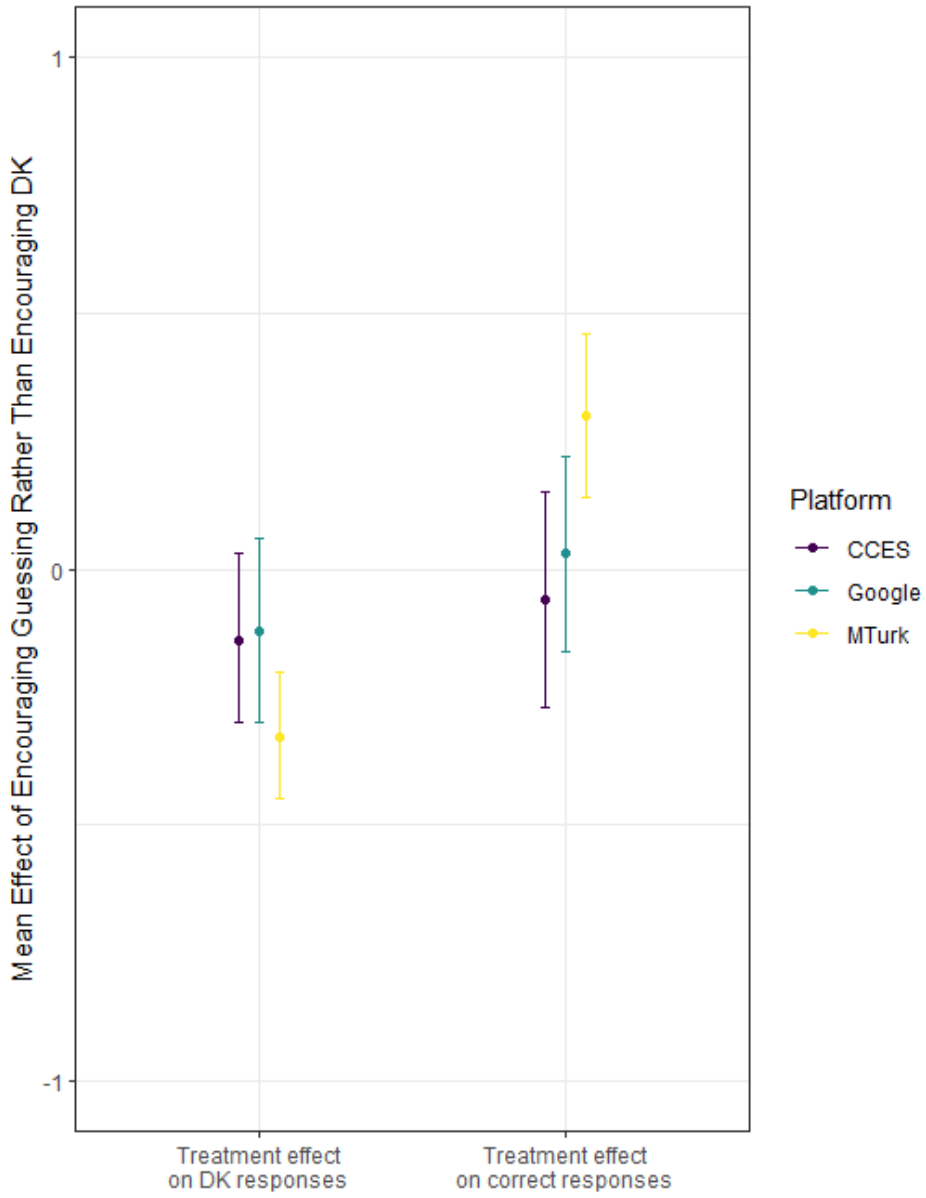
Tourangeau, Roger, Aaron Maitland, and H. Yanna Yan. 2016. "Assessing the Scientific Knowledge of the General Public: The Effects of Question Format and Encouraging or Discouraging Don't Know Responses," *Public Opinion Quarterly* 80:741-760.

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

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

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

Figure 1. Mean Treatment Effect of Encouraging Guessing on Correct and DK Responses



Supplemental Appendix

This supplement contains additional information to support the findings presented in the main manuscript. Table A1 shows the results of four probit models that predict respondents' assignment to the *encourage DK* condition (as opposed to the *encourage guessing* condition) using a series of demographic variables. The first three columns present separate models for CCES respondents, GS respondents, and MTurk respondents respectively. The final column presents a pooled model. Table A2 presents the same models, but applying sampling weights to the CCES and GS respondents. None of these models raises any concerns that respondents assigned to one condition differed meaningfully from those assigned to the other.

Table A3 presents ordinary least squares regressions of the treatment effects. Table A4 presents the same models, but applying sampling weights to the CCES and GS respondents. Tables A5 and A6 present the same models as Tables A3 and A4 (respectively), but using negative binomial regression rather than OLS. In all these tables, the first two models predict the number of items each respondent answered correctly, the latter two models predict the number of DK responses, and the second and fourth models include demographic controls. CCES respondents and GS respondents are indicated with separate variables; MTurk respondents are the omitted category. The *encourage guessing* variable indicates respondents assigned to that condition as opposed to the *encourage DK* condition. Because this dichotomous indicator is interacted with the CCES and GS variables, the baseline *encourage guessing* coefficient serves as an estimate of the treatment effect on MTurk respondents only. To obtain the treatment effect on CCES or GS respondents, one must add this baseline effect to the interaction estimates. Thus, in the first model of Table A3, we estimate that the *encourage guessing* condition is associated with a 0.30 increase

in the number of correct answers among MTurk respondents, but essentially zero change among CCES (0.30 – 0.36 = −0.06) and GS (0.30 – 0.27 = 0.03) respondents. For discussion of the results of these supplemental analyses, see footnote 6 in the main manuscript.

Figure A1 depicts the mean number of DK responses by condition and platform. Figure A2 does the same for the mean number of correct responses. The patterns identified through the more complicated analysis presented in the foregoing tables are plainly visible in these simple figures. Respondents from all three platforms (but especially MTurk) are somewhat less likely to select DK when encouraged to guess (Figure A1), but only MTurk respondents give significantly more correct responses as a result (Figure A2). Importantly, MTurk respondents do not stand out from CCES or GS respondents in the *encourage DK* condition, but they differ markedly in the *encourage guessing* condition. This pattern suggests that something bigger than mere sample selection effects is driving these results. MTurk users are not simply demographically different from CCES or GS users; their previous MTurk experience has changed the way they respond to our prompt.

Finally, Figure A3 replicates Figure 1 from the main manuscript, but with survey weights applied. The results are similar.

Table A1. Probit Models Predicting Assignment to *Encourage DK* Condition (Unweighted).

	CCES	GS	MTurk	Pooled
Male	-0.0074	-0.036	-0.017	-0.020
	(0.093)	(0.095)	(0.081)	(0.050)
Democrat	0.085	0.0069	-0.047	0.010
	(0.13)	(0.12)	(0.11)	(0.074)
Republican	-0.26	-0.11	0.042	-0.10
	(0.13)	(0.13)	(0.12)	(0.074)
4-year college degree	0.16	0.0015	0.066	0.090
	(0.097)	(0.098)	(0.081)	(0.051)
Age 25-34	0.16	-0.083	0.10	0.080
	(0.24)	(0.16)	(0.12)	(0.086)
Age 35-44	0.13	0.12	-0.12	-0.0028
	(0.24)	(0.17)	(0.13)	(0.095)
Age 45-54	0.15	0.31	-0.12	0.096
	(0.23)	(0.17)	(0.17)	(0.097)
Age 55-64	0.14	0.22	0.0073	0.079
	(0.23)	(0.17)	(0.19)	(0.096)
Age 65+	0.26	0.26	0.11	0.17
	(0.23)	(0.18)	(0.39)	(0.10)
Constant	-0.20	-0.11	-0.026	-0.089
	(0.23)	(0.16)	(0.14)	(0.093)
N	787	739	995	2,521

* $p \leq 0.05$ (two-tailed). Probit coefficients shown with standard errors in parentheses. Females, independents, and respondents aged 18-24 are the omitted categories. Rounding to two significant digits.

Table A2. Probit Models Predicting Assignment to *Encourage DK* Condition (Weighted).

	CCES	GS	MTurk	Pooled
Male	-0.091	0.038	-0.017	-0.029
	(0.12)	(0.11)	(0.081)	(0.058)
Democrat	0.11	0.055	-0.047	0.035
	(0.16)	(0.14)	(0.11)	(0.077)
Republican	-0.28	-0.078	0.042	-0.099
	(0.17)	(0.15)	(0.12)	(0.085)
4-year college degree	0.12	-0.044	0.066	0.052
	(0.12)	(0.11)	(0.081)	(0.059)
Age 25-34	0.27	-0.083	0.10	0.11
	(0.30)	(0.19)	(0.12)	(0.097)
Age 35-44	0.32	0.23	-0.12	0.080
	(0.29)	(0.20)	(0.13)	(0.11)
Age 45-54	0.18	0.39	-0.12	0.13
	(0.28)	(0.20)	(0.17)	(0.11)
Age 55-64	0.18	0.17	0.0073	0.087
	(0.27)	(0.20)	(0.19)	(0.11)
Age 65+	0.13	0.16	0.11	0.063
	(0.27)	(0.21)	(0.39)	(0.12)
Constant	-0.12	-0.15	-0.026	-0.077
	(0.28)	(0.19)	(0.14)	(0.11)
N	787	605	995	2,387

* $p \leq 0.05$ (two-tailed). Probit coefficients shown with standard errors in parentheses. CCES and GS respondents include sampling weights. Females, independents, and respondents aged 18-24 are the omitted categories. Rounding to two significant digits.

Table A3. Ordinary Least Squares Estimates of Treatment Effects (Unweighted).

	Correct	Correct	Don't Know	Don't Know
Encourage guessing	0.30*	0.31*	-0.33*	-0.33*
	(0.088)	(0.079)	(0.074)	(0.067)
CCES	-0.067	-0.18*	0.083	0.08
	(0.094)	(0.11)	(0.079)	(0.076)
× Encourage guessing	-0.36*	-0.35*	0.19†	0.19†
	(0.13)	(0.12)	(0.11)	(0.10)
Google Surveys	-0.098	-0.15†	0.12	-0.0047
	(0.093)	(0.090)	(0.079)	(0.076)
× Encourage guessing	-0.27*	-0.22†	0.21†	0.21*
	(0.13)	(0.12)	(0.11)	(0.10)
Male		0.70*		-0.36*
		(0.050)		(0.042)
Democrat		0.28*		-0.25*
		(0.067)		(0.056)
Republican		0.37*		-0.34*
		(0.073)		(0.062)
4-year college degree		0.60*		-0.34*
		(0.051)		(0.043)
Age 25-34		0.14		-0.17*
		(0.086)		(0.073)
Age 35-44		0.36*		-0.32*
		(0.095)		(0.080)
Age 45-54		0.57*		-0.35*
		(0.099)		(0.083)
Age 55-64		0.68*		-0.32*
		(0.099)		(0.084)
Age 65+		0.90*		-0.49*
		(0.11)		(0.092)
Constant	2.9	1.7	0.91	1.7
	(0.062)	(0.11)	(0.053)	(0.089)
N	2,613	2,521	2,613	2,521
R^2	0.01	0.18	0.02	0.09

* $p \leq 0.05$ (two-tailed), † $p \leq 0.05$ (one tailed). Ordinary least squares coefficients shown with standard errors in parentheses. MTurk is the omitted platform; females, independents, and respondents aged 18-24 are the omitted categories. Rounding to two significant digits.

Table A4. Ordinary Least Squares Estimates of Treatment Effects (Weighted).

	Correct	Correct	Don't Know	Don't Know
Encourage guessing	0.30*	0.31*	-0.33*	-0.33*
	(0.083)	(0.078)	(0.065)	(0.063)
CCES	-0.15	-0.17	0.21†	0.16
	(0.12)	(0.11)	(0.12)	(0.11)
× Encourage guessing	-0.20	-0.29*	-0.027	0.045
	(0.16)	(0.15)	(0.14)	(0.13)
Google Surveys	-0.14	-0.15	0.13	0.014
	(0.098)	(0.091)	(0.091)	(0.083)
× Encourage guessing	-0.29*	-0.28*	0.20	0.22*
	(0.14)	(0.12)	(0.12)	(0.11)
Male		0.66*		-0.32*
		(0.058)		(0.051)
Democrat		0.31*		-0.26*
		(0.082)		(0.075)
Republican		0.39*		-0.38*
		(0.087)		(0.078)
4-year college degree		0.67*		-0.36*
		(0.057)		(0.049)
Age 25-34		0.16		-0.15†
		(0.10)		(0.089)
Age 35-44		0.36*		-0.35*
		(0.11)		(0.092)
Age 45-54		0.60*		-0.35*
		(0.12)		(0.097)
Age 55-64		0.67*		-0.26*
		(0.13)		(0.12)
Age 65+		0.86*		-0.49*
		(0.13)		(0.11)
Constant	2.9	1.7	0.91	1.7
	(0.062)	(0.12)	(0.052)	(0.11)
N	2,456	2,387	2,456	2,387
R^2	0.01	0.17	0.02	0.10

* $p \leq 0.05$ (two-tailed), † $p \leq 0.05$ (one tailed). Ordinary least squares coefficients shown with standard errors in parentheses. MTurk is the omitted platform; females, independents, and respondents aged 18-24 are the omitted categories. CCES and GS respondents include sampling weights. Rounding to two significant digits.

Table A5. Negative Binomial Estimates of Treatment Effects (Unweighted).

	Correct	Correct	Don't Know	Don't Know
Encourage guessing	0.099*	0.099*	-0.44*	-0.44*
	(0.036)	(0.036)	(0.092)	(0.087)
CCES	-0.023	-0.065	0.087	0.087
	(0.040)	(0.043)	(0.089)	(0.090)
× Encourage guessing	-0.12*	-0.12*	0.30*	0.27*
	(0.055)	(0.056)	(0.13)	(0.13)
Google Surveys	-0.034	-0.058	0.12	-0.027
	(0.040)	(0.042)	(0.088)	(0.091)
× Encourage guessing	-0.088	-0.069	0.32*	0.29*
	(0.055)	(0.056)	(0.13)	(0.13)
Male		0.24*		-0.45*
		(0.024)		(0.0053)
Democrat		0.10*		-0.26*
		(0.032)		(0.066)
Republican		0.13*		-0.39*
		(0.035)		(0.075)
4-year college degree		0.20*		-0.44*
		(0.024)		(0.056)
Age 25-34		0.057		-0.18*
		(0.043)		(0.084)
Age 35-44		0.13*		-0.36*
		(0.046)		(0.095)
Age 45-54		0.20*		-0.40*
		(0.048)		(0.10)
Age 55-64		0.24*		-0.38*
		(0.048)		(0.098)
Age 65+		0.31*		-0.61*
		(0.051)		(0.11)
Constant	1.01	0.64	-0.36	-0.81
	(0.026)	(0.052)	(0.088)	(0.10)
N	2,613	2,521	2,613	2,521

* $p \leq 0.05$ (two-tailed), † $p \leq 0.05$ (one tailed). Negative binomial coefficients shown with standard errors in parentheses. MTurk is the omitted platform; females, independents, and respondents aged 18-24 are the omitted categories. Rounding to two significant digits.

Table A6. Negative Binomial Estimates of Treatment Effects (Weighted).

	Correct	Correct	Don't Know	Don't Know
Encourage guessing	0.098*	0.099*	-0.44*	-0.44*
	(0.027)	(0.026)	(0.087)	(0.086)
CCES	-0.052	-0.063	0.21†	0.15
	(0.044)	(0.040)	(0.11)	(0.11)
× Encourage guessing	-0.063*	-0.093†	0.065	0.13
	(0.057)	(0.051)	(0.15)	(0.15)
Google Surveys	-0.049	-0.059†	0.13	-0.0087
	(0.035)	(0.032)	(0.092)	(0.092)
× Encourage guessing	-0.096*	-0.089*	0.31*	0.32*
	(0.048)	(0.042)	(0.14)	(0.14)
Male		0.23*		-0.40*
		(0.020)		(0.061)
Democrat		0.11*		-0.27*
		(0.030)		(0.073)
Republican		0.14*		-0.43*
		(0.032)		(0.084)
4-year college degree		0.22*		-0.46*
		(0.020)		(0.063)
Age 25-34		0.064		-0.16
		(0.040)		(0.089)
Age 35-44		0.13*		-0.41*
		(0.042)		(0.10)
Age 45-54		0.22*		-0.39*
		(0.045)		(0.11)
Age 55-64		0.24*		-0.29*
		(0.048)		(0.12)
Age 65+		0.31*		-0.59*
		(0.048)		(0.13)
Constant	1.07	0.62	-0.095	-0.96
	(0.021)	(0.049)	(0.057)	(0.15)
N	2,456	2,387	2,456	2,387

* $p \leq 0.05$ (two-tailed), † $p \leq 0.05$ (one tailed). Negative binomial coefficients shown with standard errors in parentheses. MTurk is the omitted platform; females, independents, and respondents aged 18-24 are the omitted categories. CCES and GS respondents include sampling weights. Rounding to two significant digits.

Figure A1. Mean Number of DK Responses, by Condition and Platform (Unweighted)

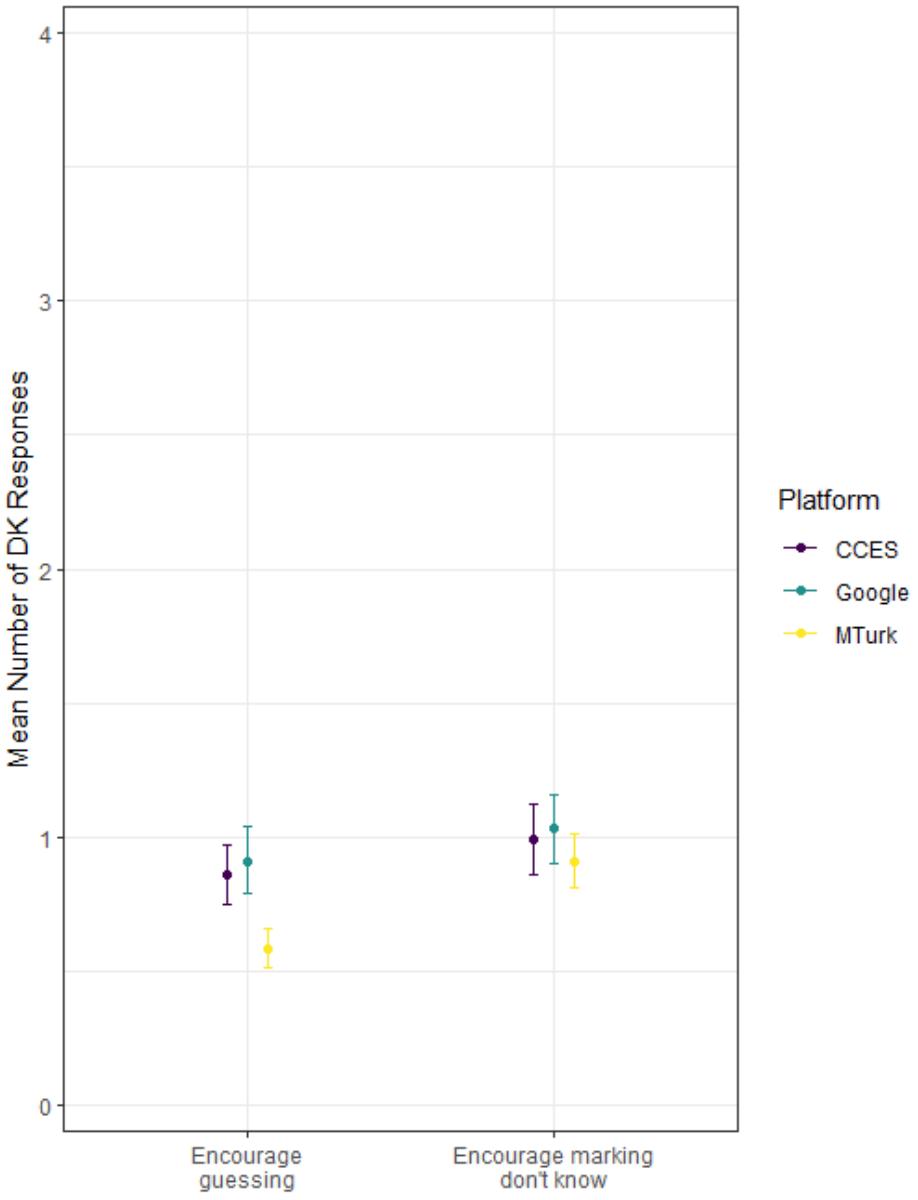


Figure A2. Mean Number of Correct Responses, by Condition and Platform (Unweighted)

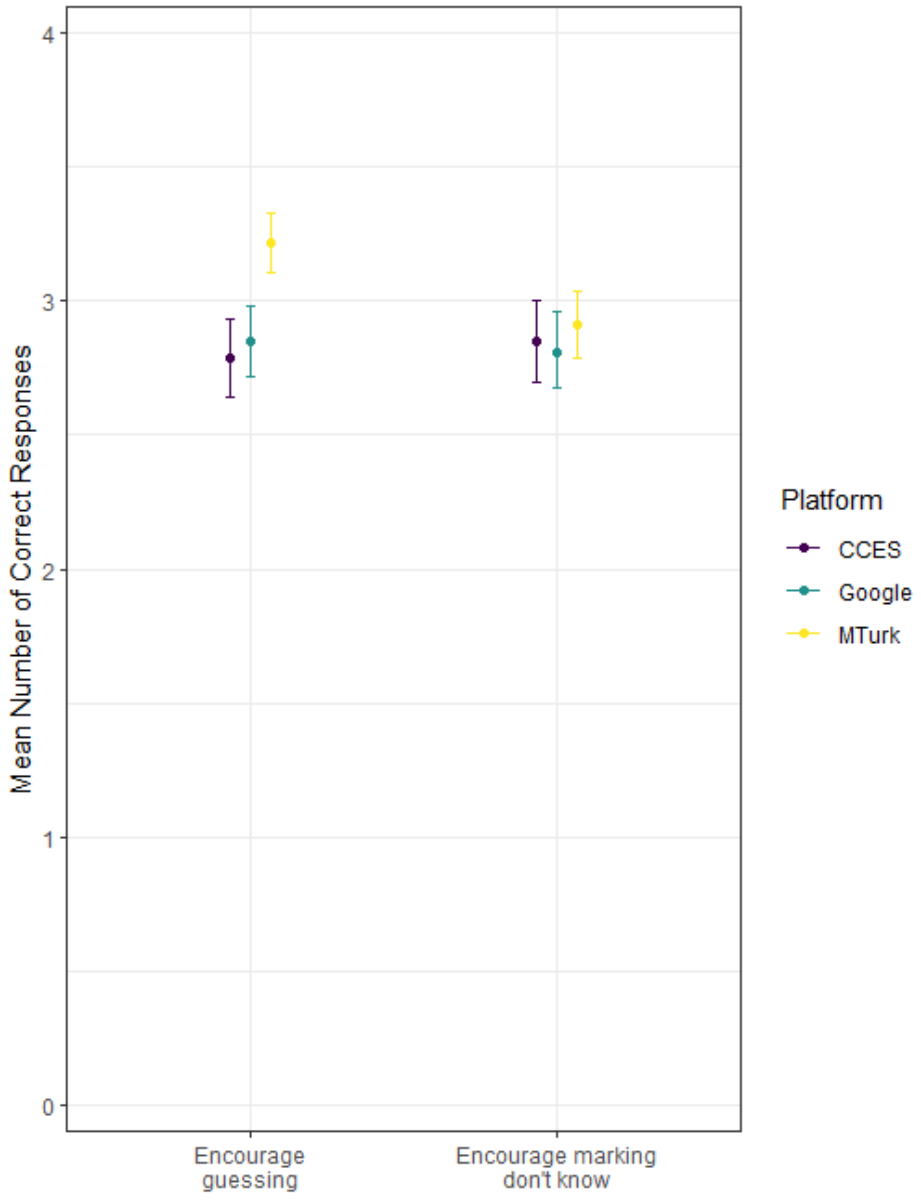


Figure A3. Mean Treatment Effect of Encouraging Guessing on Correct and DK Responses (Weighted)