






RESEARCH NOTE

## Who's cheating on your survey? A detection approach with digital trace data

Simon Munzert<sup>1\*</sup> , Sebastian Ramirez-Ruiz<sup>1</sup> , Pablo Barberá<sup>2</sup> , Andrew M. Guess<sup>3</sup>   
and JungHwan Yang<sup>4</sup> 

<sup>1</sup>Hertie School, Berlin, Germany, <sup>2</sup>Political Science and International Relations, University of Southern California, Los Angeles, USA, <sup>3</sup>Department of Politics, Princeton University, Princeton, USA and <sup>4</sup>University of Illinois at Urbana-Champaign, Urbana, USA

\*Corresponding author. Email: [munzert@hertie-school.org](mailto:munzert@hertie-school.org)

(Received 1 October 2021; revised 10 May 2022; accepted 18 May 2022; first published online 28 November 2022)

### Abstract

In this note, we provide direct evidence of cheating in online assessments of political knowledge. We combine survey responses with web tracking data of a German and a US online panel to assess whether people turn to external sources for answers. We observe item-level prevalence rates of cheating that range from 0 to 12 percent depending on question type and difficulty, and find that 23 percent of respondents engage in cheating at least once across waves. In the US panel, which employed a commitment pledge, we observe cheating behavior among less than 1 percent of respondents. We find robust respondent- and item-level characteristics associated with cheating. However, item-level instances of cheating are rare events; as such, they are difficult to predict and correct for without tracking data. Even so, our analyses comparing naive and cheating-corrected measures of political knowledge provide evidence that cheating does not substantially distort inferences.

**Keywords:** Measurement; survey design; measurement; digital trace data; political knowledge

The increasing popularity of self-administered online surveys in social science research has fueled concerns about data quality. One major consideration is that online surveys present a conducive environment for consultation of outside sources on the web. Without the constraints of a researcher-controlled environment, respondents can “google” the answers to factual knowledge questions. This potentially inflates estimates and distorts models of political knowledge.

In this note, we implement a direct yet unobtrusive approach to detect cheating behavior in the wild. By combining an online self-administered survey with passive tracking data of respondents' browsing histories, we can detect cheating with high precision at the item level. Our main goal is to assess the prevalence of cheating as precisely as possible and with high granularity, overcoming some of the limitations of previous cheating detection approaches. We then use this evidence to explore predictors as well as practical consequences of cheating.

In a quota-sampled German online panel, we find that 23 percent of respondents engage in cheating behavior, and we observe cheating rates ranging from 0 to 12 percent at the item level. Further evidence from a US online panel suggests that prompts that discourage cheating can be very effective. Results from multilevel models show that while there are robust person- and item-level determinants of cheating, the act of cheating per se is difficult to predict. As a consequence, in the absence of passive tracking data it is hard to correct for cheating behavior. At the same time, we find limited evidence that cheating distorts inferences from empirical models of political knowledge.

## 1 A passive tracking approach to detect cheating

Previous approaches to detect cheating in online surveys have provided mainly indirect evidence (e.g. Clifford and Jerit, 2016; Gummer and Kunz, 2019; Smith *et al.*, 2020; Höhne *et al.*, 2021; Style and Jerit, 2021, see online Appendix A for a detailed discussion of previous approaches and findings). We are the first to catch respondents engaging in this behavior *in flagrante*. Our approach relies on combining digital trace data and individual survey responses (Guess, 2021; Stier *et al.*, 2022). On the one hand, we have surveys with start and end timestamps and answers to the knowledge items, in addition to a set of political and demographic covariates. On the other, we obtain respondents' web navigation data collected through passive metering.

The granularity of our data allows us to pinpoint browsing behavior during the survey-taking interval. We can detect whether, how, and on which items respondents engage with outside sources. To identify cheating, we began by flagging suspicious navigation during the survey intervals through a set of keyword queries. Subsequently, we manually validated the flagged entries. In a third step, we screened all the parallel navigation during the survey intervals to the domains resulting from the keyword-validated cheating instances (see Figure C1 in the online Appendix).<sup>1</sup>

The evidence we use allows us to relax some of the assumptions that come with the use of previously employed approaches to infer cheating, such as self-reports, logs of window switching, or response times. Furthermore, the granularity of our data enables us to explore dynamics within survey sessions, such as variation between question types.

## 2 Data and results

To identify cheating in the wild, we use data from an online panel survey recruited from the German YouGov Pulse panel. The panel enables passive metering of individuals' web usage on their registered laptop, desktop, and mobile devices. This allows us to observe every URL that respondents visited on these devices during the study period. Respondents were quota-sampled based on age, gender, and education to match target marginals of the German population that use the internet using the Best for Planning study (Best for Planning, 2017) as a benchmark. The present study was launched on 13 July 2017, and included five waves that were completed by 4 October 2017. Waves 2–5 contained measures of political knowledge. We exclude respondent observations for which the survey-taking process does not appear in the tracking data. This reduces our sample to  $n_{2,3,4,5} = \{545, 557, 553, 519\}$  respondents per wave and  $n_{\text{tot}} = 685$  unique respondents in total.<sup>2</sup> More information on the survey setup, placement of questions, recruitment of survey respondents, and deployment of the passive metering software as well as privacy and ethical considerations are reported in online Appendix B.

Across survey waves, we asked eight different knowledge questions with a total of 68 item options (knowledge indicators). We differentiate between questions and item options because both knowledge and cheating behavior is measured at the item level. For instance, the visual elites knowledge question asks respondents to match several pictures of high-profile politicians with their party in a grid. Each politician–party pair represents one item. Section F in the online Appendix gives an overview of the items included in each of the waves. In total, four different question types were implemented varying in content, format, difficulty, and robustness against cheating. The question types are:

<sup>1</sup> A full list of keywords used to gather suspicious browsing, as well as tables presenting all instances of validated cheating, can be found in online Appendix E.

<sup>2</sup> If users who want to cheat were more likely to turn off web tracking, this would call our measurement strategy into question. While we cannot measure cheating propensity for non-tracked respondents, we model the availability of Pulse data using a set of respondent-level predictors and find no substantive differences between the subjects for whom the navigation data were available and those who are dropped (see also Figure C3 in the online Appendix).

- Two *factual knowledge* questions on voting procedures in the German federal electoral system. The questions were asked with a closed-ended format (3 or 5 response options) and are often used in German election surveys.
- One *visual elites knowledge* question (Prior, 2014) asking respondents to match nine pictures of high-profile politicians with their party in a grid.
- One *verbal elites knowledge* question (open-ended) asking respondents to name one or several leading candidates for each of the six main parties in the 2017 election.
- Four *event knowledge* questions asking respondents to select from a list of ten events that they believe happened in the past few weeks using a multiple-choice format. The event list was designed to cover political and non-political events, with half of the events being real.

Furthermore, we also analyze data from a companion survey fielded in the United States between 23 April 2018, and 5 February 2019, to 1494 respondents recruited from the US YouGov Pulse panel. Due to differences in the passive metering software used, we observe only domain-level data and not the full URLs for many respondents, which reduces the sample to  $n = 409$  respondents for whom cheating can be reliably detected. The panel included a total of 35 knowledge items spread across five waves. The US study also differed in that it included commitment questions (Clifford and Jerit, 2016) and a self-report measure of cheating.

### **2.1 How prevalent is cheating in the wild?**

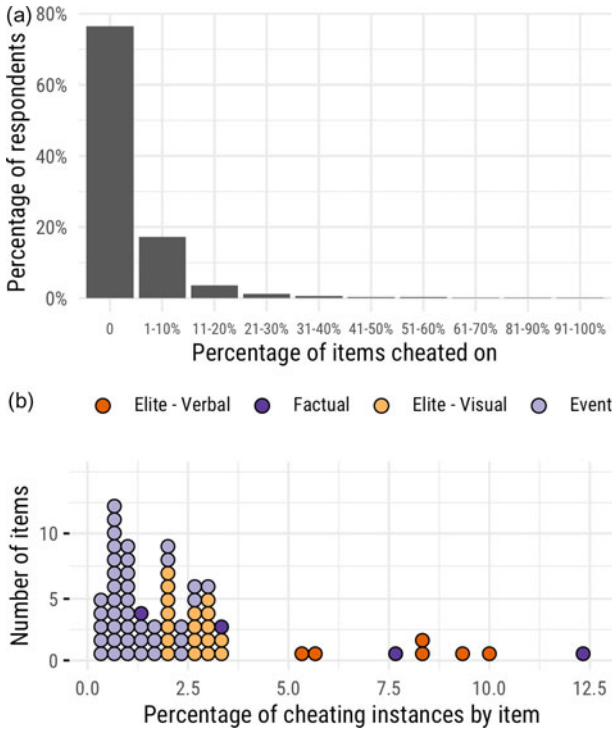
Studying cheating at the respondent level, we find that 23 percent of participants engaged with outside sources to answer at least one of the knowledge items (see Figure 1, panel (a)). The distribution of cheating instances is right-skewed, suggesting that serial cheaters are rare. Analyzing the tracking data beyond cheating activity, we find that 66 percent of respondents engaged with other websites while completing the survey in at least one of the waves. Though this suggests page switching is common while taking surveys, we could only validate 35 percent of the respondents that navigated in parallel to cheat. The rest of the traffic during the survey-taking interval went largely to search engines, video streaming platforms, social media, email, and online-shopping (see Figure C9 in the online Appendix). This suggests that approaches leveraging screen-switching as a proxy measure (e.g. Gummer and Kunz, 2019; Höhne *et al.*, 2021) may suffer from a non-negligible rate of false positives.

Turning to the question and item levels, panel (b) of Figure 1 shows that the prevalence of cheating instances varies considerably across question types. Information seeking from outside sources ranges from, on average, 1 percent in event items to 8 percent in open-ended elite-verbal items. The most cheated-on item asked which body of the government was tasked with electing the chancellor in Germany (factual knowledge). The heterogeneity of identified cheating at the question-type level illustrates possible limitations of relying on catch questions to explore the validity of political knowledge batteries as a unit. Further examination of the prevalence of cheating at the item level can be found in Figure C7 in the online Appendix.

### **2.2 Do anti-cheating commitment pledges help?**

In the US survey, we presented panelists with one visual-elites knowledge, one factual knowledge, and three event knowledge questions, totaling 35 knowledge items. Additionally, this setup contained an anti-cheating commitment pledge just before the knowledge questions were asked and a lookup pledge at the end of the survey.<sup>3</sup>

<sup>3</sup>The wording of the pledge was as follows: “When reading these next questions, please do not consult outside sources or other people for the answers. We are interested in what you believe. If you are unsure, please just take your best guess. Will



**Fig. 1.** Distribution of instances of cheating at the respondent and item level. (a) Respondent-level cheating distribution. (b) Item-level cheating distribution. *Note:* Number of respondents: 666; number of items: 68. Panel (a) is based on instances of cheating identified in the tracking data at the respondent level, showing that for 77 percent of the respondents we find no evidence of cheating in the tracking data. Panel (b) is based on instances of cheating identified in the tracking data at the item level, showing systematic variation of propensities to cheat by item type.

In contrast to the previous results, US tracking data suggest that the prevalence of cheating, in this setting, was minimal. A total of four respondents engaged with outside sources in six instances, all of them concerning event items. That is to say, fewer than 1 percent of respondents cheated and 0.01 percent of items were cheated on. Furthermore, the frequency of parallel navigation for this set of respondents was lower than that of the German survey, with 35 of the respondents engaging with other websites while completing the survey. Concerning the self-report item, none of the panelists who were caught red-handed reported having engaged with outside resources to respond to the survey. Overall, eight respondents stated having used outside information. We found through manual validation that two of them did use search engines to look up information, but these instances concerned questions outside of the knowledge battery. The remaining six did not have any suspicious activity recorded in the tracking data, though we cannot discard the possibility of cheating outside of the logged devices. Further insights can be obtained from the broader US sample of  $n_{domain} = 963$  at the domain level, which includes all respondents with domain-level or full URL data. An additional five respondents, for a total of 13 (1.4 percent of respondents in this sample), reported having utilized outside sources to complete the survey. Finally, parallel navigation from the broader sample was 61 percent.

Since the anti-cheating commitment pledge was not randomly assigned, we cannot identify its effect on cheating. Other factors that could explain the large difference in cheating prevalence between the two samples include differences in person or item characteristics that predict cheating (see next section). That being said, our evidence is consistent with existing evidence on the effect of commitment items (Clifford and Jerit, 2016; Smith *et al.*, 2020), although the

you answer the following questions without help from outside sources?" The wording of the lookup pledge was as follows: "It is essential for the validity of this study that we know whether participants looked up any information online during the study. Did you make an effort to look up information during the study? Please be honest; you will still collect incentives and you will not be penalized in any way if you did."

effectiveness reported in previous experimental studies was substantially lower and relied on different sample types (see Table A1 in the online Appendix).

### 2.3 What predicts cheating?

Despite its limited prevalence, is cheating predictable? If respondent or item characteristics that are commonly collected in social science surveys have predictive value, they could help correct for cheating even in the absence of tracking data. Furthermore, this evidence could be used to inform item usage and design as well as substantive models of political knowledge.

To assess the predictability of cheating, we model item-level cheating using a Bayesian logistic mixed-effects model with person- and item-level random effects and a set of person- and item-level fixed effects. At the person level, we consider gender, age, level of school education, internal efficacy, political knowledge, and survey-taking patterns as predictors. Internal efficacy is measured by conducting a principal components analysis on five (both internal and external) efficacy questions and taking the first principal component.<sup>4</sup> Political knowledge is measured as the fraction of knowledge items answered correctly after adjusting for cheating by coding validated instances of cheating as incorrectly answered. Further, we derive a binary measure of habitual survey-taking behavior based on the average weekly time a respondent spent on the online survey platforms that fell within the top 100 domains in our data (more than 2 h per week).<sup>5</sup> At the item level, we consider item type (Gummer and Kunz, 2019) and difficulty (Motta *et al.*, 2017; Smith *et al.*, 2020; Style and Jerit, 2021). Difficulty is measured as the fraction of correct answers to an item after adjusting for cheating.

Figure 2 reports the results of our main model.<sup>6</sup> Panel (a) presents fixed-effects estimates, panel (b) reports marginal effects (predicted probabilities) by predictor. In line with previous research, we find robust evidence for a positive relationship between education and cheating. Gender and age are also associated with cheating. We find an interesting difference between perceived and actual competence: Cheating is more likely among those with high levels of internal efficacy, but less likely among those who actually know more about politics. We find no relationship between habitual survey-taking and cheating. At the item level, cheating is most likely to happen on the open-ended (verbal) elite questions, followed by factual (closed) questions about the political process and harder-to-cheat elite questions with visual cues. We find that the least cheating happens on event knowledge items. Item difficulty does not seem to matter.<sup>7</sup>

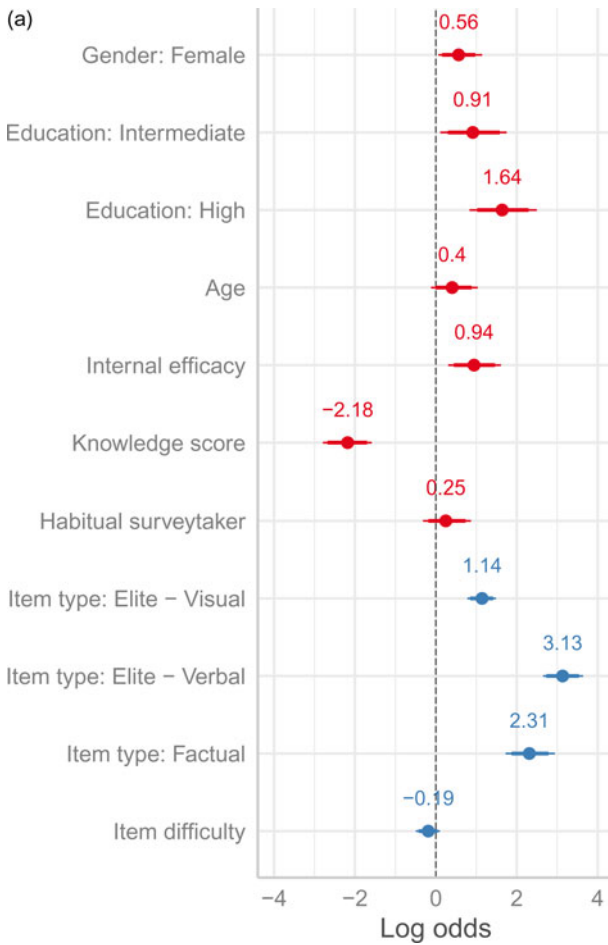
We also explored whether response times could be used as a proxy for cheating. Response times, as a frequently available alternative form of metadata (Read *et al.*, 2021), have been used to infer cheating behavior (Clifford and Jerit, 2016; Munzert and Selb, 2017; Gummer and Kunz, 2019; Marquis, 2021), building on the assumption that respondents who look up information elsewhere should, on average, take longer to submit an answer. One limitation of this approach is that it only allows us to infer at the question, rather than the item level, because often many items are presented on one page and response times can only be measured at the page level. Nevertheless, in our data, we do find a strong and very robust relationship between log time spent on a question and whether cheating occurred, controlling for a set of commonly available respondent and question characteristics (see Figure C5 in the online Appendix for the full

<sup>4</sup>The items contributing most to the internal efficacy score are (1) “Political issues are often difficult for me to understand”, (2) “I am very well versed in politics in general”, and (3) “I am very well informed about current political events”.

<sup>5</sup>This metric is based on the average time spent per week on rewards and survey-taking platforms in the top 100 domains in the tracking data. Figure C10 in the online Appendix provides the distribution of average time spent on these platforms per week for the respondents with tracking data.

<sup>6</sup>We estimate the model via MCMC using the No-U-Turn-Sampler (Hoffman and Gelman, 2014), employing default prior choices as implemented in *rstanarm* (Goodrich *et al.*, 2020).

<sup>7</sup>Respondent-level models reveal very similar relationships between predictors and cheating (see Figure C4 in the online Appendix).



**Fig. 2.** Estimated effects of respondent and item characteristics on response-level cheating incidence. (a) Fixed-effects estimates. (b) Predicted probabilities. *Note:* Results from a Bayesian logistic mixed-effects model with person and item random effects. Posterior means along with 80 and 95 percent credible intervals reported. Number of observations: 35,486; number of respondents: 656; number of items: 68. To compute the predicted probabilities, numeric covariates are held at their means and the other covariates are set to: female, intermediate education, item type “Elite - Verbal”, and habitual survey-taker.

model). Median response times for questions that were cheated on are over a minute longer than for honestly answered questions (see Figure C6 in the online Appendix). However, despite these stark differences, our predictive model that combines time with respondent and question-level information has an F1 score of just 60 percent (precision = 60 percent and recall = 60 percent), far from giving an accurate classification of cheating instances. We conclude that relying on response times alone to approximate cheating potentially generates many false negatives and potentially even more false positives.

**2.4 Does cheating distort models of political knowledge?**

If cheating happens, does it matter for our understanding of political knowledge? In a further step, we explore whether cheating distorts models of political knowledge. First, we compare estimates for predictors of political knowledge using unadjusted and cheating-adjusted measures. In line with previous research, we study gender, age, political interest, internal efficacy, and education as determinants of political knowledge (Delli Carpini and Keeter, 1996; Prior, 2014). We then run a multiverse of linear models using all possible 31 permutations of the covariates to predict question-specific as well as composite measures of political knowledge. For pairs of otherwise identically specified models that only differ in the outcome measure (naive or cheating-adjusted political knowledge), we then test for the equality of coefficients of each covariate. Figure C8 in

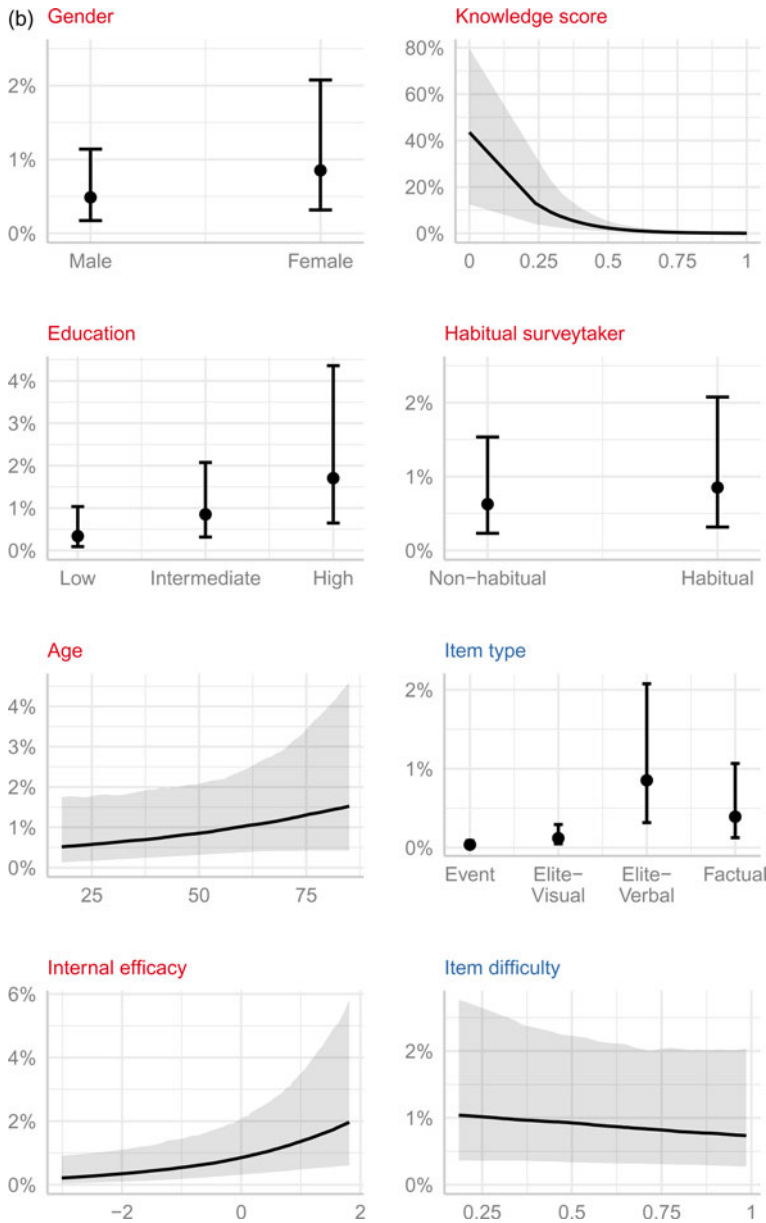


Fig. 2. Continued.

the online Appendix reports the distribution of Z-scores of the difference-in-coefficients tests. We do not find substantive differences in the estimates for any of the covariates in any of the models.

Next, we explore how the effect of political knowledge differs between cheaters and non-cheaters in predicting various related outcomes, including political interest, vote certainty, internal efficacy, and likelihood to vote. To that end, we regress these outcomes on our different measures of political knowledge (uncorrected for cheating) and interact the knowledge indicator with the cheating indicator. This gives us an estimate of how the estimated effect of political knowledge differs between both groups, while also adjusting for baseline differences between cheaters and non-cheaters. We also adjust for gender, education, and age in these models.

Tables C1 to C4 in the online Appendix report the results. In most cases the differences are substantively minor, and there is no instance of a statistically significant difference. Both analyses suggest that the distortions in measures of political knowledge induced by cheating have no severe downstream consequences for either the estimated relationships between political knowledge and some of its key predictors or estimates of political knowledge as a predictor of related outcomes.

### 3 Discussion and conclusion

In this note, we leverage a novel data collection format to explore the prevalence of cheating in online surveys of political knowledge. By inspecting the parallel web navigation of survey respondents, we uncover first-hand evidence of cheating: We find that 23 percent of respondents in our German survey panel engage at least once with outside sites for answers, though serial cheaters are rare. Our model of item-level cheating reveals that subjects who report being knowledgeable about politics, have higher levels of formal education, and have lower actual knowledge scores are more likely to cheat. Cheating is also particularly likely for open-ended items that ask for verbal input. Despite these patterns, due to the low base rate of verified cheating instances, it remains difficult to reliably classify.

It is important to note that though our cheating detection approach provides a precise method for retrieving first-hand evidence of cheating, it is not without limitations. First, lacking information about offline behavior and thought processes, we have to assume that our *corpus delicti* in the shape of trace data represents instances of cheating and not, e.g. of respondents checking their response after answering. Also, we cannot identify whether respondents consult devices not registered for passive metering—or even offline sources—to search for information. Additionally, not all survey-takers show up in the tracking data, so our sample is subset to respondents whose survey logs we can match. Nevertheless, we overcome some of the limitations of previous research utilizing self-reports, catch questions, and window switching logs by parsing the web navigation of survey respondents observed during the survey-taking interval.

Even though cheating occurs with some regularity in survey samples, optimism may be warranted: Our findings from the US raise the possibility that commitment pledges to refrain from looking up information (Clifford and Jerit, 2016) could potentially reduce the rates of cheating. Further, we fail to observe any significant differences in tests of coefficient correspondence on determinants of political knowledge for naive and cheating-adjusted measures.

Our data framework provides us with a rare opportunity to observe the prevalence of cheating with a high level of detail. Our empirical strategy offers an effective strategy to correct for cheating post hoc in survey samples with linked digital trace data. Nevertheless, this data collection setup is unique and not available to most researchers in the social sciences. Given these constraints, we recommend that researchers take measures against cheating before it occurs. Our findings suggest that normative commitment pledges discouraging the use of external sources and question types that are more robust to cheating, such as visual knowledge questions, are two viable options. Finally, while researchers face tradeoffs between self- and interviewer-delivered survey modes, the validity of political knowledge measures does not need to be one of them, since it can to a large extent be addressed through survey design features.

**Supplementary material.** The supplementary material for this article can be found at <https://doi.org/10.1017/psrm.2022.42>. To obtain replication material for this article, please visit <https://doi.org/10.7910/DVN/WGWUK2>.

**Acknowledgments.** This research was generously funded by a grant from the Volkswagen Foundation Computational Social Science Initiative [92 143 to S.M., P.B., A.G., and J.Y.]. The study was approved by the Princeton University Institutional Review Board (#8327, 10014, 10041).

### References

- Best for Planning** (2017) *Berichtsband b4p 2017*. Gesellschaft für integrierte Kommunikationsforschung.
- Clifford S and Jerit J** (2016) Cheating on political knowledge questions in online surveys: an assessment of the problem and solutions. *Public Opinion Quarterly* **80**, 858–887.



- Delli Carpini MX and Keeter S** (1996) *What Americans Know About Politics and Why it Matters*. New Haven: Yale University Press.
- Goodrich B, Gabry J, Ali I and Brilleman S** (2020) RSTANARM: Bayesian applied regression modeling via Stan. R package version 2.21.1. <https://mc-stan.org/rstanarm>.
- Guess AM** (2021) (Almost) Everything in moderation: new evidence on Americans' online media diets. *American Journal of Political Science* **65**, 1007–1022.
- Gummer T and Kunz T** (2019) Relying on external information sources when answering knowledge questions in web surveys. *Sociological Methods & Research* **51**, 816–836.
- Hoffman MD and Gelman A** (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623.
- Höhne JK, Cornesse C, Schlosser S, Couper MP and Blom AG** (2021) Looking up answers to political knowledge questions in web surveys. *Public Opinion Quarterly* **84**, 986–999.
- Marquis L** (2021) Using response times to test the reliability of political knowledge items in the 2015 Swiss post-election survey. *Survey Research Methods* **15**, 79–100.
- Motta MP, Callaghan TH and Smith B** (2017) Looking for answers: identifying search behavior and improving knowledge-based data quality in online surveys. *International Journal of Public Opinion Research* **29**, 575–603.
- Munzert S and Selb P** (2017) Measuring political knowledge in web-based surveys: an experimental validation of visual versus verbal instruments. *Social Science Computer Review* **35**, 167–183.
- Prior M** (2014) Visual political knowledge: a different road to competence? *The Journal of Politics* **76**, 41–57.
- Read B, Wolters L and Berinsky AJ** (2021) Racing the clock: using response time as a proxy for attentiveness on self-administered surveys. *Political Analysis* 1–20. <https://doi.org/10.1017/pan.2021.32>.
- Smith B, Clifford S and Jerit J** (2020) How internet search undermines the validity of political knowledge measures. *Political Research Quarterly* **73**, 141–155.
- Stier S, Mangold F, Scharkow M and Breuer J** (2022) Post post-broadcast democracy? News exposure in the age of online intermediaries. *American Political Science Review* **116**, 768–774.
- Style H and Jerit J** (2021) Does it matter if respondents look up answers to political knowledge questions? *Public Opinion Quarterly* **84**, 760–775.