Detecting and Deterring Information Search in Online Surveys

Matthew H. Graham*

9,732 words

April 24, 2022

This paper introduces a framework for measuring information search in online surveys and evaluating methods of combatting it. The results indicate that information search is a serious but manageable problem. The frequency of search is highly variable, ranging from 2 to 30 percent on a battery of general political knowledge questions. Deterrence works: a pledge not to cheat reduces search by half. Detection also works: paradata generated by the respondent's web browser identifies 70-85 percent of search, while 60-85 percent of search on knowledge questions is undertaken by respondents who correctly answer "catch" questions about obscure Supreme Court cases. Detection and deterrence are complements: deterrence reduces search *ex ante*, while detection quantifies success and provides options for dealing with undeterred search *ex post*. In combination, the three methods (pledge, paradata, catch) deter or detect more than 90 percent of search, leaving search to affect about 0.5 percent of the remaining observations.

*Institute for Data, Democracy, & Politics, George Washington University. mattgraham@gwu.edu.

The shift to online surveys has made it easier for respondents to look up the answers to questions designed to test factual knowledge (Clifford and Jerit 2014; Liu and Wang 2014; Strabac and Aalberg 2011; Shulman and Boster 2014). Researchers have developed several methods for addressing the information search problem (hereafter, "search"). Likely searchers can be detected using self-reports (Jensen and Thomsen 2014), "catch" questions that would be difficult to answer correctly without looking them up (Bullock et al. 2015; Motta et al. 2016), or paradata methods that observe the respondent's engagement with the survey (Diedenhofen and Musch 2017). Search can be deterred using requests (Vezzoni and Ladini 2017), pledges or commitment devices (Clifford and Jerit 2016), and admonitions to respondents who are caught in the act (Diedenhofen and Musch 2017).

This paper introduces a framework for accounting for measurement error in estimates of the prevalence of information search. This opens the door for newly specific evaluation of methods of combatting it, including how well they perform, why they fall short, and how performance changes when methods are layered on top of one another. The framework is applied to one deterrence method, a pledge not to cheat, and two detection methods, catch questions and paradata detection. Three empirical studies (total $N > 14{,}000$) yield the following key findings:

1. **Information search is common in political knowlege surveys.** At baseline, search was estimated to be present in 7.8 percent of answers to widely-used political knowledge questions in Study 1, 17.6 percent in Study 2, and 11.6 percent in Study 3. These estimates are adjusted for false positives and false negatives using a bias correction.

2. **Question content affects search.** Search frequency varies considerably between questions. On knowledge questions, search ranged from 2.5 to 11.5 percent in Study 1, 8.7 to 29.8 percent in Study 2, and 1.9 to 21.3 percent in Study 3.

3. **Response scales affect the prevalence of search.** In split ballot experiments, search was 30 to 100 percent more common on open-ended questions than on multiple choice questions. Despite this, *fewer* respondents assigned to open-ended questions answered correctly.

4. **Deterrence works.** A randomly assigned pledge not to look up the answers reduced search by about 50 percent in each study.

5. **Detection works.** The paradata method detected between 70 and 85 percent of search in all three studies. The catch method detected 60 to 85 percent.

6. **Detection and deterrence are complements.** A combination of detection and deterrence can eliminate more search from the data, at a lower cost, than detection alone can achieve. This is because deterrence eliminates search *ex ante* without affecting the detection methods' ability to detect what remains.

7. **Paradata detect search more efficiently than catch questions.** The catch method produces many more false positives than the paradata method. In each study, about 90 percent of those flagged in paradata looked up the answer to the knowledge questions, compared with 30 to 45 percent of those who answered a catch question correctly.

8. **Catch questions are an unreliable proxy for the prevalence of search on knowledge questions.** In all three studies, catch questions see more than twice as much search as the average knowledge question. Yet when used to approximate the proportion of subjects who search at least once, catch questions underestimate in every case.

9. **Multi-method approaches can all but eliminate search.** In each study, combining the pledge, paradata, and catch methods reduced search to 0.5 percent or less of the unflagged observations. This falls as low as 0.1 percent for questions with the lowest base rates of search and rises to 0.9 percent on the questions with the highest base rates. The catch method adds the least marginal value. With the pledge and paradata alone, search can be reduced to 0.7 percent of the unflagged observations.

A few practical takeaways can be distilled from these findings. First, search is a manageable problem, especially for questions with low base rates of search. Second, detection and deterrence have complementary strengths and weaknesses. Deterrence is especially valuable because it eliminates search *ex ante*. But without detection, researchers are forced to tolerate an unknown amount of search. Detection assesses the extent of the problem and gives researchers options for dealing with search that was not successfully deterred. Third, question content and response scales affect the prevalence of search. Because the costs and benefits of detection methods vary with the base rate of search, a combination of strategies that is satisfactory in one case may be insufficient in another. This calls for question-specific tests in the wide range of domains in which search is a threat. Finally, paradata methods are preferable to catch questions. Though both methods catch a similar proportion of those who search (sensitivity), the paradata identify searchers with much greater precision

(specificity). This makes the researcher's options for dealing with information search more attractive: fewer observations to drop, better options for dealing with missing data. As a diagnostic tool, catch questions also fall short. They overestimate the amount of search on the average knowledge question, underestimate the proportion of subjects who search on at least one question, and provide no information on how search varies between questions.

These results have broad implications. Among the 100 most recently published articles that contain a measure of political knowledge, less than one-third examine the United States (Appendix A.6, page A16). In addition to political science and public opinion research, the fields of communication, psychology, and sociology are prominently represented among this set of articles.[1] Looking beyond traditional political knowledge scales, information search looms as a threat for any survey question with a correct answer that can easily be looked up. These include other domain-specific knowledge scales, e.g. those used to measure the public's knowledge of history (Starratt et al. 2017; Pevnaya et al. 2019), science literacy (Cooper and Farid 2016; Drummond and Fischhoff 2017), legislators' policy positions (Ansolabehere and Jones 2010), and knowledge of issues of importance to one's social groups (Dolan 2011; Cohen and Luttig 2020); factual questions used to measure perceptions and misperceptions, e.g. government policy (Gilens 2001; Pasek et al. 2015), the economy (Bartels 2002), and matters of public health (Lunz Trujillo et al. 2020); and measures of personal traits like cognitive reflection (Pennycook and Rand 2019), numeracy (Schwartz et al. 1997; Peters et al. 2006), and outgroup bias (Ahler and Sood 2018). In all of these domains, the approach developed below can diagnose the extent of the problem and evaluate solutions.

The following sections review existing approaches and describe the methodology. The results are then presented in four parts: detecting the baseline prevalence of information search, evaluating methods of deterring it, evaluating the benefits and costs of combining multiple methods, and heterogeneous effects.

[1]Within political science, political knowledge measures are also used in a wide range of subfields. The phrase "political knowledge" appears in 5 to 10 percent of recently published articles in several high-impact comparative politics journals, including *Comparative Political Studies*, the *European Journal of Political Research*, and *West European Politics*. See Appendix A.6, page A16.

The Problem of Information Search

Relative to traditional interviewer- or lab-administered surveys, online surveys make information search easier for respondents to undertake and harder for researchers to prevent. Most immediately, this threatens to inflate the general public's apparent knowledge. Existing estimates often suggest that search is alarmingly prevalent; published estimate based on catch questions and self-report measures range from 15 to 25 percent (Bryson 2020; Jensen and Thomsen 2014; Motta et al. 2016; Style and Jerit 2021). Published estimates based on detection methods are more varied, ranging from 3 to 30 percent (Diedenhofen and Musch 2017; Gummer and Kunz 2019; Höhne et al. 2021). For the sort of general political knowledge scale studied here, search threatens to exaggerate the public's civic competence, political awareness, and political sophistication. Similarly, information search could inflate the public's apparent knowledge of other domains (e.g., science and history), distort widely-used measures of personal traits (e.g., numeracy and cognitive reflection), and conceal misperceptions and misinformed beliefs (which are often measured using factual questions). Simply put, information search threatens any survey measure with answers that can easily be looked up.

Information search also threatens to distort relationships between variables. Existing research finds that search affects the observed relationship between political knowledge and age, interest in politics, ideological constraint, political participation, and urbanicity (Bryson 2020; Clifford and Jerit 2014; Marquis 2021; Smith et al. 2020; Style and Jerit 2021). Adding to these concerns, this paper finds that search is more common among men, young people, strong partisans, those who are more interested in politics, those who are more cognitively reflective, and those who endorse conspiratorial beliefs at high rates. These relationships mean that the tendency for any group to appear more knowledgeable than another, or less prone to misperception than another, could be an artifact of information search. In any domain, understanding how knowledge relates to other factors requires one to grapple with the problem of information search.

The combination of a menacing threat and a lack of assurance in most cases that it has been successfully addressed breeds skepticism about the whole enterprise. For example, Rapeli (2022) argues that information search has rendered it "impossible to reliably ask knowledge questions" in unsupervised survey environments (97). This study's findings suggests that such pessimistic conclusions are premature. With a combination of currently available detection and deterrence tools, researchers can conduct analyisis or robustness checks using data that are proven to be minimally affected by search.

Approaches to Countering Information Search

Researchers use two classes of methods to deal with information search: detection and deterrence. Detection methods seek to identify respondents who look up the answers. These include self-reported admissions (Jensen and Thomsen 2014), "catch" questions that should rarely be answered correctly by chance (Berinsky et al. 2012; Motta et al. 2016), examining paradata generated by the subject's web browser (Diedenhofen and Musch 2017; Gummer and Kunz 2019; Höhne et al. 2021), and collecting the subject's browsing history (Gooch and Vavreck 2019). Deterrence methods seek to dissuade respondents from searching in the first place. These include requests (Motta et al. 2016), pledges (Clifford and Jerit 2016), timers (Domnich et al. 2015), and in-survey admonitions to respondents who are identified as likely searchers (Diedenhofen and Musch 2017).

Detection and deterrence have complementary pros and cons. The key advantage of deterrence is that it eliminates search *ex ante*, avoiding the tradeoffs posed by *ex post* methods of dealing with suspected searchers (e.g., dropping observations or imputing missing data). The more search can be prevented, the less must be tolerated or dealt with. The chief disadvantages of deterrence are that (1) not all search is deterred and (2) on its own, deterrence provides little sense of the scope of the problem. Detection complements deterrence by (2) diagnosing the extent of the problem and (1) providing researchers with *ex post* options for dealing with search that they were unable to deter. This paper tests one deterrence method,

a pledge not to look up the answers (Clifford and Jerit 2016).

The advantages and disadvantages of detection methods are aptly captured by terminology from classification problems in medicine (James et al. 2021, 145-49). A detection method may either yield false negatives by failing to detect some who search or yield false positives by falsely accusing some who do not. Detection methods that detect all search are highly *sensitive* ($P(\text{flag}|\text{search}) \approx 1$), while methods that do not incorrectly flag respondents are highly *specific* ($P(\text{flag}|\neg\text{search}) \approx 0$). For example, consider self-report measures, which flag respondents as suspected searchers if they admit to having searched. Self-reports are likely to be highly specific, meaning that few who did not search will claim to have search ($P(\text{flag}|\neg\text{search}) \approx 0$). Yet self-reports are likely to have low sensitivity, meaning that many who search will not flag themselves by admitting it ($P(\text{flag}|\text{search}) < 1$).[2]

The two detection methods used in this paper are the *catch method*, which is defined as using a catch question to predict who will search on knowledge questions, and the *paradata method*, which collects data on respondents' engagement with the survey. The catch questions ask respondents to to name the year in which an obscure Supreme Court case was decided (Motta et al. 2016); this style of catch question was included in the 2020 American National Election Study (ANES). Internally, such items are likely to be highly specific, as few who do not search will answer it correctly ($P(\text{flag}|\neg\text{search})$ close to 0).[3] As long as the answer is easy to look up, they will also be highly sensitive ($P(\text{flag}|\text{search})$ close to 1). More serious threats emerge when the catch method is used to flag respondents who search on the knowledge questions. To the extent that respondents who search on catch questions do not search on knowledge questions, the catch method may be externally underspecific. To the extent that

[2]Evidence on the sensitivity of self report measures is mixed. Below, the discussion of Diedenhofen and Musch (2017) suggests that sensitivity may be quite low. Similarly, self reported rates of search often fall below catch-based estimates (e.g., compare Figure 1 in Clifford and Jerit (2016) to the Motta et al. (2016)), but do not always (e.g., the estimate of self-reported search prevalence in Jensen and Thomsen (2014) is similar to published catch-based estimates).

[3]Catch questions about the year in which Supreme Court cases were decided are likely to be more specific than difficult multiple choice items (Berinsky et al. 2012) or questions about the number of home runs hit by a baseball player (Bullock et al. 2015) because the Supreme Court questions have a larger number of plausible responses.

Table 1: Anti-search methods evaluated in this paper.

Method	Purpose	Flag	Sources of false negatives (undersensitivity)	Sources of false positives (underspecificity)
Catch	Detect	Correct answer to specially designed question	· *Internal:* Failing to find the correct answer. · *External:* Respondents who search on knowledge but not catch questions.	· *Internal:* Lucky guesses. · *External:* Respondents who search on catch but not knowledge questions.
Paradata	Detect	Survey ceases to be visible on screen	· Using a different device to look up the answer. · Using an incompatible browser to take the survey. · Asking someone for help.	· Non-search behavior that obscures the survey (e.g., reading an unrelated text message).
Pledge	Deter	None	Not applicable.	Not applicable.

respondents who search on knowledge questions do not search on catch questions, the catch method may be externally undersensitive.

The paradata method uses a short snippet of JavaScript to measure a behavior that is likely to indicate search: obscuring the survey with another browser window or application. Paradata may be underspecific if respondents trigger the flag for reasons other than search. For example, one may view a text message on their mobile phone. Paradata may be undersensitive if respondents search in some way that the method cannot detect. For example, respondents may use a different device to look up the answer, take the survey using an incompatible web browser, or ask another person for help. These sources of undersensitivity are artificially limited in supervised settings like laboratories, wherein survey-takers complete surveys using researcher-provided devices and have limited access to alternative means of search. This limits the generalizability of laboratory-based audits.

The paper does not make use of another detection method that appears in published research, examining subjects' browsing histories (Gooch and Vavreck 2019). Though browsing histories surely have untapped potential, this paper's paradata method has some relative advantages. Ethically, paradata methods have the advantage of only collecting data about the respondent's engagement with the survey itself. In this sense, the method is comparable

to other paradata that are routinely collected without raising ethical concerns, such as click counts and timers. By contrast, collecting browsing histories requires installing software that monitors respondents outside of the survey. Moreover, samples that include browsing history are more expensive to collect and are less likely to be representative of the general population (due to the need to consent to install software or visit a laboratory). For these reasons, indirect paradata methods like the one deployed here will are likely to be a better choice for cost-constrained researchers, as well as for higher-cost, general purpose surveys that value national representativeness (e.g., the ANES).

Among previously published research, this paper's approach is closest to that of Diedenhofen and Musch (2017, hereafter DM). This paper improves on this approach in two respects. First, DM's most detailed performance assessments of their paradata method, PageFocus, were conducted in a laboratory. As just noted, this setting artifically paradata methods' greatest vulnerabilities. By contrast, this paper's respondents had full access to modes of search that paradata methods cannot detect. Second, for the survey not conducted in a lab, DM use self-reports to measure the ground truth. The quantity DM report for the paradata's sensitivity is $P(\text{flag}|\text{subsequently admitted to searching})$. These self-admissions appear to be remarkably undersensitive: in the search-discouraged group, 18 respondents are flagged as having searched but only 3 admit it (see their Tables 1 and 2), suggesting a sensitivity rate of 17 percent or less.[4] By contrast, this paper only uses self-reports in an audit that verifies the interpretation of the two detection methods. Self-reports never enter the quantitative estimates of the two methods' performance.

More generally, this paper overcomes key limitations in existing research on countering information search in online surveys. First, it explicitly accounts for systematic measurement error by deriving an estimator that corrects for the resulting bias. By contrast, existing research on paradata methods rarely quantifies measurement error and never adjusts estimates to account for it. Second, this approach to error enables a more detailed and specific

[4]Given the reported data, the possible values of the self-reports' sensitivity are 3/18, 2/19, 1/20, or 0/21.

look at the efficacy of efforts to counter information search. There exists convincing evidence that some methods of detecting and deterring information search are *likely to help* in standard online settings, but less as to *how much* they help. Past research has shown that respondents who answer catch questions correctly are more likely to answer knowledge questions correctly (Gummer and Kunz 2019; Höhne et al. 2021) and lie about voting (Style and Jerit 2021). This strongly suggests that catch questions successfully identify search but does not quantify how much search actually occurs on knowledge questions among those who answer catch questions correctly. Previous evaluations of deterrence methods similarly rely on circumstantial evidence like a reduced number of correct answers (Vezzoni and Ladini 2017), fewer self-reports of searching (Clifford and Jerit 2016), or fewer correct answers to catch questions (Motta et al. 2016; Style and Jerit 2021). Third, this paper is the first to analyze the marginal effects of combining strategies. Whereas existing research examines each evaluated method in isolation, this paper examines the effects of layering methods atop one another.

Methodology

The empirical analysis examines three online surveys fielded in 2020 and 2021. For Studies 1 and 3, 2,176 and 6,687 respondents were recruited online by Lucid using a quota sampling procedure based on Census demographic benchmarks. For Study 2, 5,411 respondents were recruited through Amazon Mechanical Turk (MTurk). All subjects completed a captcha and provided informed consent, and all recruited on Lucid passed an attention check (Aronow et al. 2020; Peyton et al. 2021). Studies 2 and 3 were preregistered. Full details appear in Appendix B.1 (page A18).

Each study followed the same sequence. First, respondents were told that they would complete a short knowledge quiz. At this stage, one randomly selected group was asked to promise not to look up the answers. The other group's instructions omitted the pledge

but were otherwise identical.[5] Second, all respondents completed the knowledge quiz, which consisted of five political knowledge questions in Study 1 and seven in Studies 2 and 3. Additionally, Studies 2 and 3 randomized the format of two questions between closed-ended (i.e., multiple choice) or open-ended, bringing the total number of knowledge items considered in those studies to nine. Third, after answering some unrelated questions, all respondents completed a "pay to look" task. Each subject was asked to look up the answer to a catch question in exchange for a chance to win a $100 or $200 Amazon gift card, then to self-report whether and how they had looked up the answer.[6]

Each survey's political knowledge battery was outfitted with the paradata method and a catch question. Although the paradata method is similar in many respects to those described by Diedenhofen and Musch (2017) and Permut et al. (2019), it was developed independently. Rather than using one long block of code for the whole survey, users append a short snippet of code to each question and choose their own variable names for the output. This simplifies some aspects of the implementation and aids interpretability of the resulting data, but it also increases the time cost of adding additional questions. Appendix B.3 (page A22) and the replication file each contain one-page instructions for implementing the paradata method in Qualtrics.

The bias-correction derived in the next section consists of multiple parameters that are estimated using different survey items, but always with the same respondents. Accordingly, standard errors and confidence intervals are computed using the block bootstrap, which accounts for dependence between observations by randomly resampling at the respondent level. For example, it is used in analysis of conjoint experiments (Hainmueller et al. 2015), time series data (Bertrand et al. 2004), and analysis that pools across many knowledge questions (Graham 2020). Whenever the prevalence of search is estimated conditional on

[5]Simple random assignment, $p = 0.5$ (Gerber and Green 2012). Full text of the pledge appears in Appendix B.2 (page A19).

[6]In Study 1, the pay to look task was randomly assigned along with a task that *discouraged* the respondent from looking up the answer on the same question. This task was later judged not to add much value to the analysis, but is reported below for transparency.

another variable, all components of (2) are estimated within the subgroup of interest.[7]

Dealing with Measurement Error

Researchers do not directly observe information search in self-administered online surveys. Consequently, researchers use indirect measures to "flag" instances of suspected information search. Error in these measures adds systematic bias to estimates of the prevalence of search. The bias may be large or small, depending on the method's sensitivity and specificity. This section introduces a framework for quantifying and correcting these biases.

A bias-correction for the prevalence information search

Indirect methods of detecting search suffer from two basic problems: you may miss some people and you may wrongly accuse some poeple. To worry that a method does not flag all search is to worry that undersensitivity causes false negatives, i.e. that $P(\text{flag}|\text{search}) < 1$. To worry that a method flags people who do not search is to worry that underspecificity causes false positives, i.e. that $P(\text{flag}|\neg\text{search}) > 0$.

This paper's first point of departure from existing research is to quantify these sources of error and incorporate them into estimates of the prevalence of search. To begin, rewrite what researchers observe ($P(\text{flag})$) in terms of the estimand ($P(\text{search})$). By the law of total probability,

$$P(\text{flag}) = P(\text{flag}|\text{search})P(\text{search}) + P(\text{flag}|\neg\text{search})(1 - P(\text{search})), \tag{1}$$

where $P(\text{flag}|\text{search})$ is sensitivity and $P(\text{flag}|\neg\text{search})$ is the complement of specificity. Solv-

[7] In practice, the components of the bias correction hardly vary across subgroups, making this analytic choice empirically unimportant. However, it is important as a theoretical matter. For example, if the pledge had affected respondents' behavior in the pay-to-look task, failing to estimate the components of the bias correction separately for those assigned and not assigned to the pledge could introduce post-treatment bias (Montgomery et al. 2018).

ing for $P(\text{search})$ gives

$$P(\text{search}) = \frac{P(\text{flag}) - P(\text{flag}|\neg\text{search})}{P(\text{flag}|\text{search}) - P(\text{flag}|\neg\text{search})}. \tag{2}$$

Throughout the analysis, empirical versions of the right-hand side of (2) are used to estimate $P(\text{search})$.

In (2), the terms other than P(flag) amount to a bias correction. When one assumes that a measure is perfectly sensitive and specific (i.e., $P(\text{flag}|\text{search} = 1$ and $P(\text{flag}|\neg\text{search}) = 0$), the remaining terms disappear and (2) simplifies to $P(\text{search}) = P(\text{flag})$. By definition, to interpret the probability of being flagged as equivalent to the probability of search is to assume that one's measure is perfectly sensitive and specific.

Putting (2) into practice requires one to either estimate sensitivity and specificity or to assume that these problems can be safely ignored. The next section explains how this paper handles the necessary assumptions and approximations. Ultimately, the bias correction suggests that taking paradata-based estimates at face value slightly underestimates the prevalence of search. Appendix A.1 compares the corrected and uncorrected estimates in more detail (page A1).

Estimating sensitivity ($P(\text{flag}|\text{search})$)

To learn about the detection methods' failures to flag instances of search, the end of each study featured a "pay-to-search" task in which respondents were asked to look up the answer to a catch question. In exchange, respondents were offered entry into a drawing for a $100 (Studies 1 and 2) or $200 (Study 3) Amazon gift card. Immediately afterward, all respondents were asked to describe their search behavior: had they looked up the answer using the same device they are using to take the survey, looked some other way, or did they not look it up.[8] Table 2, which cross-tabulates the joint distribution of flags (correct answer and/or detected by the paradata) and self-reported search behavior, serves as the basis for

[8]Full text of the pay-to-search task appears in Appendix B.2 (page A19).

Table 2: Audit of detection methods based on pay-to-search task.

Self-reported search by detection status	Study 1 N	Study 1 % of total	Study 1 % of group	Study 2 N	Study 2 % of total	Study 2 % of group	Study 3 N	Study 3 % of total	Study 3 % of group
Correct + paradata flag	532	49.2		4394	77.8		3233	47.5	
Looked, same device	492	45.5	92.5	3896	69.0	88.7	2840	41.7	87.8
Looked, different device	19	1.8	3.6	164	2.9	3.7	154	2.3	4.8
Did not look	21	1.9	3.9	334	5.9	7.6	239	3.5	7.4
Correct + no paradata flag	166	15.3		721	12.8		1434	21.1	
Looked, same device	43	4.0	25.9	335	5.9	46.5	264	3.9	18.4
Looked, different device	99	9.1	59.6	221	3.9	30.7	1010	14.8	70.4
Did not look	24	2.2	14.5	165	2.9	22.9	160	2.4	11.2
Incorrect + paradata flag	52	4.8		108	1.9		252	3.7	
Looked, same device	36	3.3	69.2	64	1.1	59.3	171	2.5	67.9
Looked, different device	5	0.5	9.6	16	0.3	14.8	19	0.3	7.5
Did not look	11	1.0	21.2	28	0.5	25.9	62	0.9	24.6
Incorrect + no paradata flag	332	30.7		427	7.6		1885	27.7	
Looked, same device	44	4.1	13.3	114	2.0	26.7	146	2.1	7.7
Looked, different device	51	4.7	15.4	66	1.2	15.5	261	3.8	13.8
Did not look	237	21.9	71.4	247	4.4	57.8	1478	21.7	78.4
Total	1082			5650			6804		

the following analysis.

Large majorities of respondents in each study complied with the request to look up the answer. The percentage of respondents who either answered correctly or was flagged in the paradata was 67.0 in Study 1, 93.5 in Study 2, and 72.3 in Study 3. Among those flagged by both technologies, about 90 percent self-reported that they looked up the answer using the same device they used to take the survey (Table 2, first group of rows).

The paradata's failures are represented by the "correct + no paradata flag" category, which indicates that the respondent reported the correct answer to the pay-to-search question but was not flagged in the paradata (Table 2, second group of rows). The self-report question captures two reasons why this group might not be flagged: browser incompatibility or using a different device. In Studies 1 and 3, a substantial majority of the "correct + no paradata

flag" group claimed to have used a different device. In Study 2, about one-third reported the same. Most remaining respondents reported using the same device, suggesting browser incompatibility or misreporting.

For the paradata method, $P(\text{flag}|\text{search})$ is calculated as the proportion of likely search that is flagged.[9] This equals 0.78 in Study 1, 0.86 in Study 2, and 0.71 in Study 3. The analysis assumes that this quantity is constant across knowledge questions, which is reasonable given that the reasons for undersensitivity are either constant for the duration the survey (e.g., browser incompatibility) or plausibly thought of as reflecting individual-level dispositions (e.g., a tendency to use a different device). This assumption does have vulnerabilities. In particular, subjects may not try as hard to avoid detection when told that search is allowed. Relative to the prevailing practice of analyzing paradata methods as if they are error-free, however, the assumptions that error exists and is constant across questions relaxes stronger, less credible assumptions.[10]

The catch method's failures to detect search are represented by the "incorrect + paradata flag" category (Table 2, third group of rows). Among this group, substantial majorities report having looked up the answer using the same device in all three studies. By contrast, few in the "incorrect + no paradata flag" category self-report having looked up the answer, suggesting that the flagged respondents who answer incorrectly really did try to look up the answer. Evidence that catch questions can be internally undersensitive emerged due to an informative accident in Study 2, wherein some respondents were fooled by incorrect answers in search results (e.g., reporting the date of the district court case Oliver v. Alexander County Housing Authority (1982) rather than the Supreme Court case Oliver v. Alexander (1832)). To sidestep the need to correct for this source of undersensitivity, the analysis treats

[9]Specifically, $P(\text{paradata flag}|\text{paradata flag or answered correctly})$. For example, in Study 1, the calculation is $(492 + 54)/(492 + 166 + 54)$.

[10]More specifically, analyzing paradata as though they are perfectly sensitive amounts to an assumption that $\Pr(\text{flag}|\text{search})=1$. The approximation used in this paper allows one to assume that $\Pr(\text{flag}|\text{search})<1$, which increases estimates of the prevalence of search for any reasonable detection method (see Appendix A.1, page A1). The concern that subjects do not try as hard to avoid detection on pay-to-search tasks amounts to a concern that the value used for $\Pr(\text{flag}|\text{search})$ is still too large. In this case, the bias correction would constitute an improvement over existing practice but would still under-correct.

the catch question's estimand as the probability of successful search rather than attempted search. Because the definition of success (a correct answer) is the same as the flagging procedure, undersensitivity cannot exist.

As noted above, the catch method is more vulnerable to external undersensitivity, when it is used to approximate search on the knowledge battery rather than on the catch question itself. Further below, the paper introduces a strategy for using bias-corrected estimates from the paradata method to estimate the catch method's external sensitivity.

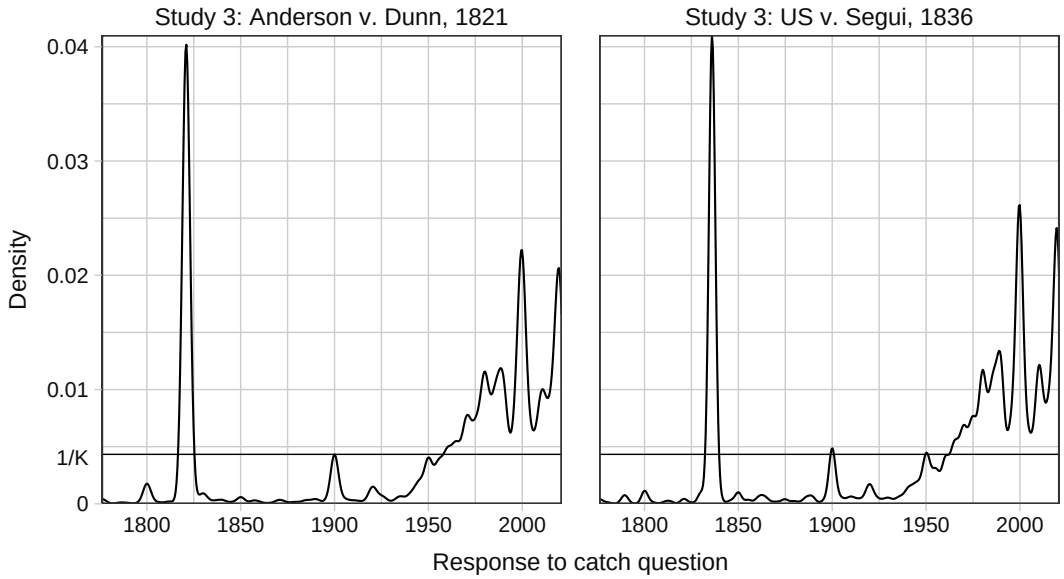
Estimating underspecificity ($P(\textbf{flag}|\neg\textbf{search})$)

The paradata method flags all instances in which the survey becomes partially or fully obscured on the respondent's screen. This will produce false positives, and consequently be underspecific, if behavior other than information search triggers the flag. To approximate $P(\text{flag}|\neg\text{search})$ for the paradata, it was added to a set of baseline questions where looking up the answer is unlikely to be necessary (e.g., age) or undefined due to the question's subjective nature (e.g., interest in politics).[11] In all three studies, false positives are heavily concentrated among a small percentage of respondents who are repeatedly flagged. Consequently, the baseline items were divided into two sets, one for screening out those likely to inflate the number of false positives and a second for estimating $P(\text{flag}|\neg\text{search})$ among those not screened out. Among the remaining respondents, $P(\text{flag}|\neg\text{search})$ was estimated to be 0.007 in Study 1, 0.015 in Study 2, and 0.010 in Study 3. These rates are roughly constant across the baseline questions.

The catch method flags all respondents who answer catch questions correctly. Internally, lucky guesses are the key source of the false positives that cause underspecificity. Researchers endeavor maximize specificity by choosing catch questions with many plausible response options, thereby minimizing the probability of a correct guess. For example, if

[11] When search is non-existent, the percentage of respondents who are flagged as searchers is equal to the underspecificity rate, i.e. if $P(\text{search}) = 0$ then $P(\text{flag}) = P(\text{flag}|\neg\text{search})$). To see this, plug $P(\text{search}) = 0$ into (1).

Figure 1: Response distribution for catch questions, Study 3.



Note: Figure displays the PDF of responses to the catch questions in Study 3. Equivalent figures for Studies 1 and 2 appear in Appendix A.2 (page A7).

guessers choose among plausible response options with equal probability, the probability of correctly answering a question about the date of a Supreme Court case (Motta et al. 2016) is $\frac{1}{2021-1776} \equiv 0.004$.[12] It is possible to reduce this further. Outside of those who look up the answer, responses concentrate in recent years and at multiples of 5 (Figure 1). To estimate the expected rate of lucky guessing, Appendix A.2 uses local linear regression to estimate the probability distribution function (PDF) of incorrect answers to the catch questions in all three studies (page A6). These estimates suggest that $P(\text{flag}|\neg\text{search})$ is about 0.001 for catch questions that avoid commonly-guessed correct answers. For simplicity, this is rounded down to 0 in all analysis.

Once again, the catch method is more specific internally than externally; that is, the catch method is more vulnerable to underspecificity when it is used to predict who will look up answers to knowledge questions. A strategy for using bias-corrected estimates from the paradata to examine the catch method's external sensitivity is introduced further below.

[12]The author is aware that the Constitution was ratified in 1789, but some survey respondents are not.

Detecting Information Search

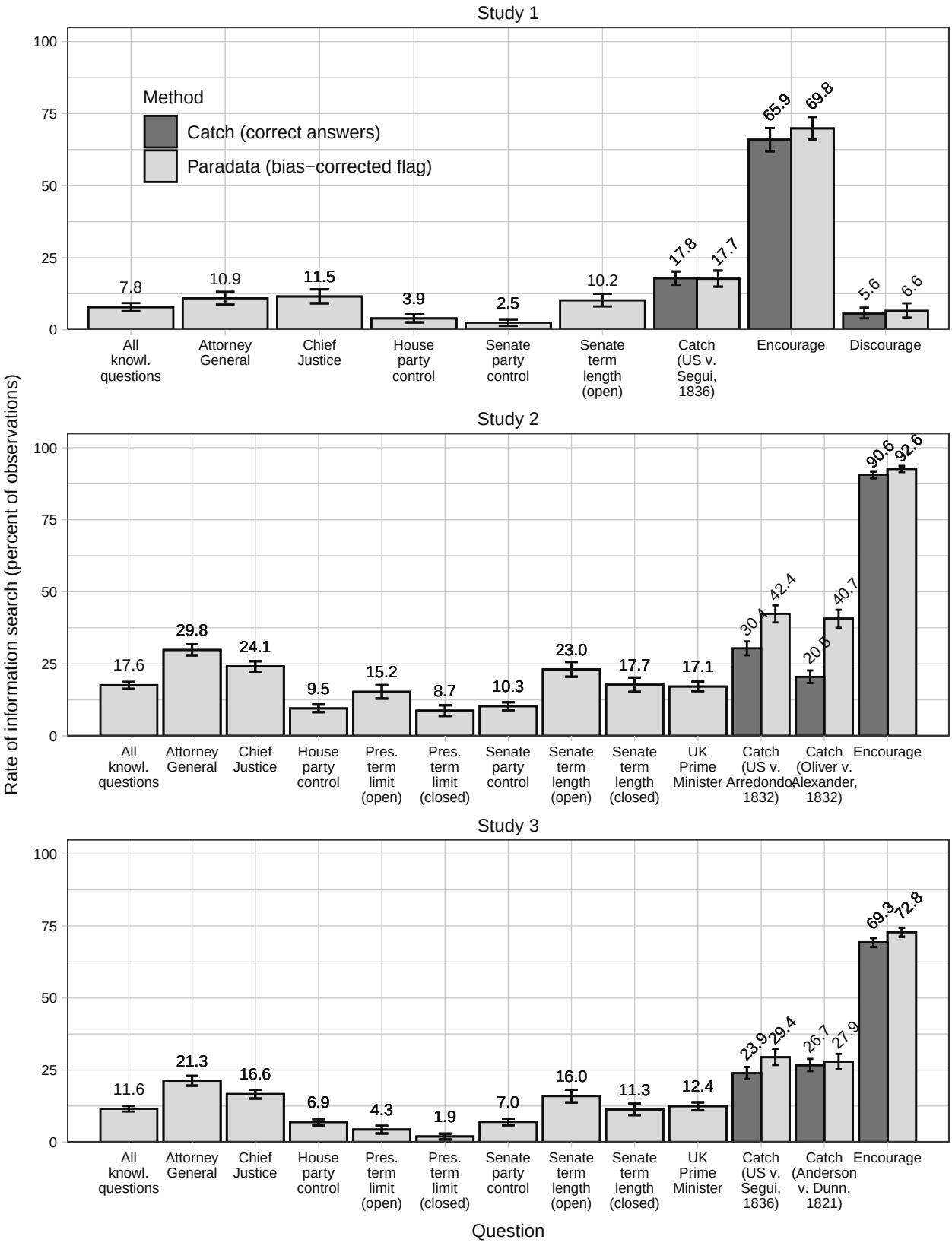
The bias correction approach is first applied to the prevalence of search when no methods of combatting it are employed. Figure 2 presents these base-rate estimates for each question in all three studies. The x-axis labels questions and the y-axis is an empirical estimate of (2). For catch questions, dark grey bars show the proportion of correct answers, which should equal (2) if catch questions have perfect internal sensitivity and specificity.

The estimates indicate that search is common. Across all instances in which a respondent answered a knowledge question, search occurred in 7.8 percent of observations in Study 1, 17.6 percent in Study 2, and 11.6 percent in Study 3. Though this is a large share of the data, it is less than published estimates based on catch questions.

The overall average hides substantial variability between questions. In Study 1's knowledge battery, search ranged from about 2.5 percent (on a question about party control of the Senate) to about 11.7 percent (on a question about Chief Justice John Roberts' job or political office). In Study 2, search ranged from 8.8 percent (the closed-ended version of the presidential term limit question) to about 29.9 percent (on a question about Attorney General Merrick Garland's job or political office); the questions bringing up the extremes in Study 1 were the third-smallest and second-largest in Study 2. Study 3, exactly replicated Study 2's question battery. Here, search was most rare and most common on the same survey items (2.0 and 21.8 percent), but fell below the Study 2 level in every instance. At least in these samples, MTurk respondents were more likely to look up answers than Lucid respondents.

To learn about the degree to which variation in search is a function of response scales as opposed to question content, Studies 2 and 3 each randomized the response format for two of the questions, regarding presidential term limits and the length of a Senate term. Search was two to five percentage points (p.p.; 30 to 100 percent) more common on open-ended than closed-ended questions. This is despite the fact that respondents were two to

Figure 2: Estimated rate of information search by question and study.



four p.p. less likely to answer each closed-ended question correctly.[13] This suggests that the greater prevalence of search on open-ended questions roughly halves the difference in apparent knowledge between open-ended and multiple choice questions.

Search is much more common on catch questions than on knowledge questions (Figure 2, center-right). In all three studies, the rate of search on catch questions more than doubled the rate for the knowledge battery: 17.7 versus 7.8 in Study 1, 42.4 and 40.7 versus 17.6 in Study 2, and 29.4 and 27.9 versus 11.6 in Study 3.

The catch method fares modestly better as an approximation for the proportion of *subjects* who engage in search behavior. Table 3 compares the percentage of correct answers to the catch question to the paradata-based estimate of the percentage of individuals who searched on at least one question. The estimates are separated by study and the presence or absence of the pledge. All estimates of the difference in proportions are negative, indicating that the catch method tends to underestimate the percentage of subjects who search. The differences are small in Study 1 (-2.1 and -2.7 p.p.), large in Study 2 (-15.8 and -11.1 p.p.), and in between in Study 3 (-5.8 p.p. and -7.0. p.p.). These results likely depend to some extent on the battery's content and length. Had it omitted the questions with the highest rates of search, perhaps the catch method would have produced more accurate estimates.

The catch questions' inconsistent performance in Table 3 is partly attributable to internal undersensitivity in the catch questions used in Study 2. Recall that whereas the paradata method detects failed searches, the catch method only counts those who successfully look up the answer. Figure 2 compares the catch- and paradata-based estimates for the catch questions. The estimates are usually similar, with the percentage of correct answers falling slightly below the paradata flag in most cases. The exceptions are the two catch questions used in Study 2, Oliver v. Alexander (1832) and United States v. Arredondo (1832). These Supreme Court cases were selected based on the false positive-minimizing principles developed in Study 1 (early 1800s, no multiples of 5). However, these questions

[13] A similar pattern is reported by Höhne et al. (2021).

Table 3: Estimated percentage of subjects searching by study.

	No pledge			Pledge		
	% correctly answering catch question	% searching on at least one knowl. question	Difference	% correctly answering catch question	% searching on at least one knowl. question	Difference
Study 1	17.8	19.9	-2.1	8.5	11.2	-2.7
	(1.2)	(1.4)	(1.5)	(0.8)	(1.2)	(1.2)
Study 2	25.7	41.5	-15.8	14.9	26.0	-11.1
	(0.9)	(1.1)	(1.1)	(0.7)	(0.9)	(0.8)
Study 3	25.3	30.8	-5.5	9.6	15.7	-6.1
	(0.8)	(1.0)	(0.9)	(0.5)	(0.8)	(0.7)

Note: Figure compares the percentage of respondents who correctly answer a catch question to the paradata-based estimate of the proportion who searched on at least one knowledge question. First column is the percentage of correct answers on the catch questions. Second column is an estimate of the percentage of individuals searching on at least one knowledge question based on equation (2). Third column is the difference between them.

also had a feature that drove up the share of false negatives: incorrect answers appearing in search queries that, to a satisficing searcher,[14] could look like the correct answer (see Appendix A.2, page A6). For example, the most common incorrect answer to the Oliver v. Alexander question, 1982, is the date of a district court case by the name of Oliver v. Alexander County Housing Authority. Such responses explain most of the difference between the catch- and paradata-based estimates.

In summary, the results so far suggest that search is common in political knowledge quizzes. Search varies considerably between questions and also appears to vary across survey platforms. Catch questions see more search than any other question type, but still manage to underestimate the proportion of respondents who search at least once. Catch questions thus constitute a misleading proxy for prevalence of search on knowledge questions.

[14]That is, one who looks up the answers but does so in a quick and haphazard manner. On satisficing, see Krosnick (1991).

Deterring Information Search

The framework is next applied to the efficacy of deterrence methods. Applying the bias correction requires one to estimate (2) for two groups, then take the difference between them. Suppose the groups are defined by $X \in 0, 1$. The resulting difference in means,

$$P(\text{search}|X = 1) - P(\text{search}|X = 0) \tag{3}$$

can be used to compare the rate at which any two groups of respondents look up answers in a political knowledge survey. Here, the groups are defined by a randomly assigned pledge not to look up the answer.

Table 4a presents difference in means estimates for each knowledge question in all three studies. Overall, the pledge reduced search by about half in all three studies: 50.7 percent in Study 1, 47.7 percent in Study 2, and 56.5 percent in Study 3. Although this amounts to a substantial reduction, it also leaves a substantial amount of search in the data. Among respondents who took the pledge, search was estimated to have occurred in 3.8 percent of responses in Study 1, 9.2 percent of responses in Study 2, and 5.0 percent of responses in Study 3.

The pledge's effect is similar from question to question. All but one point estimate of the effect of the pledge is both negative and statistically significant, suggesting that the pledge works for a range of questions. For nearly all questions, the percentage reduction ranges from roughly 35 to 65 percent of the baseline. The largest outlier is the question with the lowest bases rate of search, which makes the percentage reduction difficult to estimate with precision. Variation in the percentage point effects is somewhat more pronounced. Excluding the outlier, this ranged from 2.4 p.p. (Study 1, House party control) to 14.7 p.p. (Study 2, Attorney General), a sixfold difference. This variation in percentage point effects reflects variation in the base rate of search, not any clear interaction between the pledge and question content. Across all questions in all three studies, the correlation between the base

Table 4: Deterrent effect of pledge.

(a) Knowledge battery.

	Study 1				Study 2				Study 3			
Question	No pledge	Pledge	Effect	%	No pledge	Pledge	Effect	%	No pledge	Pledge	Effect	%
All knowl. questions	7.8 (0.7)	3.8 (0.5)	-3.9 (0.9)	-50.7	17.6 (0.6)	9.2 (0.5)	-8.4 (0.7)	-47.7	11.6 (0.5)	5.0 (0.4)	-6.6 (0.6)	-56.7
Attorney General	10.9 (1.1)	5.8 (0.9)	-5.1 (1.4)	-46.6	29.8 (1.0)	14.7 (0.8)	-15.0 (1.2)	-50.5	21.3 (0.9)	9.0 (0.6)	-12.3 (1.0)	-57.9
Chief Justice	11.5 (1.2)	5.6 (0.9)	-5.9 (1.5)	-51.5	24.1 (0.9)	13.4 (0.7)	-10.7 (1.2)	-44.4	16.6 (0.8)	6.9 (0.5)	-9.7 (1.0)	-58.6
House party control	3.9 (0.7)	1.5 (0.5)	-2.4 (0.9)	-60.6	9.5 (0.7)	3.7 (0.4)	-5.8 (0.8)	-61.6	6.9 (0.6)	2.5 (0.4)	-4.4 (0.7)	-64.1
Pres. term limit (open)					15.2 (1.2)	9.7 (0.9)	-5.5 (1.5)	-36.1	4.3 (0.7)	2.3 (0.5)	-2.0 (0.9)	-47.2
Pres. term limit (closed)					8.7 (0.9)	5.6 (0.8)	-3.1 (1.2)	-35.5	1.9 (0.5)	2.5 (0.6)	0.6 (0.8)	31.5
Senate party control	2.5 (0.6)	0.2 (0.3)	-2.2 (0.7)	-91.1	10.3 (0.7)	3.2 (0.4)	-7.0 (0.8)	-68.5	7.0 (0.6)	2.5 (0.4)	-4.5 (0.7)	-64.6
Senate term length (open)	10.2 (1.1)	6.1 (0.9)	-4.1 (1.5)	-40.6	23.0 (1.3)	13.7 (1.0)	-9.3 (1.7)	-40.4	16.0 (1.1)	6.9 (0.8)	-9.0 (1.4)	-56.5
Senate term length (closed)					17.7 (1.2)	8.2 (0.9)	-9.6 (1.5)	-53.9	11.3 (1.0)	5.8 (0.8)	-5.5 (1.2)	-48.9
UK Prime Minister					17.1 (0.8)	10.7 (0.7)	-6.4 (1.1)	-37.5	12.4 (0.7)	5.5 (0.5)	-7.0 (0.9)	-56.1

(b) Catch questions.

		Paradata method				Catch method (% correct)			
Study	Question	No pledge	Pledge	Effect	%	No pledge	Pledge	Effect	%
Study 1	US v. Segui, 1836	17.7 (1.4)	8.7 (1.0)	-9.0 (1.8)	-50.9	17.8 (1.2)	8.5 (0.8)	-9.3 (1.4)	-52.3
Study 2	US v. Arredondo, 1832	42.4 (1.5)	27.5 (1.3)	-14.9 (2.0)	-35.1	30.4 (1.3)	20.0 (1.0)	-10.4 (1.6)	-34.3
	Oliver v. Alexander, 1832	40.7 (1.6)	18.7 (1.2)	-22.0 (2.1)	-54.1	20.5 (1.1)	9.1 (0.8)	-11.4 (1.4)	-55.5
Study 3	US v. Segui, 1836	29.4 (1.4)	11.5 (1.0)	-17.9 (1.7)	-60.9	23.9 (1.1)	9.0 (0.7)	-14.9 (1.3)	-62.3
	Anderson v. Dunn, 1821	27.9 (1.3)	11.7 (1.0)	-16.2 (1.7)	-58.2	26.7 (1.1)	10.1 (0.7)	-16.6 (1.3)	-62.2

Note: Cell entries are estimates of the prevalence of search. Block bootstrapped standard errors in parentheses. In subtable (a), all estimates use the paradata detection method. In subtable (b), column headers indicate whether the estimates are based on the paradata detection method or the catch method (i.e., flagging all correct answers as suspected search).

rate of search and the treatment effect of the pledge is -0.94. The more search there is to begin with, the more is eliminated by the pledge.

The catch questions provide an opportunity to cross-check the paradata-based estimates of the pledge's efficacy. In Study 1, the paradata-based estimate suggests that the pledge reduced search on the catch questions by 9.0 p.p. (50.9 percent). Similarly, the proportion of correct answers declined by 9.3 p.p. (52.3 percent). In Study 2, the paradata-based estimates suggest larger absolute reductions (14.9 and 22.2 versus 10.4 and 11.4 p.p.). However, as a percentage of the base rate, the two sets of estimates are similar (35.1 and 54.1 versus 34.3 and 55.5 percent). In Study 3, the estimates are once again similar in magnitude and percentage point terms (about a 60 percent reduction). This suggests that even though catch questions are an untrustworthy proxy for the prevalence of search, they are a reasonable barometer for the efficacy of deterrence methods.

Although the pledge offers significant value as a deterrent, its failure to fully eliminate search leaves something to be desired. Even net of a 50 percent reduction, search remains fairly common, representing 10 percent or more of the sample on several questions in Studies 2 and 3. The next section examines the benefits and costs of going further.

Eliminating Information Search

Efforts to deter information search are often paired with detection, enabling researchers to identify observations that are affected by search despite the researcher's best efforts to discourage it. This gives researchers the option to take further, *ex post* steps to eliminate search, either in the main analysis or as a robustness check. In particular, researchers may treat contaminated responses as missing data, then handle the missingness by dropping observations or imputing values. The missing data constitutes a cost that is not present with deterrence measures.

This section builds on the framework above to define and estimate quantiites that capture the tradeoff between eliminating search and creating missing data. The key new

step is to use the paradata to evaluate the catch method's external specificity. Rather than accept paradata flags as a direct measure of search, the analysis conditions estimates of (2) on whether or not the subject is flagged by the catch method. Combined with further algebra, this enables estimates of quantities like the amount of search on knowledge questions that is not detected by catch questions.

The paradata method is superior to the catch method in one sense that is not captured by the quantities estimated below: it generates partial information about respondents who search. In each study, roughly one-third of suspected searchers were flagged only once, and another third were flagged on less than half of the items.[15] A reasonable knowledge score can be estimated for individuals like this using a model that makes use of the other items in a principled manner, e.g. an item response theory (IRT) model. In this case, the analysis below substantially overstates the missing data cost of the paradata method. By contrast, the catch method measures it at the subject level only. This limits the *ex post* solution toolkit to dropping suspected searchers from the analysis entirely or imputation based on variables that are not part of the knowledge battery.

Quantifying tradeoffs

This section will use four quanities. In the language of consumerism, the first two help researcher-shoppers understand what they're buying, the third is the price, and the fourth is the end product.

The first quantity is sensitivity, which has been defined. Sensitivity answers the question, "what percentage of search does this method detect?" For the paradata method, this was approximated earlier using the pay-to-search task. For the catch method, the key threat to external sensitivity is the possibility that it is not always the same individuals who search.

[15] In Study 1, 45 percent of suspected searchers were flagged on one of the five questions, with another 24 percent flagged on a second question. In Study 2, about 32 percent of suspected searchers were flagged on only one of the seven questions, with another 34 percent flagged on either two or three questions. In Study 3, about 37 percent of suspected searchers were flagged only once in sevel questions, with another 34 percent flagged two or three times.

The ability to use the paradata to estimate the rate of search for different subgroups of subjects (e.g., (3)) provides an opportunity to examine this empirically. The catch method's external sensitivity will be estimated as

$$P(\text{flag}|\text{search}) = \frac{P(\text{search}|\text{flag})P(\text{flag})}{P(\text{search})}. \tag{4}$$

One of the quantities on the right-hand side is directly measured ($P(\text{flag})$), i.e. answering the catch question correctly). The other two can be estimated using the paradata. The denominator is simply (2), while the first term in the numerator is (2) for the subset of respondents who answered the catch question correctly.

The second quantity is the *positive predictive value* (James et al. 2021, 145-49),which can be written as $P(\text{search}|\text{flag})$. This quantity answers, "how much of what is flagged is actually search?" For the catch method, the positive predictive value can be estimated by calculating (2) among those who answered the catch question correctly, just as it appears in the denominator of (4). For the paradata, the positive predictive value will be estimated as

$$P(\text{search}|\text{flag}) = \frac{P(\text{flag}) - P(\text{flag}|\neg\text{search})P(\neg\text{search})}{P(\text{flag})}. \tag{5}$$

Of the three unique quantities that constitute the right-hand side, $P(\text{flag})$ is observed, $P(\neg\text{search})$ is the complement of (2), and $P(\text{flag}|\neg\text{search})$ is the pay-to-search estimate of the paradata's underspecificity.

The third quantity is the probability of being detected, $P(\text{flag})$. This answers, "what's the price?" or equivalently, "how much missing data will I create if I take this approach?" This is simply the percentage of respondents who trigger the flag, which is equivalent to the percentage of observations lost if all instances of suspected search are treated as missing.

The fourth quantity is the amount of search remaining in the data, i.e. the probability of search among unflagged observations ($P(\text{search}|\neg\text{flag})$). This quantity could also be called the complement of the negative predictive value. It answers the question, "how much search

Figure 3: Sensitivity by detection method.

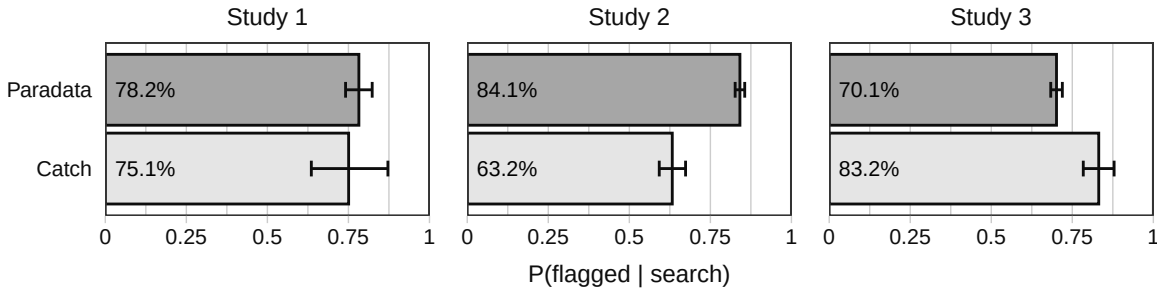
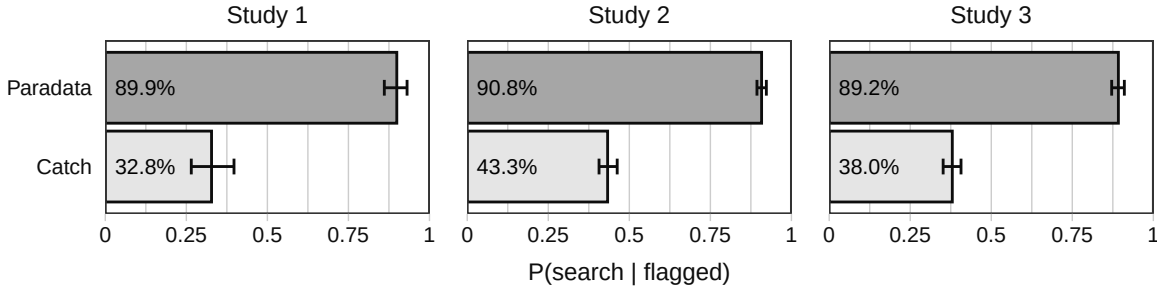


Figure 4: Positive predictive value by detection method.



will be left in the data I do not treat as missing?" This can be estimated by

$$P(\text{search}|\neg\text{flag}) = \frac{P(\neg\text{flag}|\text{search})P(\text{search})}{P(\neg\text{flag})}, \tag{6}$$

which is simply a combination of estimators that have been defined above. The first term in the numerator is the complement of (4). The second is (2). The denominator is observed.

Estimates

To begin understanding the search elimination-missing data tradeoff, examine the sensitivity estimates that appear in Figure 3. In all three studies, the paradata identified more than two-thirds of search: 78.2 percent in Study 1, 84.1 percent in Study 2, and 70.1 percent in Study 3. The catch method was about equally sensitive, flagging 75.1 percent of search in Study 1, 63.2 percent in Study 2, and 83.2 percent in Study 3.

The estimates of the positive predictive value appear in Figure 4. In all three studies, the paradata method is well-targeted, with search estimated to have taken place in at least

90 percent of flagged observations. For every nine subjects who are correctly flagged as having searched, one is incorrectly flagged. The catch method is less precise, with predictive values that fall between 30 and 45 percent in all three studies. This means that for every two instances of search on knowledge questions that are correctly identified by the catch method, three or four observations that were not affected by search are generated. Put simply, even though the catch method flags just as high a percentage of true positives, this comes at a higher cost in terms of false positives.

To understand the implications for practice, it is helpful to examine the tradeoff between price and the bottom line. Figure 5 plots the third quantity, the percentage of data lost when instances of suspected search are treated as missing data, against the fourth quantity, the amount of search present in the remaining data. The x-axis is a combination of detection technologies. The leftmost point on the x-axis represents a survey in which no steps were taken to eliminate search. For this scenario, search sits at the same base rates reported in Figure 2, while the percentage of data eliminated sits at 0 percent. Each point to the right represents some combination of the pledge, paradata, and catch methods.

The left column of Figure 5 examines the marginal costs and benefits of the paradata and catch methods in the absence of a pledge. In Study 1, treating suspected search as missing data shrinks the number of observations by 6.8 percent in order to reduce the rate of search among the remaining observations from 7.8 to 1.7 percent. The catch method yields less benefit at a higher cost. After flagging the 17.8 percent of respondents who answered the catch question correctly, search would still affect 3.1 percent of the remaining observations. If the paradata method is already in place, the catch method reduces search by another 1.1 percent of the remaining data (from 1.7 to 0.6 percent) at a marginal cost of 13.8 percent of the initial observations (from 6.7 to 20.5 percent missing). The results of Studies 2 and 3 are similar, but shifted upward in magnitude and reflective of Study 2's relatively undersensitive catch questions. In Study 2, the paradata method flags 16.1 percent of the data to reduce search to 2.8 percent; the catch method, 25.7 percent to reduce search to 9.3 percent. The

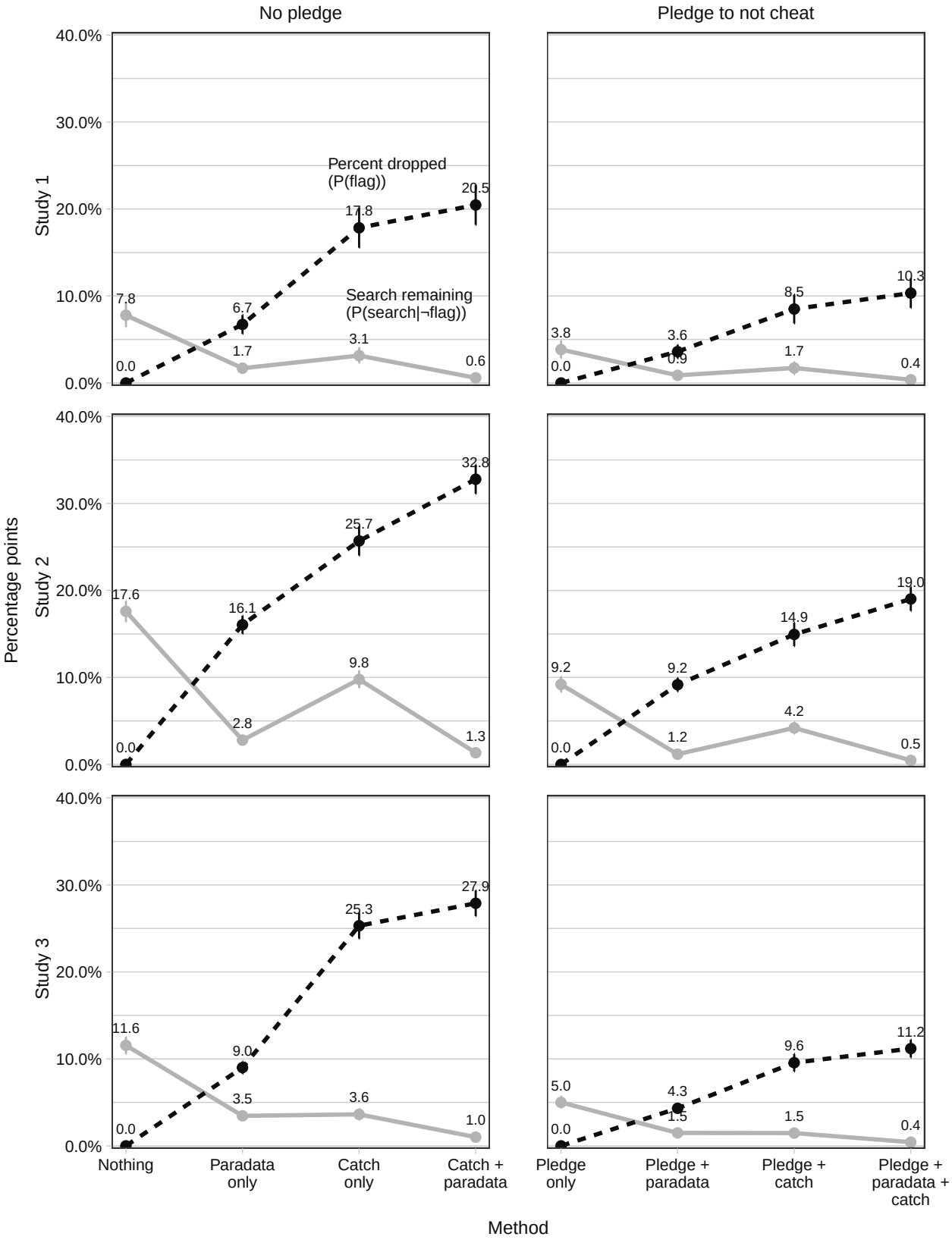
marginal benefit of the catch method is 1.5 percent less search (from 2.8 to 1.3) at a cost of 16.7 percent of observations (from 16.1 to 32.8). In Study 3, the paradata method costs 9.0 percent of the data to reduce remaining search to 3.5 percent, while the catch method costs 25.3 percent of the data to reduce search to 3.6 percent. Adding the catch method costs 18.9 percent of the data (from 9.0 to 27.9) to reduce search by 2.5 percent of the remaining data (from 3.5 to 1.0).

The right column of panels in Figure 5 examines the same tradeoff in the presence of a pledge not to search. Broadly speaking, the pledge flattens the two curves. The less search occurs to begin with, the lower the absolute cost of treating suspected instances of search as missing data. By contrast, the pledge does not have much effect on the per-unit cost. Instead, when a pledge is present, detection yields about half the benefit at about half the cost. For example, in Study 1, adding the paradata method still cuts search by about three-quarters regardless of whether a pledge is present (from 4.8 to 1.2 percent with a pledge, versus 9.6 to 2.2 without one).

The reader may note what could appear as a contradiction in the Study 3 results: despite being more sensitive than the paradata (Figure 3), the catch method leaves a slightly larger percentage of search among the observations that are not flagged for suspected search (Figure 5). This highlights a subtle but important point regarding how sensitivity and specificity interact to shape the bottom line. The catch method achieves its high rate of sensitivity by casting a wide net; recall that search is far more common on catch questions than on any knowledge question. Relative to the paradata, the catch method sometimes throws out a bit more of the bad, but it always throws out a lot more of the good. In this way, the catch method's lack of external specificity undermines the benefit of its impressive external sensitivity. Even when the catch method is better than the paradata at detecting search, using the catch method to purge data of search leaves behind fewer observations which are at least as affected by search on a per-unit basis.

The combination of the pledge, paradata, and catch methods constitutes a remarkably

Figure 5: Tradeoff between eliminating search and missing data.



effective strategy for combatting information search. To see this, compare the leftmost and rightmost estimates in Figure 5. The leftmost estimates ("nothing") represent the situation when nothing is done to combat search, and the rightmost ("pledge + paradata + catch") represent what is achieved by all three methods in combination. In Study 1, implementing all three methods reduces search from 7.8 percent to 0.4 percent of observations, at a cost of converting 10.3 percent of observations to missing. In Study 2, reducing search from 17.6 percent to 0.5 percent costs 19.0 percent of the data. In Study 3, reducing search from 11.6 percent to 0.4 percent costs 11.2 percent of the data. The size of the benefit is comparable to the cost due to the presence of the pledge, which eliminates half of search *ex ante* without any cost in terms of missing data.

The costs and benefits of supplementing paradata with the catch method depend on the base rate of search. Whereas the paradata flag fewer respondents when search is less common, the catch method always flags the same group of respondents. This means that adding the catch method to a lower-search question requires one to treat *more* observations as missing (because fewer are already flagged by the paradata) in order to obtain a *smaller* benefit (because there is less search to eliminate). To illustrate this, Appendix A.3 presents the same information for every question in all three studies (page A11). When search is common, the catch method offers some marginal benefit. In the most efficient case, the Attorney General question in Study 3, adding the catch method reduces search by 1.8 percent of the remaining observations (from 2.7 to 0.9) at a cost of 5.2 percent of the data (from 7.1 to 12.3). This equals about one unit of search eliminated for every three units of missing data. By contrast, when search is rare, the benefits decline while the costs simultaneously rise. For example, on Study 3's House party control question, adding the catch method reduces search by 0.6 p.p. of the data (from 0.7 to 0.1) at a cost of 8.0 percent of the original data (2.6 to 10.6). This equates to one unit of search eliminated for every 13 units of missing data. In Studies 1 and 2, the cost per unit on the House and Senate party control questions is even higher, in some cases exceeding one unit of search reduction for every 50 units of

missing data. At that price, researchers may prefer to tolerate a few fractions of a percentage point of additional search.

Heterogeneous Effects

Researchers deciding how to handle information search must also consider how the effects of these strategies vary with subject characteristics. This section presents an exploratory analysis of how the detection and deterrence methods shape these two issues, representativeness and between-group differences in search. To do so, it examines the each measure in isolation, comparing the four scenarios labelled "nothing," "paradata only," "catch only" and "paradata only" in the analysis above. As in the previous section, the methods are compared in terms of the proportion of search remaining in the data after the method has been used to eliminate search.

The implications for deterrence and detection vary as a function of their respective *ex ante* and *ex post* natures. Because deterrence can eliminate search without creating missing data, it has no effect on sample composition. However, if groups that are more prone to search are also more resistant to deterrence, deterring search could actually increase the influence of search on estimated between-group differences in knowledge. This would be a lost opportunity, but has no implications for sample representativeness. By contrast, eliminating search through detection forces researchers into a zero-sum tradeoff between dropping observations and reducing between-group differences in the prevalence of search. Dropping observations can only reduce between-group differences if it disproportionately drops groups that search more, altering the composition of the sample. This can only be avoided if groups are dropped at the same rate, in which case between-group differences would not decrease.

The subgroups are defined by ten pre-treatment measures that appeared in all three surveys. For binary characteristics, between-group differences are simply the difference between the two groups. For all other measures, the estimates reflect the difference between a

respondent who is one standard deviation above and below the mean on that measure.[16]

At baseline, the data reflect several between-group differences in the prevalence of information search (Figure 6; difference in means tests in Appendix Table A7, page A14).[17] Several demographic differences exist: younger respondents are more likely to search than older respondents (+10.4 p.p), men are more likely to search than women (+1.7 p.p.), and non-white and Hispanic respondents are more likely to search than their counterparts (+7.4 and +8.3 p.p.). Other attitudes and traits also predict search. Search is more common among respondents who are more cognitively reflective (+2.9 p.p.), endorse more conspiratorial beliefs (+7.9 p.p.), are more interested in politics (+2.7 p.p.), prefer the Democratic party (+2.7 p.p), and are stronger partisans (+2.7 p.p.). In some cases these differences in search behavior reinforce established differences in measured knowledge (e.g., gender, interest in politics), while in other cases they counteract such differences (e.g., age, conspiracy beliefs).

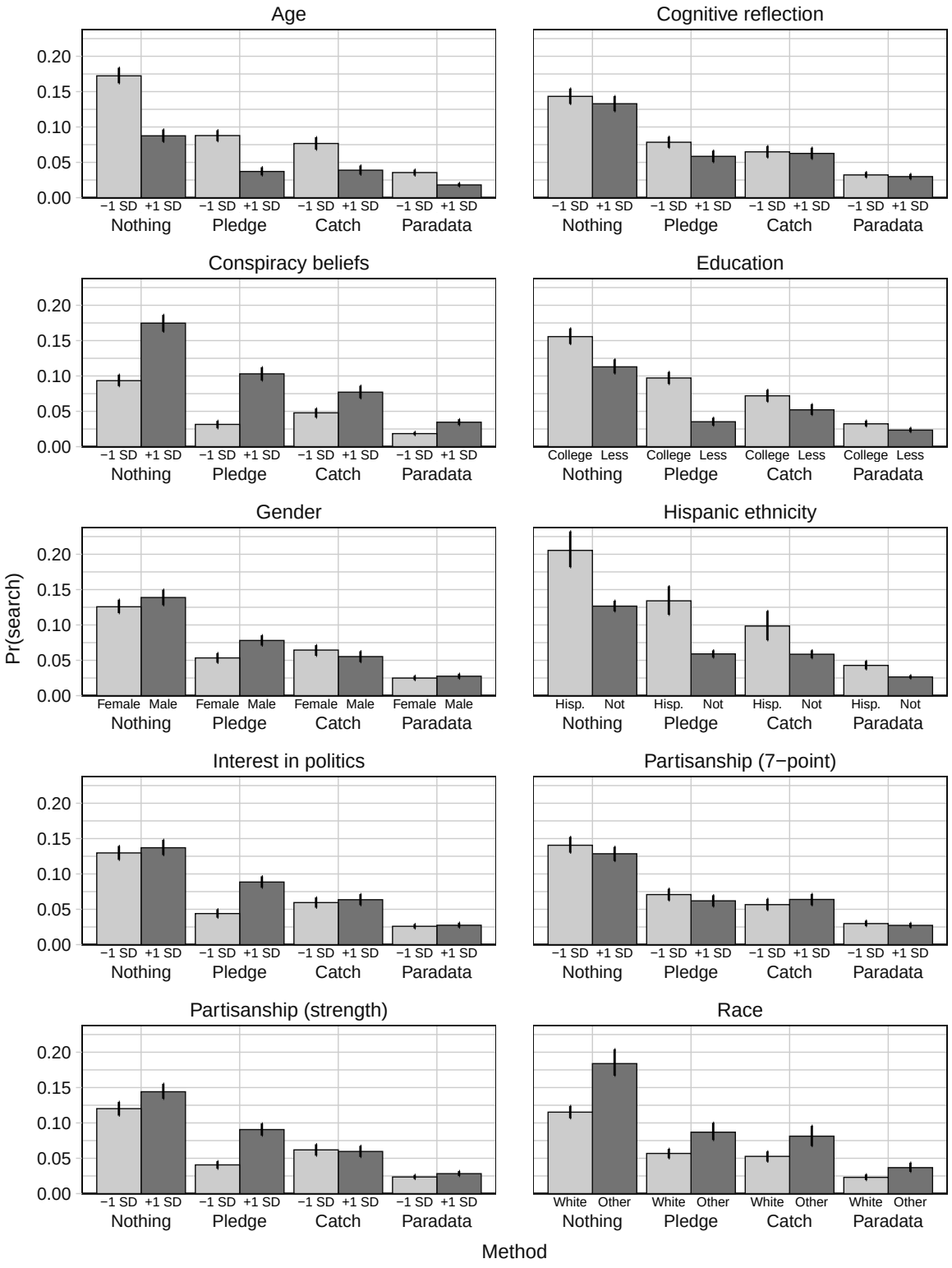
The detection methods (catch and paradata) have fairly even effects across groups. For every respondent characteristic examined, the difference in the prevalence of information search among unflagged respondents is smaller than the baseline difference. This means that researchers who treat detected respondents as missing data succeed, at least to some extent, at reducing between-group differences in search in the remaining data. The flip side of this success is a cost in terms of sample composition. Groups that are more likely to search at baseline are disproportionately dropped in approaches that treat instances of suspected search as missing data.

The deterrence method, the pledge, has mixed success at reducing between-group differences in the prevalence of search. In particular, respondents who are more educated, more interested in politics, and stronger partisans are both more likely to search *and* less deterred by the pledge than their counterparts. College graduates are 4.3 p.p. more likely to search at baseline and 6.2 p.p. more likely with a pledge (difference = 1.9, s.e. = 0.9; see Appendix

[16] To compute these estimates, a prediction for each component of the bias correction formula was generated using OLS regression. These predicted values were plugged into the bias correction formula. For hypothesis testing, the entire procedure was repeated in every block bootstrap replicate.

[17] As all of this section's results are consistent across studies, the data are pooled for simplicity.

Figure 6: Estimated prevalence of information search by mitigation method and respondent characteristic.



Note: Figure displays the porportion of suspected searchers (y-axis) by mitigation method (x-axis) and demographic characteristic (fill color). Error bars represent 95 percent confidence intervals. The appendix includes a table of estimates with difference in means tests (Table A7, page A14).

Table A7, page A14). High-interest respondents are just 0.7 p.p. more likely at baseline but 4.5 p.p. more likely with a pledge (difference = 3.7, s.e. = 0.8). Stronger partisans are also 2.4 p.p. more likely at baseline and 5.0 p.p. more likely with a pledge (difference = 2.6, s.e. = 0.8). In two more cases (gender and cognitive reflection), statistically insignificant estimates also suggest a slightly wider gap among those assigned to the pledge (difference = 1.2 and 1.0, s.e. = 0.9 and 0.9). Success at shrinking between-group differences is found for the other five variables. Given its *ex ante* nature, it is inconvenient that deterrence appears to be less consistent than detection in shrinking between-group differences in search.

The finding that pledges increase differences between these three sets of groups stands out in the context of existing research on the search problem. For different reasons, the tendency for the interested, the educated, strong partisans, and men to score better on political knowledge batteries are all well-established in the literature. On the positive side, the tendency for these subgroups of respondents to resist the pledge is consistent with Style and Jerit's (2021) argument that cheating is self-deceptive. The educated, interested, and partisan are likely to have the most self-image at stake when answering a quiz question about politics. On the negative side, evaluations often treat a stronger relationship in the expected direction as a sign that a methord improves a scale's validity (Smith et al. 2020; Marquis 2021). This is not necessarily the case. At baseline, search may either reinforce or attenuate between-group differences in knowledge, and deterrence methods may either counteract or reinforce the baseline difference. This complicates assessments of how search affects construct validity, which often assume that search can only work against expected relationships (e.g., Marquis 2021, 91-92).

A more detailed look at the results suggests no modification to the previous section's conclusions regarding the costs and benefits of combining measures. In conjunction with one another, the pledge and the paradata get most of the job done, getting all subgroups to 3 percent or less. Catch questions adding a bit of marginal value, getting this figure to 2 percent at a high cost in terms of missing data. The pledge's uneven performance at

reducing between-group differences persists in the presence of multiple measures.

Implications

Information search looms as a threat to the validity of any survey measure with answers that can easily be looked up. This paper shows that through a combination of detection and deterrence, researchers can manage the threat. As a starting point, question-level estimates of the prevalence of search are essential. The prevalence of search varies as a function of question content and response scales, and also appears to vary across survey platforms. Depending on the context, search may be a larger threat or no threat at all. Future research should estimate the prevalence of search using a wider range of questions and countries, as well as in higher-quality samples. Whereas the baseline level of information may depend on sample quality, extant research on the generalizability of treatment effects suggest that deterrence mechanisms are likely to be similarly effective across contexts (Mullinix et al. 2016; Coppock et al. 2018).

Deterrence is a researcher's first line of defense against information search. Its *ex ante* nature of avoids most of the downsides of *ex post* strategies for dealing with search, e.g. dropping suspected searchers as a robustness check. At the cost of one screen worth of survey space, the pledge tested here reduced search by 50 percent. Despite this, three shortcomings prevent it from fully solving the problem. First, it does not completely eliminate information search. Especially for questions with high base rates of search, a substantial amount of search still occurrs. Second, even as the pledge brings down the overall rate of search, it exacerbates differences in search between some subgroups. Third, on their own, deterrence methods like the pledge provide no information about these shortcomings.

Detection methods serve two purposes: to diagnose the prevalence of search and to provide the researcher with *ex post* options for dealing with it. Among existing methods, this paper considered the two that offer the best combination of cost and credibility: catch

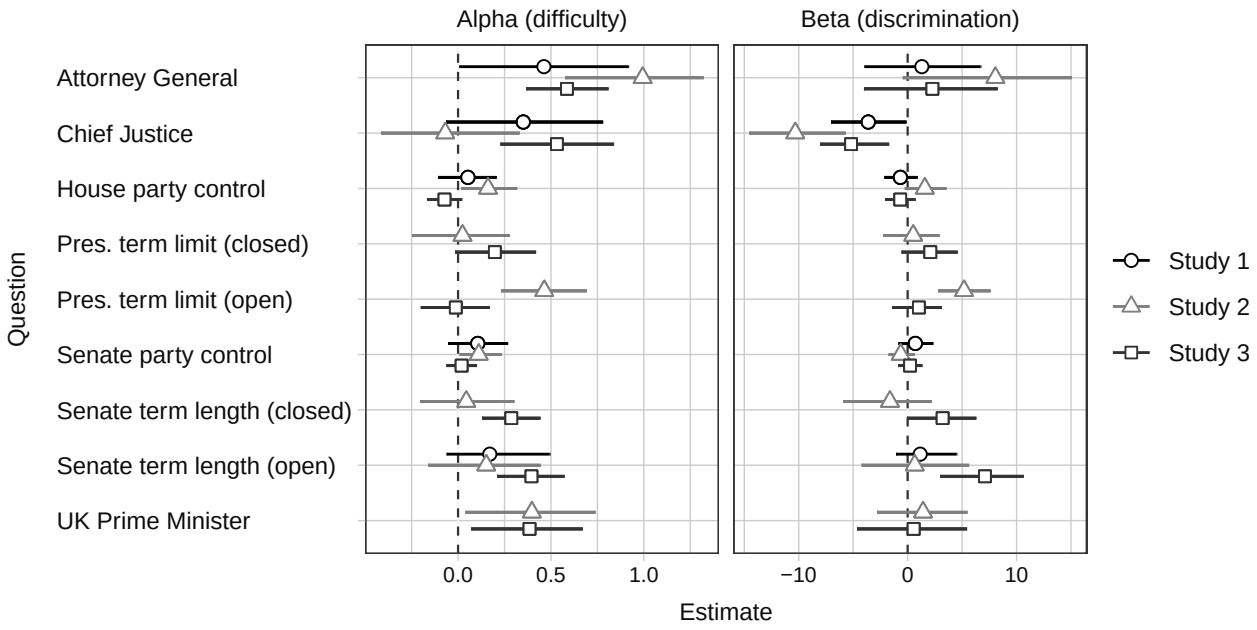
questions and paradata.[18] These two methods were compared in terms of their sensitivity (proportion of search detected), specificity (ability to avoid false positives), and the bottom line (the proportion of search in the unflagged data). Relative to pardata, catch questions were about as sensitive but less specific. Consequently, a higher proportion of the observations that go unflagged by the catch method are contaminated by search. The fundamental reasons for the paradata's superior performance are (1) it measures search at the item level rather than the individual level and (2) it does so for the knowledge questions themselves rather than using a separate question. Casting a wide net enables the catch method to detect a lot of search, but renders it an unreliable means of diagnosing the prevalence of search and a costly means of eliminating search *ex post*.

Evaluations of other measurement properties of knowledge scales also stand to benefit from item-level detection methods like this study's paradata method. Consider the assumption, common in validation studies, search works against expected relationships between variables. This paper showed that the baseline distribution of search may either reinforce or undercut expected differences in knowledge between groups, and that deterrence methods may either mitigate or exacerbate this source of confounding. Future evaluations of deterrence methods can address this complication by examining between-group differences in search and how these change in response to the intervention being tested.

To make these benefits more concrete, consider the effect of information search on the construction of knowledge scales. Though a full analysis is beyond the paper's scope, an exploratory test of one simple question was conducted: how does a pledge affect the difficulty and discrimination of knowledge items when they are combined into a scale using an IRT model? The results are presented in Figure 7. The most consistent evidence for an effect emerges in the two questions with the highest rates of search: the Attorney General question becomes more difficult and the Chief Justice question becomes less discriminating.

[18]Two other detection methods are discussed above. Self reports have low implementation costs but questionable sensitivity. Browsing histories are likely to be highly sensitive and specific but are costly in terms of money and sample representativeness.

Figure 7: Effect of pledge on IRT estimates.



Note: Figure displays the effect of the pledge on the two item-level parameters in an IRT model. Horizontal bars represent block bootstrapped 95 percent confidence intervals. A table of estimates appears in the appendix (Table A8, page A15).

Inconsistent results for the Senate and presidential term questions can be explained by differences between Studies 2 and 3 in the base rate of search and effect of the pledge (see Table 4). A set of results that might be written off as statistical noise or small effects instead looks like what one would expect if search matters most when it is most common. Such a conclusion can only be reached with the aid of question-level detection.

Despite its shortcomings, the catch method offers some value. When paradata are unavailable and the baseline prevalence of search is high, the catch method can eliminate search at a reasonably efficient rate. Moreover, a combination of two findings—that catch questions generate more search than knowledge questions (Figure 2), and that the pledge eliminated about the same proportion of search on both question types (Table 4)—suggests that catch questions are well-suited for use as "lab rats" for testing the relative efficacy of deterrence methods. Given their high base rate of search, treatments that are equally effective on a per-unit basis will result in larger effects that are easier to statistically distinguish from the control group and from one another.

In combination with one another, the complementary nature of detection and deterrence can allow researchers to conduct analysis in which they provide assurance, not hope, that information search has been eliminated from the data. Though this is an optimistic conclusion for the future of online surveys, it also raises the bar for analysis that claims to have addressed the search problem. Rather than asking whether the chosen methods *help reduce* search, researchers and audiences can begin to ask whether the chosen methods *successfully eliminate* search and whether they do so at a reasonable cost. The chief cost comes in the form of treating contaminated observations as missing data, which reduces statistical power and alters sample composition. Future research can strive to avoid these costs by paying attention to between-question variation in the rate of search and by seeking to identify more effective deterrence strategies.

This room for improvement notwithstanding, this paper's findings suggest that reasonable solutions to the information search problem are within our grasp and that even more complete solutions are within our reach. This is encouraging for online survey measures of knowledge and beliefs in all areas of research.

References

Ahler, Douglas J. and Guarav Sood. 2018. "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences." *The Journal of Politics* 80(3).

Ansolabehere, Stephen and Philip Edward Jones. 2010. "Constituents' responses to congressional roll-call voting." *American Journal of Political Science* 54(3):583–597.

Aronow, Peter M, Josh Kalla, Lilla Orr, John Ternovski and Broockman January May. 2020. "Evidence of Rising Rates of Inattentiveness on Lucid in 2020." *OSF Preprints* (unpublished manuscript).
URL: *https://osf.io/preprints/socarxiv/8sbe4/*

Bartels, Larry M. 2002. "Beyond the Running Tally: Partisan Bias in Political Perceptions." *Political Behavior* 24(2):117–150.

Berinsky, Adam J., Gregory A. Huber and Gabriel S. Lenz. 2012. "Evaluating online labor markets for experimental research: Amazon.com's mechanical turk." *Political Analysis* 20(3):351–368.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119(1):249–275.

Bryson, Bethany P. 2020. "When Survey Respondents Cheat: Internet Exposure and Ideological Consistency in the United States." *International Journal of Communication* 14:5351–5374.

Bullock, John G, Alan S Gerber, Seth J Hill and Gregory A Huber. 2015. "Partisan Bias in Factual Beliefs about Politics." *Quarterly Journal of Political Science* 10:1–60.

Clifford, Scott and Jennifer Jerit. 2014. "Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies." *Journal of Experimental Political Science* 1(2):120–131.

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(4):858–887.

Cohen, Cathy J. and Matthew D. Luttig. 2020. "Reconceptualizing Political Knowledge: Race, Ethnicity, and Carceral Violence." *Perspectives on Politics* 18(3):805–818.

Cooper, Emily A. and Hany Farid. 2016. "Does the Sun revolve around the Earth? A comparison between the general public and online survey respondents in basic scientific knowledge." *Public Understanding of Science* 25(2):146–153.

Coppock, Alexander, Thomas J. Leeper and Kevin J. Mullinix. 2018. "Generalizability of heterogeneous treatment effect estimates across samples." *Proceedings of the National Academy of Sciences* 115(49):12441–12446.

Diedenhofen, Birk and Jochen Musch. 2017. "PageFocus: Using paradata to detect and prevent cheating on online achievement tests." *Behavior Research Methods* 49(4):1444–1459.

Dolan, Kathleen. 2011. "Do Women and Men Know Different Things? Measuring Gender Differences in Political Knowledge." *The Journal of Politics* 73(1):97–107.
URL: *http://www.journals.uchicago.edu/doi/10.1017/S0022381610000897*

Domnich, Alexander, Donatella Panatto, Alessio Signori, Nicola Luigi Bragazzi, Maria Luisa Cristina, Daniela Amicizia and Roberto Gasparini. 2015. "Uncontrolled web-based administration of surveys on factual health-related knowledge: A randomized study of untimed versus timed quizzing." *Journal of Medical Internet Research* 17(4):e94.

Drummond, Caitlin and Baruch Fischhoff. 2017. "Individuals with greater science literacy and education have more polarized beliefs on controversial science topics." *Proceedings of the National Academy of Sciences of the United States of America* 114(36):9587–9592.

Gerber, Alan S. and Donald Green. 2012. *Field Experiments*. New York: W.W. Norton.

Gilens, Martin. 2001. "Political Ignorance and Collective Policy Preferences." *American Political Science Review* 95(2):379–396.

Gooch, Andrew and Lynn Vavreck. 2019. "How face-to-face interviews and cognitive skill affect item non-response: A randomized experiment assigning mode of interview." *Political Science Research and Methods* 7(1):143–162.

Graham, Matthew H. 2020. "Self-Awareness of Political Knowledge." *Political Behavior* 42:305–326.

Gummer, Tobias and Tanja Kunz. 2019. "Relying on External Information Sources When Answering Knowledge Questions in Web Surveys." *Sociological Methods and Research* pp. 1–21.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2015. "Causal Inference in Conjoint Analysis: Understanding Multi-Dimensional Choices via Stated Preference Experiments." *Poltical Analysis* 22(1):1–30.

Höhne, Jan Karem, Carina Cornesse, Stephan Schlosser, Mick P Couper and Annelies G Blom. 2021. "Looking up Answers to Political Knowledge Questions in Web Surveys." *Public Opinion Quarterly* 84(4):986–999.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2021. *An Introduction to Statistical Learning*. Springer.

Jensen, Carsten and Jens Peter Frølund Thomsen. 2014. "Self-reported cheating in web surveys on political knowledge." *Quality and Quantity* 48(6):3343–3354.

Krosnick, Jon A. 1991. "Response strategies for coping with the cognitive demands of attitude measures in surveys." *Applied Cognitive Psychology* 5(3):213–236.

Liu, Mingnan and Yichen Wang. 2014. "Data Collection Mode Effects On Political Knowledge." *Survey Methods: Insights from the Field* (December 12, 2014):1–12.

Lunz Trujillo, Kristin, Matthew Motta, Timothy Callaghan and Steven Sylvester. 2020. "Correcting Misperceptions about the MMR Vaccine: Using Psychological Risk Factors to Inform Targeted Communication Strategies." *Political Research Quarterly* .

Marquis, Lionel. 2021. "Using response times to enhance the reliability of political knowledge items: An application to the 2015 swiss post-election survey." *Survey Research Methods* 15(1):79–100.

Montgomery, Jacob M., Brendan Nyhan and Michelle Torres. 2018. "How Post-Treatment Bias Can Ruin Your Experiment and What to Do about It." *American Journal of Political Science* 62(3):760–775.

Motta, Matthew P., Timothy H. Callaghan and Brianna Smith. 2016. "Looking for Answers: Identifying Search Behavior and Improving Knowledge-Based Data Quality in Online Surveys." *International Journal of Public Opinion Research* 29(4):edw027.

Mullinix, Kevin J., Thomas J. Leeper, James N. Druckman and Jeremy Freese. 2016. "The generalizability of survey experiments." *Journal of Experimental Political Science* 2(2):109–138.

Pasek, Josh, Gaurav Sood and Jon A. Krosnick. 2015. "Misinformed About the Affordable Care Act? Leveraging Certainty to Assess the Prevalence of Misperceptions." *Journal of Communication* 65(4):660–673.

Pennycook, Gordon and David G. Rand. 2019. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188(September 2017):39–50.

Permut, Stephanie, Matthew Fisher and Daniel M. Oppenheimer. 2019. "TaskMaster: A Tool for Determining When Subjects Are on Task." *Advances in Methods and Practices in Psychological Science* 2(2):188–196.

Peters, Ellen, Daniel Vastfjall, Paul Slovic, C.K. Mertz, Ketti Mazzocco and Stephan Dickert. 2006. "Numeracy and Decision Making." *Psychological Science* 17(5):2006.

Pevnaya, Maria, Tatyana Bystrova and Elizaveta Pevnaya. 2019. "The Knowledge of City History as a Basis of Youth Engagement for Urban Sustainability." *European Conference on Management, Leadership & Governance* 1:315–330.

Peyton, Kyle, Gregory A Huber and Alexander Coppock. 2021. "The Generalizability of Online Experiments Conducted During The COVID-19 Pandemic." *Journal of Experimental Political Science* (forthcoming).
URL: *https://osf.io/preprints/socarxiv/s45yg/*

Rapeli, Lauri. 2022. Accuracy of Self-Assessments of Political Sophistication. In *Perspectives on Political Awareness: Conceptual, Theoretical and Methodological Issues*, edited by T. Denk, N. Norgaard Kristensen, M. Olson and T. Solhaug. Springer pp. 97–114.

Schwartz, Lisa M., Stephen Woloshin, William C. Black and H. Gilbert Welch. 1997. "The Role of Numeracy in Understanding the Benefit of Screening Mammography." *Annals of Internal Medicine* 127(11):966–72.

Shulman, Hillary C. and Franklin J. Boster. 2014. "Effect of Test-Taking Venue and Response Format on Political Knowledge Tests." *Communication Methods and Measures* 8(3):177–189.

Smith, Brianna, Scott Clifford and Jennifer Jerit. 2020. "How Internet Search Undermines the Validity of Political Knowledge Measures." *Political Research Quarterly* 73(1):141–155.

Starratt, Gerene K., Ivana Fredotovic, Sashay Goodletty and Christopher Starratt. 2017. "Holocaust knowledge and Holocaust education experiences predict citizenship values among US adults." *Journal of Moral Education* 46(2):177–94.

Strabac, Zan and Toril Aalberg. 2011. "Measuring political knowledge in telephone and web surveys: A cross-national comparison." *Social Science Computer Review* 29(2):175–192.

Style, Hillary and Jennifer Jerit. 2021. "Does it Matter if Respondents Look up Answers to Political Knowledge Questions?" *Public Opinion Quarterly* 84(3):760–775.

Vezzoni, Cristiano and Riccardo Ladini. 2017. "Thou shalt not cheat: How to reduce internet use in web surveys on political knowledge." *Rivista Italiana di Scienza Politica* 47(3):251–265.

Appendix to

Detecting and Deterring Information Search in Online Surveys

Contents

A Supplemental Results

A.1 Understanding the bias correction

The paper introduced an estimator for $P(\text{search})$, (2), that accounts for the bias introduced by undersensitivity and underspecificity in measures of suspected search. This section examines this bias correction's influence theoretically and empirically. In brief, it concludes that for methods like this paper's paradata method — with very high specificity and moderately high sensitivity — researchers generally underestimate search when they take the percentage of suspected search at face value. There is potential for overestimation, but only where there is very little search to begin with.

A.1.1 *Analytic results*

Researchers commonly use the probability of being flagged, $P(\text{flag})$, as an estimator for the probability of search, $P(\text{search})$. Abbreviate these as $P(f)$ and $P(s)$. The bias, i.e. the difference between the estimator and the true value, is

$$P(f) - P(s). \tag{7}$$

Plugging in (2) gives

$$P(f) - \frac{P(f) - P(f|\neg s)}{P(f|s) - P(f|\neg s)}, \tag{8}$$

which can usefully be rewritten as

$$\frac{P(f|\neg s)(1 - P(f)) - (1 - P(f|s))P(f)}{P(f|s) - P(f|\neg s)}. \tag{9}$$

This expression isolates the effect of underspecificity in the first term of the numerator and second term of the denominator, and undersensitivity's effect in the second term of the numerator and first term of the denominator.

First consider the denominator. It will be positive for any method of information search for which searchers are more likely to be flagged than non-searchers (i.e., $P(f|s) > P(f|\neg s)$). In order for this not to hold for the denominator to be negative, search and detection would have to be negatively correlated. Although it is possible to conjure rules that violate this condition (e.g., flag everyone who answers *in*correctly as a suspected searcher), users of any reasonable method can think of the denominator as being positive.[19]

Now consider the effect of underspecificity, which occurs when measures mistakenly flag some respondents who did not in fact search ($P(f|\neg s) > 0$). The first term in the numerator is

[19] Many absurd methods would also satisfy this condition, e.g. catch questions with two response options. It is a remarkably low bar.

always positive unless the measure is perfectly specific, in which case the first term equals zero. This indicates that underspecificity adds positive bias, exaggerating how much search occurred. This positive bias is magnified by the denominator, which includes the negative of $P(f|\neg s)$. The less specific the measure, the larger the number subtracted from the denominator (which, per the discussion just above, will remain positive for any reasonable detection method).

Next consider the effect of undersensitivity, which occurs when measures do not detect everyone who searches ($P(f|s) < 1$). The second term in the numerator is negative[20] unless the measure is perfectly sensitive, in which case it equals zero. This indicates that underspecificity adds negative bias, understating how much search occurred. This negative bias is magnified by the first term in the denominator, a positive number between 0 and 1. The less sensitive the measure, the smaller the denominator.

Finally consider the role that the numerator's two instances of $P(f)$ play in modulating the relative influence of specificity and sensitivity. When search is less common and fewer respondents are flagged, the influence of the underspecificity term (i.e., the first term) is greater. When search is common and more respondents are being flagged, the influence of the undersensitivity term (i.e., the first term) is greater. This means that the bias can change signs even when sensitivity and specificity are held constant across questions. When few respondents are being flagged, the bias is more likely to be positive. When more respondents are being flagged, the bias is more likely to be negative.

In the data, underspecificity tends to fall around $P(f|\neg s) = 0.01$ and undersensitivity tends to fall around $P(f|s) = 0.7$ to 0.85. This suggests that the negative bias from undersensitivity should generally be expected to be larger than the positive bias from underspecificity. However, due to the modulating effect of $P(f)$, bias may still be positive on questions with little search and correspondingly few flags.

A.1.2 *Comparison of estimates*

To demonstrate the bias correction's influence empirically, two steps are taken here.

First, for every question in the survey, corrected and uncorrected estimates of the prevalence of search and the effect of the pledge are plotted against one another (Figure A2). The x-axis displays the corrected estimate; the y-axis, the uncorrected estimate; and the dashed 45 degree line, the point at which the two estimates are equal. The results confirm that uncorrected estimates generally underestimate the prevalence of search, except at very low levels (Figure A2a). The uncorrected estimates also generally underestimate the effect of the pledge, in that they suggest the effect was less negative than the corrected estimates (Figure A2a).

Second, the two figures from the main text that report these same estimates are reproduced with the bias correction removed (Figure A3, Table A2).

[20] Equivalently, a term that is always positive is subtracted.

Figure A2: Comparison of bias-corrected and uncorrected estimates.

(a) Prevalence of search.

(b) Effect of pledge.

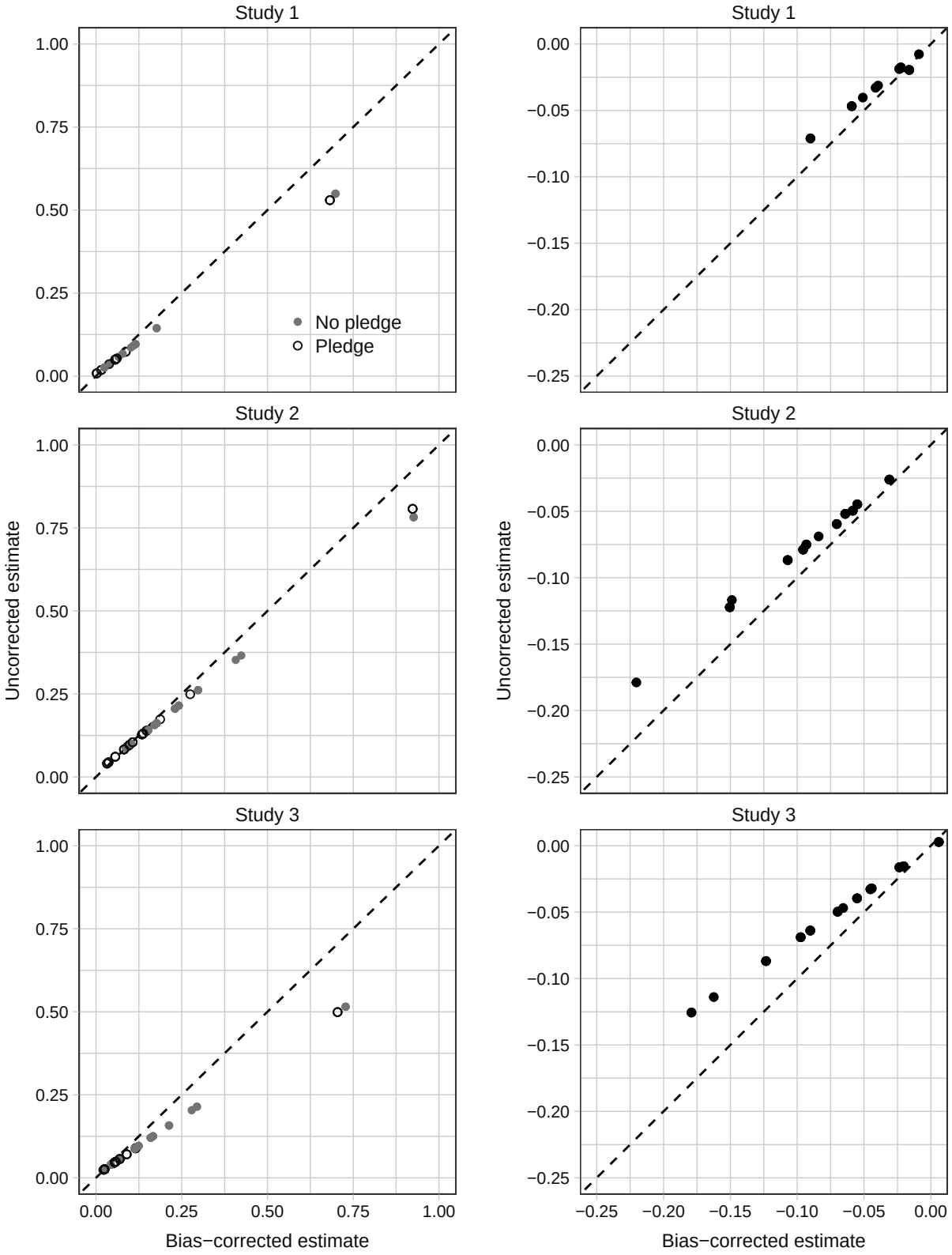
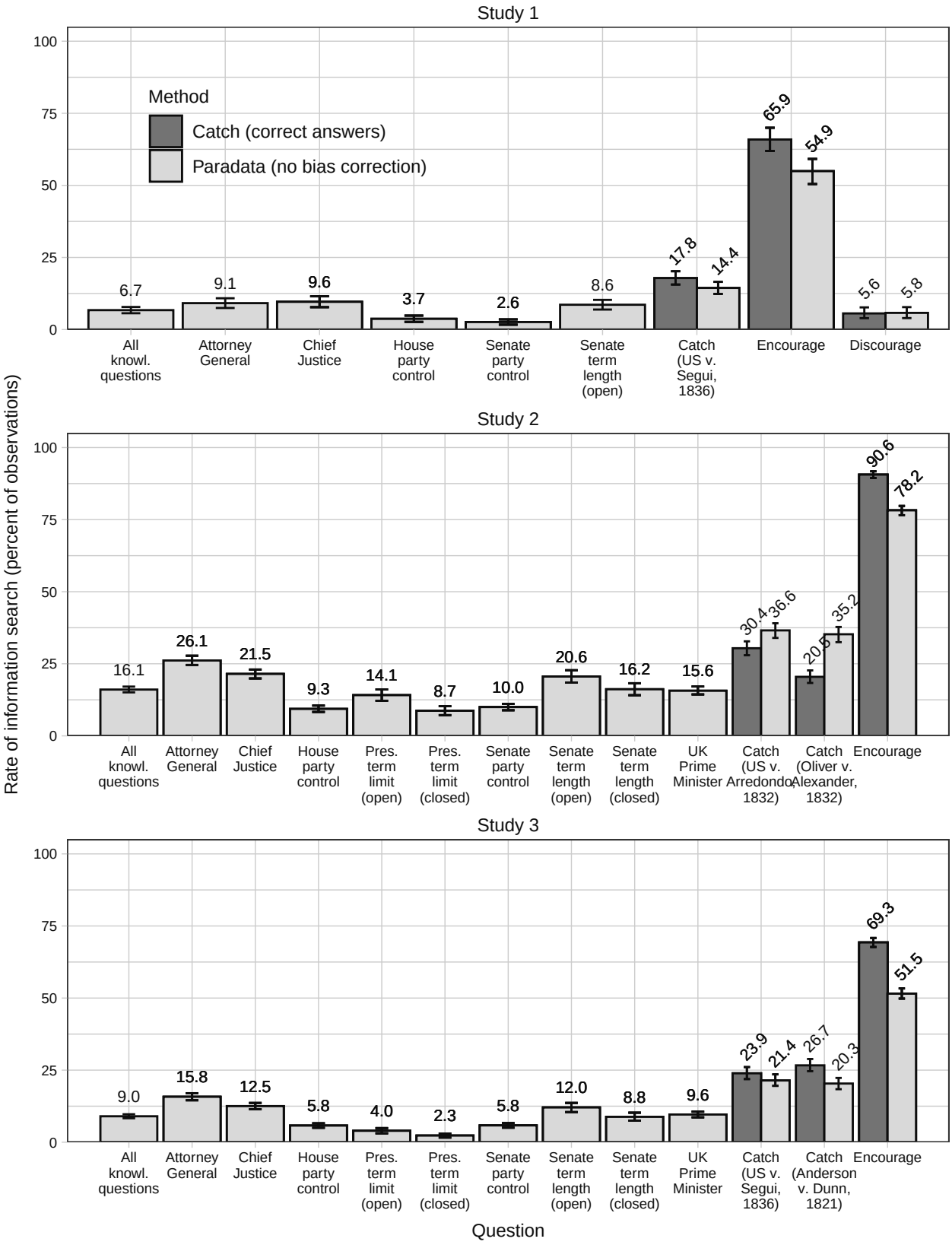


Figure A3: Estimated rate of information search, no bias correction (version of Figure 2).



Note: Figure is identical to Figure 2, but uses $P(\text{flag})$ to estimate search rather than (2).

Table A2: Deterrent effect of pledge, no bias correction (version of Table 4).

(a) Knowledge battery.

	Study 1				Study 2				Study 3			
Question	No pledge	Pledge	Effect	%	No pledge	Pledge	Effect	%	No pledge	Pledge	Effect	%
All knowl. questions	6.7 (0.6)	3.6 (0.4)	-3.1 (0.7)	-46.5	16.1 (0.5)	9.2 (0.4)	-6.9 (0.6)	-42.9	9.0 (0.4)	4.3 (0.2)	-4.7 (0.4)	-52.0
Attorney General	9.1 (0.9)	5.1 (0.7)	-4.0 (1.1)	-44.1	26.1 (0.8)	13.9 (0.7)	-12.2 (1.0)	-46.7	15.8 (0.6)	7.1 (0.4)	-8.7 (0.8)	-55.1
Chief Justice	9.6 (0.9)	4.9 (0.7)	-4.7 (1.2)	-48.6	21.5 (0.8)	12.8 (0.7)	-8.7 (1.0)	-40.4	12.5 (0.6)	5.6 (0.4)	-6.9 (0.7)	-55.0
House party control	3.7 (0.6)	1.8 (0.4)	-1.9 (0.7)	-50.5	9.3 (0.6)	4.4 (0.4)	-5.0 (0.7)	-53.0	5.8 (0.4)	2.6 (0.3)	-3.2 (0.5)	-55.6
Pres. term limit (open)					14.1 (1.0)	9.6 (0.8)	-4.5 (1.3)	-31.7	4.0 (0.5)	2.4 (0.4)	-1.5 (0.6)	-39.0
Pres. term limit (closed)					8.7 (0.8)	6.1 (0.7)	-2.6 (1.0)	-30.1	2.3 (0.4)	2.6 (0.4)	0.3 (0.6)	12.4
Senate party control	2.6 (0.5)	0.8 (0.3)	-1.8 (0.6)	-68.2	10.0 (0.6)	4.0 (0.4)	-6.0 (0.7)	-59.7	5.8 (0.4)	2.6 (0.3)	-3.3 (0.5)	-56.1
Senate term length (open)	8.6 (0.9)	5.3 (0.7)	-3.3 (1.1)	-38.3	20.6 (1.1)	13.1 (0.9)	-7.5 (1.4)	-36.5	12.0 (0.8)	5.7 (0.5)	-6.4 (1.0)	-53.0
Senate term length (closed)					16.2 (1.0)	8.3 (0.7)	-7.9 (1.3)	-48.8	8.8 (0.7)	4.8 (0.5)	-4.0 (0.9)	-45.0
UK Prime Minister					15.6 (0.7)	10.4 (0.6)	-5.2 (0.9)	-33.2	9.6 (0.5)	4.6 (0.4)	-5.0 (0.6)	-51.7

(b) Catch questions.

		Paradata detection				Catch (correct answers)			
Study	Question	No pledge	Pledge	Effect	%	No pledge	Pledge	Effect	%
Study 1	US v. Segui, 1836	14.4 (1.1)	7.3 (0.8)	-7.1 (1.3)	-49.3	17.8 (1.2)	8.5 (0.8)	-9.3 (1.4)	-52.3
Study 2	US v. Arredondo, 1832	36.6 (1.3)	24.9 (1.1)	-11.7 (1.7)	-31.9	30.4 (1.3)	20.0 (1.0)	-10.4 (1.6)	-34.3
	Oliver v. Alexander, 1832	35.2 (1.4)	17.4 (1.1)	-17.9 (1.7)	-50.7	20.5 (1.1)	9.1 (0.8)	-11.4 (1.4)	-55.5
Study 3	US v. Segui, 1836	21.4 (1.0)	8.9 (0.7)	-12.6 (1.2)	-58.6	23.9 (1.1)	9.0 (0.7)	-14.9 (1.3)	-62.3
	Anderson v. Dunn, 1821	20.3 (1.0)	9.0 (0.7)	-11.4 (1.2)	-56.0	26.7 (1.1)	10.1 (0.7)	-16.6 (1.3)	-62.2

Note: Table is identical to Table 4, but uses $P(\text{flag})$ to estimate search rather than (2).

A.2 Catch questions

A.2.1 *Distribution of responses*

The main text explains that through attention to the empirical distribution of responses to catch questions, researchers can reduce the expected rate of lucky guesses to a rate far below what would be expected if responses were uniformly distributed. In particular, catch questions about Supreme Court cases decided in the early or mid-1800s, in years not ending in 0 or 5, are likely to yield an extremely low rate of correct guessing.

To demonstrate this, Figures A4 and A5 plot the distributions of responses for each catch question, including those asked as part of the pay-to-search task. A thin horizontal line running across each figure illustrates what would be expected if responses were uniformly distributed. For all questions, incorrect answers are heavily concentrated in the past few decades and drop off significantly in years earlier than 1950. There is consistently a small spike around 1800 and 1900, suggesting that catch questions decided close to the turn of a century are a poor choice.

A.2.2 *Evidence of internal undersensitivity in Study 2*

The main text notes that in Study 2, correct answers were an undersensitive test for search on the catch questions because both questions had plausible, incorrect answers that the respondent could have found in search results.

To show this, Tables A3 through A5 display the top 10 responses to each catch question, separated by whether the respondent was flagged by the paradata method. In Studies 1 and 3, which used different catch questions, only the correct answers are more common among those who trigger the paradata flag than among those who do not.

By contrast, in Study 2, several incorrect responses are more likely among those who are flagged by the paradata method. These years can be found in search results for the cases; for example, 6.6 percent of respondents answered 1982 to the Oliver v. Alexander question, which is the date of a district court case titled Oliver v. Alexander County Housing Authority. There is also some circumstantial evidence that faced with the frustration of multiple plausible correct answers, respondents gave up and guessed something else. In particular, the answer 1900 is more among paradata-flagged respondents in Study 2 than it is among respondents who were not flagged. By contrast, in Study 3, the spike at 1900 is concentrated among respondents who were not flagged by the paradata method.

Figure A4: Distribution of responses to catch questions.

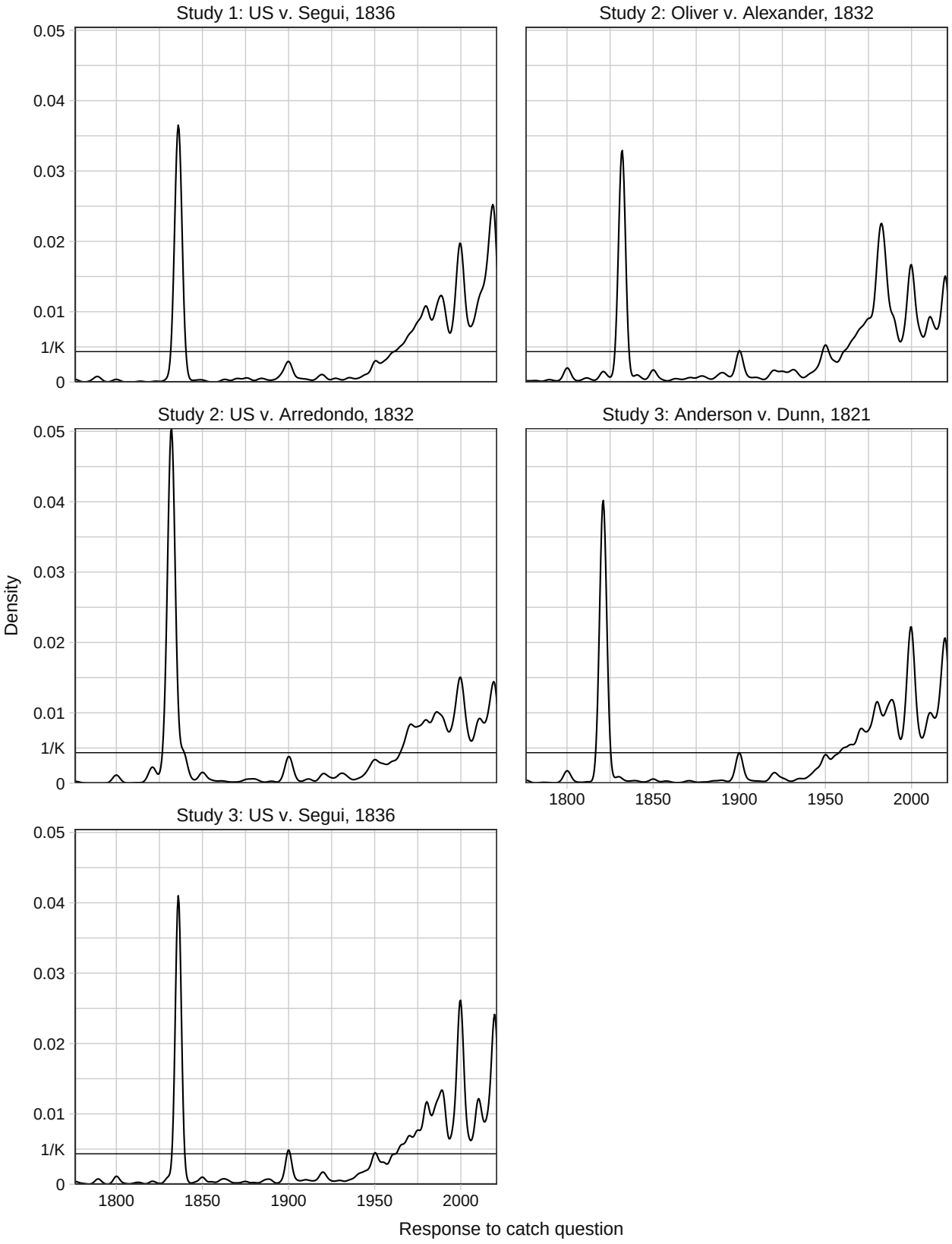


Figure A5: Distribution of responses to pay-to-search questions.

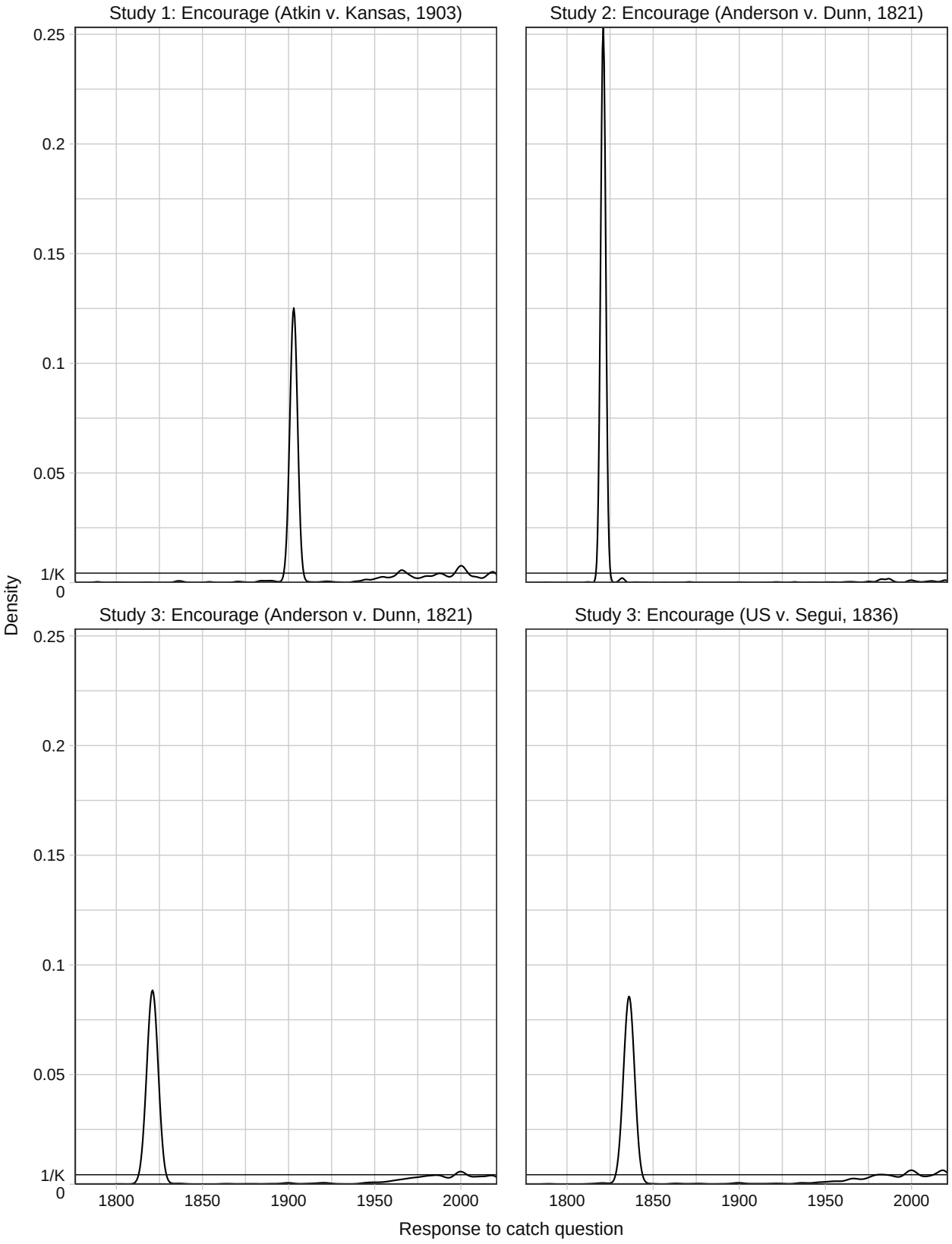


Table A3: Top 10 responses to catch questions, Study 1.

(a) US v. Segui, 1836.

	No flag		Flagged		
Response	N	%	N	%	Ratio
1836	75	10.29	116	15.91	1.55
1974	7	0.96	3	0.41	0.43
2019	30	4.12	3	0.41	0.10
1950	10	1.37	2	0.27	0.20
1970	5	0.69	2	0.27	0.40
1980	16	2.19	2	0.27	0.12
1984	6	0.82	2	0.27	0.33
1789	1	0.14	1	0.14	1.00
1814	0	0.00	1	0.14	Inf
1840	1	0.14	1	0.14	1.00

Table A4: Top 10 responses to catch questions, Study 2.

(a) US v. Arredondo, 1832.

	No flag		Flagged		
Response	N	%	N	%	Ratio
1832	123	8.88	298	21.52	2.42
1839	11	0.79	26	1.88	2.36
1900	7	0.51	13	0.94	1.86
1971	3	0.22	11	0.79	3.67
1828	7	0.51	10	0.72	1.43
1850	2	0.14	10	0.72	5.00
1837	5	0.36	7	0.51	1.40
1970	16	1.16	7	0.51	0.44
2000	33	2.38	7	0.51	0.21
2020	50	3.61	7	0.51	0.14

(b) Oliver v. Alexander, 1832.

	No flag		Flagged		
Response	N	%	N	%	Ratio
1832	67	5.36	189	15.12	2.82
1982	15	1.20	67	5.36	4.47
1984	26	2.08	27	2.16	1.04
1900	8	0.64	14	1.12	1.75
2020	34	2.72	14	1.12	0.41
1830	2	0.16	7	0.56	3.50
1980	31	2.48	6	0.48	0.19
2000	39	3.12	6	0.48	0.15
1800	5	0.40	4	0.32	0.80
1950	13	1.04	4	0.32	0.31

Table A5: Top 10 responses to catch questions, Study 3.

(a) US v. Segui, 1836.

	No flag		Flagged		
Response	N	%	N	%	Ratio
1836	137	8.46	250	15.44	1.82
1900	20	1.24	9	0.56	0.45
2019	41	2.53	7	0.43	0.17
2000	84	5.19	4	0.25	0.05
2020	81	5.00	4	0.25	0.05
1800	6	0.37	3	0.19	0.50
1850	5	0.31	3	0.19	0.60
1983	2	0.12	3	0.19	1.50
1999	39	2.41	3	0.19	0.08
2008	8	0.49	3	0.19	0.38

(b) Anderson v. Dunn, 1821.

	No flag		Flagged		
Response	N	%	N	%	Ratio
1821	173	10.27	276	16.39	1.60
1800	8	0.48	7	0.42	0.88
1900	18	1.07	7	0.42	0.39
1820	6	0.36	5	0.30	0.83
1830	2	0.12	3	0.18	1.50
1987	14	0.83	3	0.18	0.21
2020	78	4.63	3	0.18	0.04
1819	0	0.00	2	0.12	Inf
1822	1	0.06	2	0.12	2.00
1824	2	0.12	2	0.12	1.00

A.2.3 *Estimating internal underspecificity*

To estimate $P(\text{flag}|\neg\text{search})$ for catch questions, the following formula was used:

$$P(\text{not a multiple of 5}|\text{incorrect}) \times \hat{f}(\text{correct answer}|\text{not a multiple of 5, incorrect})$$

where $\hat{f}$ is an empirical estimate of the p.d.f. calculated using local linear regression. Verbally, this formula uses the distribution of incorrect answers to predict how likely a correct answer would be to occur by chance. The distribution is estimated excluding multiples of 5, then adjusted downward to correct for the exclusion of these values. The table below presents the estimates from this formula. The average of these estimates is reported in the main text.

Table A6: Estimates of underspecificity for catch and pay-to-look questions.

Study	Question	Estimate
Study 1	Catch (US v. Segui, 1836)	0.0001169
	Encourage (Atkin v. Kansas, 1903)	0.0010994
Study 2	Catch (Oliver v. Alexander, 1832)	0.0005115
	Catch (US v. Arredondo, 1832)	0.0023443
	Encourage (Anderson v. Dunn, 1821)	0.0034239
Study 3	Catch (Anderson v. Dunn, 1821)	0.0004258
	Catch (US v. Segui, 1836)	0.0002774
	Encourage (Anderson v. Dunn, 1821)	0.0003382
	Encourage (US v. Segui, 1836)	0.0003972

A.3 Marginal costs and benefits of combining measures

The main text notes that the marginal costs and benefits of using detection measures to remove suspected instances of search from the data depends on the base rate of search. Figures A6 through A8 illustrate this. Each figure presents the same information as Figure **??** in the main text. Across all studies, the catch method offers the least marginal benefit on the House and Senate party control questions in Studies 1 and 2; Study 2 is the basis for the main text's claim that the marginal benefit sometimes falls below one unit of search eliminated for every 50 units of missing data. The marginal benefit is greatest for the Attorney General question in Study 3, which is analyzed in the main text.

Figure A6: Marginal costs and benefits of deterrence and detection, Study 1.

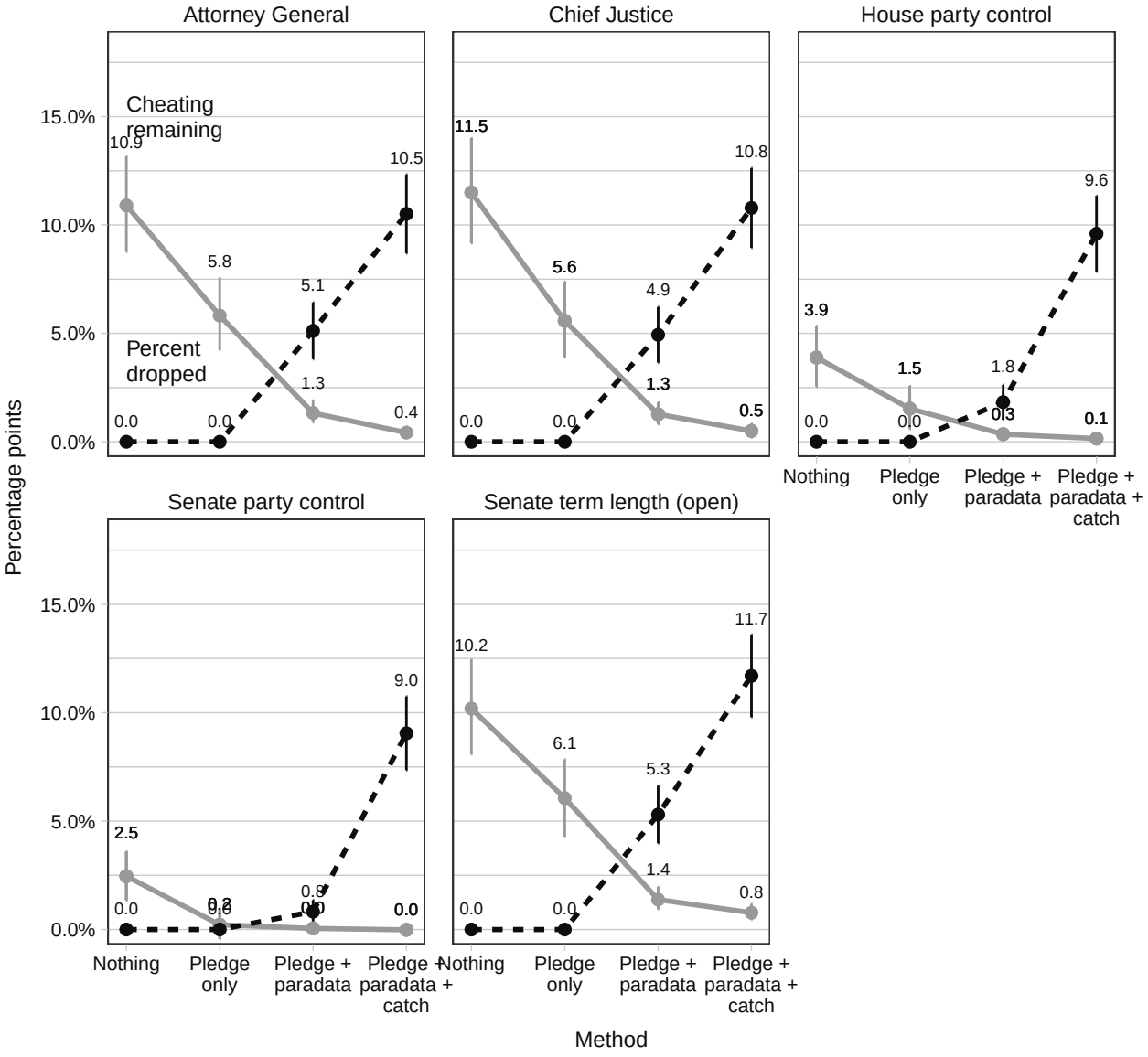


Figure A7: Marginal costs and benefits of deterrence and detection, Study 2.

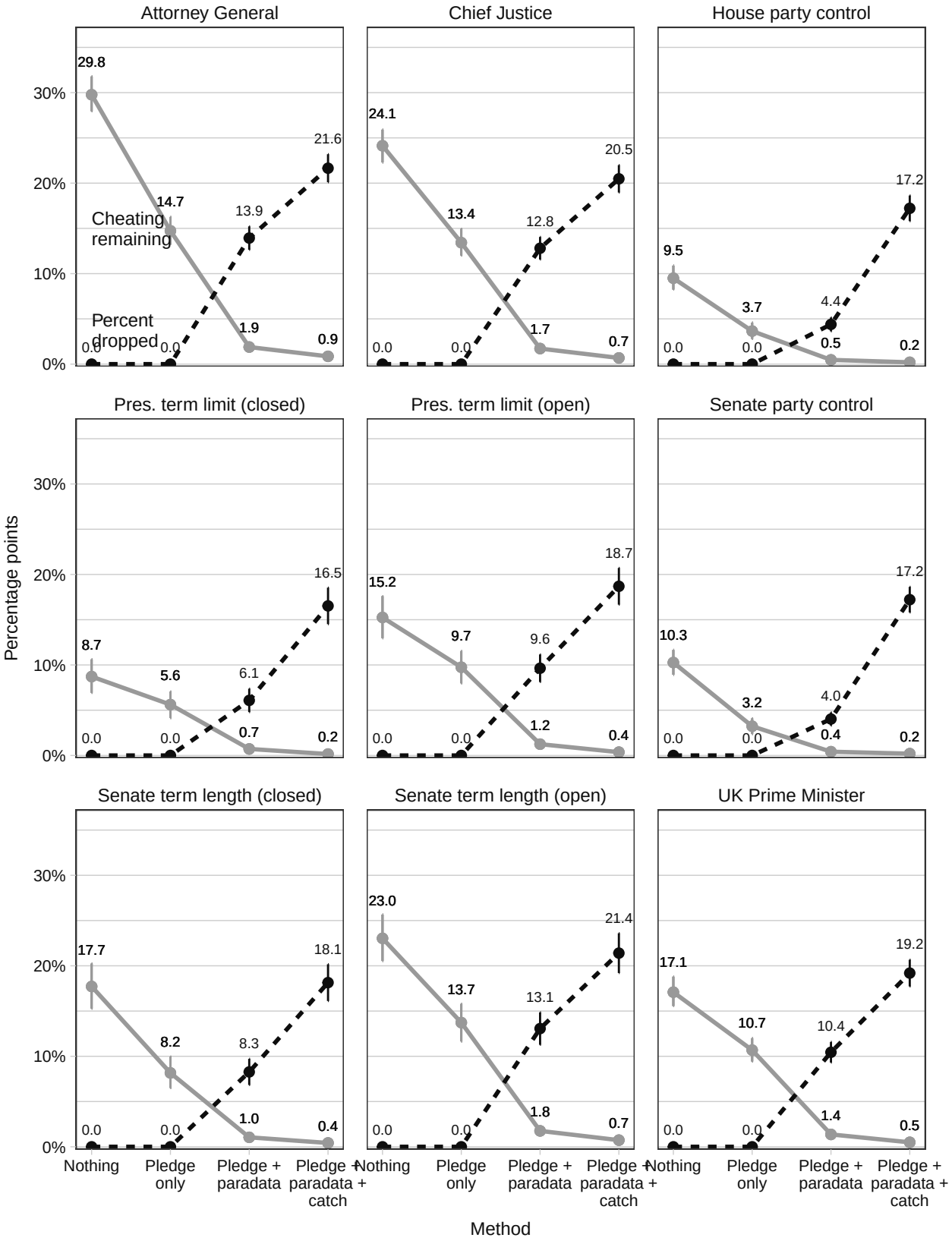
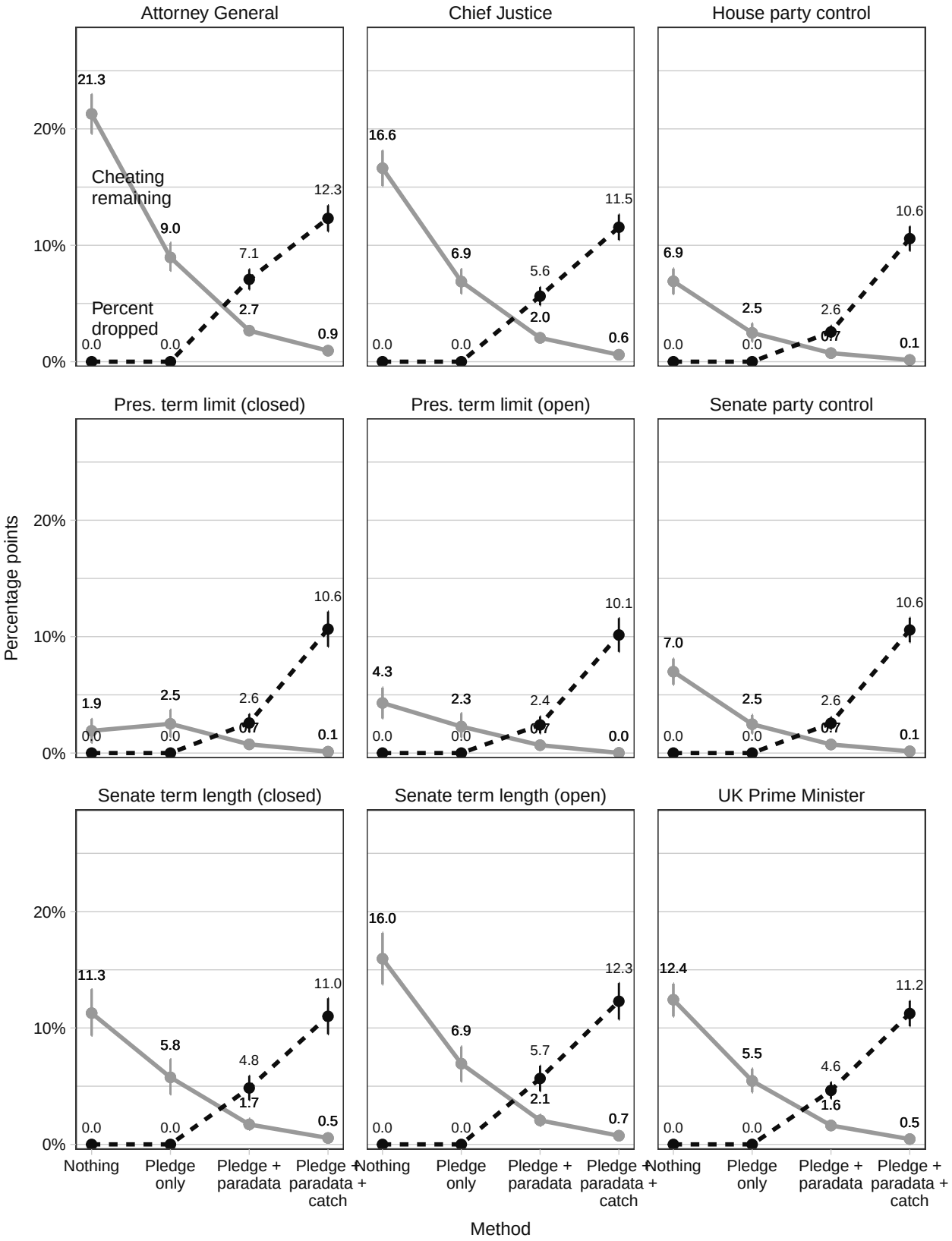


Figure A8: Marginal costs and benefits of deterrence and detection, Study 3.



A.4 Heterogeneous effects

Table A7: Estimates plotted in Figure 6 with tests for differences in proportions.

		Nothing	Pledge		Catch		Paradata	
Variable	Value	Prop.	Prop.	Diff.	Prop.	Diff.	Prop.	Diff.
Age	+1 SD	0.088	0.037	-0.050	0.039	-0.049	0.018	-0.069
		(0.004)	(0.003)	(0.005)	(0.003)	(0.003)	(0.001)	(0.003)
	-1 SD	0.172	0.088	-0.084	0.077	-0.096	0.036	-0.137
		(0.006)	(0.004)	(0.007)	(0.005)	(0.005)	(0.002)	(0.004)
	Diff	-0.085	-0.051	0.034	-0.038	0.047	-0.018	0.067
		(0.007)	(0.005)	(0.008)	(0.005)	(0.006)	(0.002)	(0.005)
Cognitive reflection	+1 SD	0.133	0.058	-0.074	0.063	-0.070	0.030	-0.103
		(0.005)	(0.004)	(0.006)	(0.004)	(0.004)	(0.002)	(0.004)
	-1 SD	0.143	0.078	-0.065	0.065	-0.078	0.032	-0.111
		(0.006)	(0.004)	(0.007)	(0.004)	(0.005)	(0.002)	(0.004)
	Diff	-0.010	-0.020	-0.010	-0.002	0.008	-0.002	0.008
		(0.007)	(0.006)	(0.009)	(0.006)	(0.006)	(0.002)	(0.006)
Conspiracy beliefs	+1 SD	0.175	0.103	-0.072	0.077	-0.098	0.035	-0.140
		(0.006)	(0.005)	(0.007)	(0.004)	(0.005)	(0.002)	(0.005)
	-1 SD	0.093	0.032	-0.062	0.048	-0.045	0.019	-0.075
		(0.004)	(0.003)	(0.005)	(0.003)	(0.003)	(0.001)	(0.003)
	Diff	0.081	0.071	-0.010	0.029	-0.052	0.016	-0.065
		(0.007)	(0.006)	(0.009)	(0.005)	(0.006)	(0.002)	(0.006)
Education	Less	0.113	0.035	-0.078	0.052	-0.061	0.023	-0.089
		(0.005)	(0.003)	(0.006)	(0.004)	(0.004)	(0.002)	(0.004)
	College	0.156	0.097	-0.059	0.072	-0.084	0.032	-0.123
		(0.006)	(0.004)	(0.007)	(0.004)	(0.005)	(0.002)	(0.004)
	Diff	-0.043	-0.062	-0.019	-0.020	0.023	-0.009	0.034
		(0.007)	(0.005)	(0.009)	(0.006)	(0.006)	(0.002)	(0.006)
Gender	Male	0.139	0.078	-0.061	0.055	-0.084	0.027	-0.111
		(0.006)	(0.004)	(0.007)	(0.004)	(0.005)	(0.002)	(0.004)
	Female	0.126	0.053	-0.072	0.064	-0.061	0.025	-0.101
		(0.005)	(0.004)	(0.006)	(0.004)	(0.004)	(0.002)	(0.004)
	Diff	0.013	0.025	0.012	-0.009	-0.022	0.003	-0.010
		(0.007)	(0.005)	(0.009)	(0.005)	(0.006)	(0.001)	(0.006)
Hispanic ethnicity	Hisp.	0.205	0.134	-0.071	0.098	-0.107	0.043	-0.163
		(0.013)	(0.010)	(0.016)	(0.011)	(0.011)	(0.003)	(0.010)
	Not	0.127	0.059	-0.068	0.059	-0.068	0.026	-0.100
		(0.004)	(0.003)	(0.005)	(0.003)	(0.003)	(0.001)	(0.003)
	Diff	-0.079	-0.075	0.004	-0.040	0.039	-0.016	0.062
		(0.013)	(0.011)	(0.017)	(0.011)	(0.012)	(0.003)	(0.011)
Interest politics	+1 SD	0.137	0.089	-0.049	0.064	-0.074	0.027	-0.110
		(0.005)	(0.004)	(0.007)	(0.004)	(0.004)	(0.002)	(0.004)
	-1 SD	0.130	0.044	-0.086	0.059	-0.070	0.026	-0.104
		(0.005)	(0.003)	(0.006)	(0.004)	(0.004)	(0.002)	(0.004)
	Diff	0.007	0.045	0.037	0.004	-0.003	0.001	-0.006
		(0.007)	(0.005)	(0.008)	(0.005)	(0.006)	(0.001)	(0.006)
Partisanship (7-point)	+1 SD	0.129	0.062	-0.067	0.064	-0.065	0.027	-0.101
		(0.005)	(0.004)	(0.006)	(0.004)	(0.004)	(0.002)	(0.004)
	-1 SD	0.141	0.071	-0.070	0.057	-0.084	0.030	-0.111
		(0.006)	(0.004)	(0.007)	(0.004)	(0.005)	(0.002)	(0.004)
	Diff	-0.012	-0.009	0.003	0.007	0.019	-0.003	0.009
		(0.007)	(0.006)	(0.010)	(0.006)	(0.006)	(0.002)	(0.006)
Partisanship (strength)	+1 SD	0.144	0.091	-0.053	0.060	-0.084	0.028	-0.116
		(0.006)	(0.004)	(0.007)	(0.004)	(0.005)	(0.002)	(0.004)
	-1 SD	0.120	0.041	-0.080	0.062	-0.058	0.024	-0.097
		(0.005)	(0.003)	(0.005)	(0.004)	(0.004)	(0.001)	(0.004)
	Diff	0.024	0.050	0.026	-0.002	-0.026	0.005	-0.019
		(0.007)	(0.005)	(0.008)	(0.006)	(0.006)	(0.001)	(0.006)
Race	White	0.115	0.057	-0.058	0.053	-0.063	0.023	-0.092
		(0.004)	(0.003)	(0.005)	(0.004)	(0.003)	(0.002)	(0.003)
	Other	0.184	0.087	-0.097	0.081	-0.103	0.037	-0.147
		(0.009)	(0.006)	(0.011)	(0.007)	(0.008)	(0.003)	(0.007)
	Diff	0.069	0.030	-0.039	0.029	-0.040	0.014	-0.055
		(0.010)	(0.007)	(0.012)	(0.008)	(0.008)	(0.002)	(0.008)

Note: Column headers describe detection and deterrence methods. Below them, "Prop." columns display the proportion of subjects searching and "Diff." columns display the difference in proportions between each intervention and no intervention (i.e., the baseline rate of cheating). Each trio of rows gives the estimated proportion of subjects cheating for two values of each variable, as well as the difference between them. Block bootstrapped standard errors in parentheses.

A.5 Effect on IRT estimates

Table A8: Effect of pledge on IRT estimates.

		Study 1			Study 2			Study 3		
Question	Parameter	No pledge	Pledge	Diff.	No pledge	Pledge	Diff.	No pledge	Pledge	Diff.
Attorney General	α	1.84	1.38	0.46*	1.54	0.55	0.99*	0.76	0.17	0.59*
		(0.17)	(0.15)	(0.23)	(0.13)	(0.13)	(0.18)	(0.09)	(0.07)	(0.11)
	β	20.45	19.15	1.30	25.40	17.35	8.05	22.97	20.70	2.27
		(1.75)	(2.05)	(2.73)	(2.29)	(3.26)	(3.99)	(2.14)	(2.17)	(3.05)
Chief Justice	α	2.17	1.82	0.35	2.52	2.58	-0.07	2.36	1.83	0.53*
		(0.18)	(0.11)	(0.21)	(0.15)	(0.11)	(0.19)	(0.09)	(0.13)	(0.15)
	β	23.45	27.07	-3.62*	30.27	40.59	-10.32*	41.74	46.93	-5.19*
		(1.28)	(1.16)	(1.73)	(1.82)	(1.17)	(2.23)	(1.13)	(1.08)	(1.61)
House party control	α	0.71	0.66	0.05	1.12	0.96	0.16*	0.82	0.89	-0.07
		(0.05)	(0.06)	(0.08)	(0.06)	(0.04)	(0.07)	(0.03)	(0.04)	(0.05)
	β	4.59	5.26	-0.67	7.57	6.00	1.58	5.55	6.23	-0.68
		(0.45)	(0.58)	(0.74)	(0.75)	(0.53)	(0.93)	(0.44)	(0.55)	(0.69)
Pres. term limit (closed)	α				1.63	1.61	0.02	1.74	1.55	0.20
					(0.09)	(0.09)	(0.13)	(0.08)	(0.07)	(0.11)
	β				7.02	6.51	0.51	7.32	5.24	2.08
					(0.97)	(0.85)	(1.32)	(0.97)	(0.87)	(1.29)
Pres. term limit (open)	α				1.75	1.29	0.46*	1.52	1.53	-0.01
					(0.10)	(0.06)	(0.12)	(0.07)	(0.07)	(0.09)
	β				9.65	4.47	5.18*	6.54	5.51	1.03
					(1.11)	(0.57)	(1.25)	(0.75)	(0.91)	(1.16)
Senate party control	α	0.83	0.73	0.11	-0.64	-0.75	0.11*	-0.60	-0.62	0.02
		(0.06)	(0.06)	(0.08)	(0.04)	(0.04)	(0.06)	(0.03)	(0.03)	(0.04)
	β	5.98	5.27	0.71	7.78	8.41	-0.63	6.19	5.97	0.22
		(0.53)	(0.59)	(0.79)	(0.41)	(0.43)	(0.59)	(0.36)	(0.40)	(0.55)
Senate term length (closed)	α				0.71	0.66	0.04	0.30	0.02	0.29*
					(0.08)	(0.09)	(0.12)	(0.06)	(0.05)	(0.08)
	β				8.61	10.24	-1.63	13.27	10.06	3.21
					(1.18)	(1.60)	(2.06)	(1.19)	(0.99)	(1.59)
Senate term length (open)	α	0.42	0.25	0.17	0.99	0.84	0.15	0.38	-0.02	0.40*
		(0.12)	(0.08)	(0.14)	(0.10)	(0.11)	(0.15)	(0.07)	(0.05)	(0.09)
	β	11.59	10.44	1.15	15.25	14.60	0.65	19.15	12.05	7.10*
		(1.01)	(0.92)	(1.37)	(1.42)	(1.87)	(2.40)	(1.41)	(1.27)	(1.96)
UK Prime Minister	α				1.88	1.49	0.40*	1.67	1.29	0.38*
					(0.12)	(0.13)	(0.18)	(0.10)	(0.11)	(0.15)
	β				12.99	11.57	1.41	20.25	19.70	0.55
					(1.37)	(1.63)	(2.13)	(1.57)	(1.99)	(2.54)

Note: Cell entries are estimates of the item-level parameters from an IRT model. Block bootstrapped standard errors in parentheses. $*p < 0.05$, two-tailed. Estimates from the "Diff." columns are displayed in the main text (Figure 7).

A.6 Political knowledge in the literature

To understand the breadth of political knowledge's applicability within political science and the social sciences more broadly, two complementary analyses were conducted.

First, the 100 most recent articles that included a survey-based measure of political knowledge were identified using Google Scholar. The author examined each article to determine whether it actually used a knowledge scale, the country to which the article's data pertained, and basic information about the article (title, author, journal, DOI). The distribution of countries and journals is summarized in Table A9. The underlying table appears in the online replication file. The table is current as of April 8, 2022.

Table A9: Summary of 100 most recently published articles containing a measure of political knowledge.

(a) Countries

	Country	N
1	United States	32
2	Multinational	12
3	Indonesia	7
4	Germany	3
5	Iran	3
6	Nertherlands	3
7	Switzerland	3
8	Australia	2
9	Belgium	2
10	China	2
11	Czech Republic	2
12	Denmark	2
13	Finland	2
14	Malaysia	2
15	Pakistan	2
16	Philippines	2
17	Poland	2
18	Spain	2
19	15 countries	1

(b) Journals

	Journal	N
1	International Journal of Public Opinion Research	4
2	Computers in Human Behavior	3
3	Political Communication	3
4	British Journal of Political Science	2
5	Current Psychology	2
6	Government and Opposition	2
7	Human Communication Research	2
8	New Media & Society	2
9	Political Behavior	2
10	Political Studies Review	2
11	Social Media + Society	2
12	Southern Communication Journal	2
13	The Hague Journal of Diplomacy	2
14	The Interational Journal of Press/Politics	2
15	68 journals	1

Note: Table displays all countries and journals with two or more articles among the 100 most recenly published.

Second, the top 15 political science journals were identified using Google's H5-index as of April 10, 2022. That same day, Google Scholar queries used to cout the total number of articles appearing in the journal, as well as the number of such articles containing the phrase "political knowledge." The former query was `source:''[journal name]''` and the latter appended `''political knowledge''`, with quotes as part of the queries. The 15-journal threshold was selected for convenience; whereas false positives were straightforward enough to eliminate for the top 15 journals, the 16th and 17th journals (*Governance* and *Political Studies*) returned large numbers of false positives due to the large number of journal names that include these phrases.

Table A10: Proportion of articles including phrase "political knowledge," top 15 political science journals.

	2010-Present			2020-Present		
Publication	Total	Knowl.	Pct.	Total	Knowl.	Pct.
American Journal of Political Science	1020	98	9.6	239	25	10.5
American Political Science Review	978	72	7.4	310	30	9.7
Journal of European Public Policy	1310	12	0.9	338	8	2.4
The Journal of Politics	1480	150	10.1	380	28	7.4
JCMS: Journal of Common Market Studies	1540	15	1.0	394	3	0.8
Comparative Political Studies	1050	51	4.9	218	13	6.0
British Journal of Political Science	875	80	9.1	321	34	10.6
Journal of Democracy	1450	11	0.8	222	0	0.0
European Journal of Political Research	973	66	6.8	270	25	9.3
West European Politics	1220	46	3.8	228	16	7.0
Annual Review of Political Science	370	33	8.9	72	7	9.7
Political Behavior	984	228	23.2	336	78	23.2
Party Politics	1800	61	3.4	408	27	6.6
Political Analysis	748	43	5.7	136	4	2.9
Democratization	3380	42	1.2	662	14	2.1

These analyses have complementary strengths and weaknesses. The first analysis assures that the article in question measures political knowledge rather than simply referring to it. The second is specific to top journals in political science. By extension, the first analysis is broader in its scope. It captures the extensive use of political knowledge measures in social science fields like communication, psychology, and sociology, as well as the use of such measures among international scholars whose valuable work is unlikely to be considered by the journals in the second analysis (e.g., due to a bias toward U.S.-focused analysis).

B Study Information

B.1 About the studies

Study 1

Platform: Lucid.

Dates: December 4-9, 2020.

Sample size: 2,176.

Compensation: $1, plus $2 for all respondents who completed an unrelated follow-up survey.

Screeners: Captcha verification, attention check.

Consent: Subjects read an IRB-approved consent form, then voluntarily consented to participate in a research study.

Refusal rate: 6.3 percent.

IRB approval: Yale University, #2000026693.

Study 2

Platform: Amazon Mechanical Turk.

Dates: April 19-May 7, 2021.

Sample size: 5,411.

Compensation: $0.75.

Screeners: Captcha verification.

Consent: Subjects read an IRB-approved consent form, then voluntarily consented to participate in a research study.

Refusal rate: 0.1 percent.

IRB approval: George Washington University, #NCR213434.

Preregistration: https://osf.io/k9d4r/?view_only=9649336cd3634f7a8c8239317c4b7683

Study 3

Platform: Lucid.

Dates: May 19-30, 2021.

Sample size: 6,687.

Compensation: $1.

Screeners: Captcha verification, attention check.

Consent: Subjects read an IRB-approved consent form, then voluntarily consented to participate in a research study.

Refusal rate: 7.1 percent.

IRB approval: George Washington University, #NCR213434.

Preregistration: https://osf.io/7b8e2/?view_only=86ca5de4073948199ca39bc098c32e12

B.2 Question text

Next you will complete a short knowledge quiz.

DISPLAY IF z_pledge = 1:

We want to measure what you already know about these questions. Please do not cheat by looking up the answers, asking someone, or getting help in any other way.

Do you promise not to cheat?

[Yes, I promise not to cheat.] [No.]

Do you happen to know which party currently has the most members in the U.S. House of Representatives in Washington?

[Democrats] [Republicans] [Same number of members]

Do you happen to know which party currently has the most members in the U.S. Senate?

[Democrats] [Republicans] [Same number of members]

What job or political office is now held by John Roberts?

[Senate Majority Leader] [Secretary of Defense] [Chief Justice of the Supreme Court] [Vice President] [Attorney General] [Speaker of the House of Representatives] [Circuit Court Judge]

What job or political office is now held by Merrick Garland?

[Senate Majority Leader] [Secretary of Defense] [Chief Justice of the Supreme Court] [Vice President] [Attorney General] [Speaker of the House of Representatives] [Circuit Court Judge]

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 U.S. Senator?

DISPLAY IF z_sen = open: Please type a number: _____

DISPLAY IF z_sen = multi (Studies 2 and 3) or to all respondents (Study 1): [Once] [Twice] [Three times] [Four times] [Unlimited number of times] [None of these]

STUDIES 2 and 3 ONLY:

How many times can an individual be elected President of the United States under current laws?

DISPLAY IF z_pres = open: Please type a number: _____

DISPLAY IF z_pres = multi: [Once] [Twice] [Three times] [Four times] [Unlimited number of times] [None of these]

STUDIES 2 and 3 ONLY:

Who is currently the Prime Minister of the United Kingdom?

[Richard Branson] [Boris Johnson] [David Cameron] [Theresa May] [Margaret Thatcher] [Winston Churchill]

STUDY 1: In what year did the U.S. Supreme Court decide United States v. Segui? STUDY 2: In what year did the U.S. Supreme Court decide [RANDOMLY ASSIGNED: Oliver v. Alexander / United States v. Arredondo]? STUDY 3: In what year did the U.S. Supreme Court decide [RANDOMLY ASSIGNED: Anderson v. Dunn / United States v. Segui]?

Please type a number: _____

[SERIES OF UNRELATED QUESTIONS]

BONUS OPPORTUNITY

Before the survey ends, we want to learn about how people search for information.

Please look up the answer to this question. At the end of the survey, one person who answers it correctly will receive a code for a [STUDIES 1 AND 2: **$100** / STUDY 3: **$200**] **Amazon gift card**.

STUDY 1: In what year did the U.S. Supreme Court decide Atkin v. Kansas?

STUDY 2: In what year did the U.S. Supreme Court decide Anderson v. Dunn?

STUDY 3: In what year did the U.S. Supreme Court decide [RANDOMLY ASSIGNED: Anderson v. Dunn / United States v. Segui]?

Please type a number: _____

Note: In Study 3, one case was always asked as the catch question and the other case was always asked as the pay-to-search question. Simple random assignment, $p = 0.5$.

How did you come up with your answer to the last question?

Please be honest. Your answer will not affect your payment. We just need to know what you did.

[I did not look it up, and just took my best guess.] [I did not look it up, and skipped the question.] [I looked it up using the same device I am using to take the survey.] [I looked it up using a different device.]

DISPLAY IF INCORRECT ANSWER TO BONUS QUESTION:

For the gift card drawing, your answer was [RESPONDENT'S ANSWER]. The correct answer was [CORRECT ANSWER]. Because you did not answer correctly, you were not eligible to win.

DISPLAY IF CORRECT ANSWER TO BONUS QUESTION + TICKET NOT EQUAL TO WINNING NUMBER:

For the gift card drawing, you answered the question correctly.

Your ticket number was [TICKET NUMBER]. The winning number was [WINNING NUMBER]. Sorry you weren't selected.

DISPLAY IF CORRECT ANSWER TO BONUS QUESTION + TICKET EQUAL TO WINNING NUMBER:

For the gift card drawing, your ticket number was [TICKET NUMBER]. The winning number was [WINNING NUMBER]. Congratulations! You won the gift card.

To redeem the gift card, please copy the code below and enter it into your Amazon.com account. Visit your "account" page, then click the "gift card" link at the top of the screen.

[GIFT CARD CODE]

If you have any problems redeeming this code, please contact us at [RESEARCHER'S EMAIL ADDRESS] and we will correct the problem. In your email, please include the gift card code and your ticket number.

B.3 Implementing the paradata method

These instructions describe the implementation of the paradata method used in this paper, using the Chief Justice question as an example. To implement it for other questions, simply change all instances of "roberts" in both steps to your desired prefix.

Step 1. Add the following JavaScript to the page containing the survey question.

```
Qualtrics.SurveyEngine.addOnload(function(){
Qualtrics.SurveyEngine.setEmbeddedData("roberts_everLeft", "0");
function recordLeave() {
var prefix = "roberts_";
if (document.hidden) {
var theStart = new Date();
var embedString = prefix + "everLeft";
Qualtrics.SurveyEngine.setEmbeddedData(embedString, "1");
var embedString = prefix + "timeLeft";
Qualtrics.SurveyEngine.setEmbeddedData(embedString, theStart.getTime()/1000);
}
};
document.addEventListener('visibilitychange', recordLeave, false);
function recordArrive() {
var prefix = "roberts_";
if (document.hidden){}else {
var theEnd = new Date();
var embedString = prefix + "timeReturn";
Qualtrics.SurveyEngine.setEmbeddedData(embedString, theEnd.getTime()/1000);
}
};
document.addEventListener('visibilitychange', recordArrive, false);
$('NextButton').onclick = function (event) {
document.removeEventListener('visibilitychange', recordLeave);
document.removeEventListener('visibilitychange', recordArrive);
Qualtrics.SurveyEngine.navClick(event, 'NextButton')
};
});
```

Step 2. Add the following embedded variables to the beginning of the survey flow, or at any point in the survey flow **before** the question.

The `_everLeft` variable will be set to 1 if the respondent ever looks away from the page. The analysis in this paper uses the `_everLeft` variables as the paradata flag.

The `_timeLeft` and `_timeReturn` variables record the time the respondent leaves the page, measured as the number of seconds since midnight on January 1, 1970. The difference between the two gives the total time spent away from the page.