AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Detecting and Deterring Information Search in Online Surveys

Matthew H. Graham Temple University

Abstract: *This article introduces a framework for evaluating methods of combatting information search in online surveys. Three empirical studies based on the framework suggest that search is a serious but manageable problem. Search frequency varies substantially according to question content, ranging from 2% to 30% in batteries of general political knowledge questions. Deterrence works: a pledge not to cheat reduces search by half. Detection also works: web browser paradata identify 70% to 85% of search, while 60% to 85% of search on knowledge questions is undertaken by respondents who correctly answer "catch" questions about obscure Supreme Court cases. Detection and deterrence are complementary: deterrence reduces search* ex ante, *while detection quantifies success and provides* ex post *options for dealing with undeterred search. In combination, the three methods tested (pledge, paradata, and catch) deter or detect more than 90% of search, leaving search to affect about 0.5% of the remaining observations.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the American Journal of Political Science Dataverse within the Harvard Dataverse Network, at: doi.org/10.7910/DVN/XNCQR8.

Unsupervised online surveys make it easy for respondents to look up the answers to questions designed to measure factual knowledge (Clifford and Jerit 2014; Liu and Wang 2014; Shulman and Boster 2014; Strabac and Aalberg 2011). Researchers have developed several methods to address information search (hereafter simply "search"). Likely searchers can be detected using self-reports (Jensen and Thomsen 2014), "catch" questions that are difficult to answer correctly without help (Bullock et al. 2015; Motta, Callaghan, and Smith 2016), and 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 those caught in the act (Diedenhofen and Musch 2017).

This article introduces a framework for accounting for measurement error in estimates of the prevalence of information search. This opens the door for more specific evaluation of methods to combat it: how well they perform, why they fall short, and how performance changes when methods are combined. 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 yield the following key findings:

1. **Information search is common in political knowledge surveys.** At baseline, search was estimated to be present in 7.8% of answers to widely used political knowledge questions in Study 1, 17.6% in Study 2, and 11.6% in Study 3. These estimates are adjusted for false positives and 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% in Study 1, 8.7% to 29.8% in Study 2, and 1.9% to 21.3% in Study 3.
3. **Response scales affect search.** In split ballot experiments, search was 30% to 100% more common on open-ended questions than on

Matthew H. Graham, Assistant Professor, Department of Political Science, Temple University, 1115 Polett Walk, Polett Walk, Philadelphia, PA 19122 (mattgraham@temple.edu).

Earlier versions of this work were presented at the Yale Cooperative Lucid Survey Mini-Conference (July 2020) and the annual meeting of the American Association for Public Opinion Research (May 2022). Study 1 was funded by the Georg Walter Leither Program in Political Economy at Yale University. Studies 2 and 3 were funded by the John S. and James L. Knight Foundation through a grant to the Institute for Data, Democracy & Politics at George Washington University.

American Journal of Political Science, Vol. 0, No. 0, xxxx 2023, Pp. 1–20

DOI: 10.1111/ajps.12786

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% in each study.
5. **Detection works.** The paradata method detected between 70% and 85% of search in all three studies. The catch method detected 60% to 85%.
6. **Detection and deterrence are complements.** Combining detection and deterrence eliminates more search from the data than either achieves alone. 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% of those flagged in paradata looked up the answer to the knowledge questions, compared with 30% to 45% 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 saw more than twice as much search as the average knowledge question. Yet when used to approximate the proportion of respondents who search at least once, catch questions underestimated 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% or less of the unflagged observations. This falls to 0.1% for questions with the lowest base rates of search and rises to 0.9% 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% of the unflagged observations at a much lower cost in terms of missing data.

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 valuable because it eliminates search *ex ante*. But without detection, researchers are forced to tolerate an unknown amount of search. Detection quantifies undeterred search and gives researchers *ex post* options for dealing with it. Third, question content and response scales affect the baseline prevalence of search and, by extension, the optimal combination of strategies for dealing with it. 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. Finally, paradata are preferable to catch questions. Though both methods catch a similar proportion of those who search (sensitivity), paradata identify searchers with much greater precision (specificity) and capture variation within respondents and across questions. This makes the researcher's options for dealing with information search more attractive: fewer observations to drop and better options for dealing with missing data. As elaborated on in the concluding section, catch questions' value appears limited to a few circumstances: when paradata are not available, when the base rate of search is unusually high, and as "lab rats" for testing the relative efficacy of detection methods.

Approaches to Countering Information Search

Relative to traditional interviewer- or lab-administered surveys, online surveys make search easier for respondents to undertake and harder for researchers to prevent. Existing research suggests that search is alarmingly prevalent. Published estimates based on catch questions and self-report measures range from 15% to 25% (Bryson 2020; Jensen and Thomsen 2014; Motta, Callaghan, and Smith 2016; Style and Jerit 2021). Estimates based on paradata are more varied, ranging from 3% to 30% (Diedenhofen and Musch 2017; Gummer and Kunz 2019; Höhne et al. 2021). This threatens the validity of knowledge scales, which are widely used in the study of politics (Appendix A.6, page A16). In addition to the general knowledge scales studied here, search threatens any survey measures with answers that can be lookedup—for example, domain-specific measures of knowledge of history (Starratt et al. 2017), science (Cooper and Farid 2016), politicians' positions (Ansolabehere and Jones 2010), and outgroup bias (Ahler and Sood 2018).

Researchers use two classes of methods to deal with 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, Huber, and Lenz 2012; Motta, Callaghan, and Smith 2016), examining paradata generated by the respondent web browser (Diedenhofen and Musch 2017; Gummer and Kunz 2019; Höhne et al. 2021), and collecting browsing history (Gooch and Vavreck 2019). Deterrence methods seek to dissuade respondents from searching in the first place. These include requests (Motta, Callaghan, and Smith 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 costs imposed by *ex post* methods of dealing with suspected search (such as 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 no sense of the problem's scope. Detection complements deterrence by (2) quantifying search and (1) providing researchers with *ex post* options for dealing with search that they were unable to deter. This article 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 (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. Methods that detect all search are highly *sensitive* ($P(\text{flag}|\text{search}) \approx 1$). 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 searched ($P(\text{flag}|\neg\text{search}) \approx 0$). Yet to the extent that those who search are reluctant to admit it, self-reports are likely to have low sensitivity ($P(\text{flag}|\text{search}) < 1$).

The two detection methods used in this article 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 name the year in which an obscure Supreme Court case was decided (Motta, Callaghan, and Smith 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 ($P(\text{flag}|\neg\text{search}) \approx 0$).[1] As long as the answer is easy to look up, they will also be highly sensitive ($P(\text{flag}|\text{search}) \approx 1$). 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 under-specific. To the extent that respondents who search on knowledge questions do not search on catch questions, the catch method may be externally under-sensitive.

The paradata method uses a snippet of JavaScript to measure a likely indicator of search: obscuring the survey with another browser window or application. Paradata may be under-specific if this occurs for reasons other than search; for example, one may view a text message on their mobile phone. Paradata may be under-sensitive if respondents search in some way that the method cannot detect. For example, one 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 under-sensitivity are artificially limited in supervised settings, like laboratories, wherein survey-takers use researcher-provided devices and have limited access to alternative means of search. This limits the generalizability of laboratory-based audits.

Among previously published research, this article's approach is closest to that of Diedenhofen and Musch (2017, hereafter DM). This article improves on DM in two respects. First, DM's most detailed assessments of their paradata method, PageFocus, were conducted in a laboratory. As just noted, this setting artificially limits paradata methods' greatest vulnerabilities. By contrast, this article's respondents had full access to modes of search that paradata methods cannot detect. Second, for the survey not conducted in a lab, DM used self-reports to measure the ground truth. The quantity DM reported for the paradata's sensitivity is $P(\text{flag}|\text{subsequently admitted to searching})$. These self-admissions appear remarkably under-sensitive: in the search-discouraged group, 18 respondents are flagged but only three admitted searching (see their Tables 1 and 2), suggesting a sensitivity rate of 17% or less.[2] By contrast, this article only uses self-reports in an audit that verifies the interpretation of conflicts between the

[1] Catch questions about Supreme Court case years 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 former have more plausible responses.

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

TABLE 1 Methods of Dealing with Information Search Evaluated in this Article

Method	Purpose	Flag	Sources of false negatives (under-sensitivity)	Sources of false positives (under-specificity)
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 an incompatible browser to take the survey Using a different device to look up the answer Asking someone for help	Non-search behavior that obscures the survey (e.g., checking email, reading a text message)
Pledge	Deter	None	Not applicable	Not applicable

Note: Table summarizes the three methods for dealing with information search evaluated in this article.

two detection methods. Self-reports never enter the quantitative estimates of the two methods' performance.

More generally, this article overcomes three limitations in existing research on countering information search in online surveys. First, it quantifies systematic measurement and corrects for the resulting bias. By contrast, existing research on cheating in online surveys rarely quantifies measurement error and never adjusts estimates to account for it. Second, this approach to error enables more precise performance assessments. Convincing evidence exists that some methods of detecting and deterring information search are likely to help but little regarding how much they help or how they fall short. For example, research shows 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 is identified or missed. Third, whereas existing research examines each evaluated method in isolation, this article examines the effects of layering methods atop one another.

Methodology

The analysis examines three surveys fielded in 2020 and 2021. For Studies 1 and 3, 2,176 and 6,687 respondents were recruited online by Lucid with quota sampling to census demographic benchmarks. For Study 2, 5,411 respondents were recruited through Amazon Mechanical Turk (MTurk). All respondents completed a captcha, and all those recruited on Lucid passed an attention check (Peyton, Huber, and Coppock 2021; Ternovski and Orr 2022). Studies 2 and 3 were preregistered. Full details appear in Appendix B.1 (page A18).

Each study followed the same sequence. Just before the knowledge quiz, 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. Next, all respondents completed the political knowledge quiz, which consisted of five questions in Study 1 and seven in Studies 2 and 3. Studies 2 and 3 randomized the format of two questions between closed-ended (that is, multiple choice) or open-ended, bringing the number of knowledge items in those studies to nine. Finally, after some unrelated questions, all respondents completed a "pay-to-search" task. Each respondent 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.[3]

Each survey included both detection methods, paradata and a catch question. Although the paradata method is similar in many respects to those described by Diedenhofen and Musch (2017) and Permut, Fisher, and Oppenheimer (2019), it was developed independently. Appendix B.3 (page A22) and the replication file each

[3] Study 1's pay-to-search task was randomly assigned along with a task that *discouraged* looking up the same answer. This was later judged not to add value but is reported below for transparency.

TABLE 2 Pay-to-Search Audit of Detection Methods

Self-reported search by detection status		**Study 1**			**Study 2**			**Study 3**	
	N	Percent of total	Percent of group	N	Percent of total	Percent of group	N	Percent of total	Percent 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		

Note: Table displays the results of the pay-to-search audit of the paradata and catch methods' shortfalls in performance. Rows display self-reported search behavior by by detection status. Columns display the frequency (N), joint probability (Percent of total), and conditional probability (Percent of group).

contain one-page instructions for implementing it in Qualtrics.

The bias correction derived in the next section consists of multiple parameters that are estimated using multiple survey questions per respondent. Accordingly, uncertainty is estimated 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, Hopkins, and Yamamoto 2015), time series data (Bertrand, Duflo, and Mullainathan 2004), and analysis that pools across many knowledge questions (Graham 2020). In tables, parenthesized values are standard errors (s.e.). Figures display 95% confidence intervals calculated using the percentile method. Whenever the prevalence of search is estimated conditional on another variable, all components of Equation 2 are estimated within the subgroup of interest. All plotted estimates appear in tabular form in Appendix A (page A1).

Dealing with Measurement Error

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

Bias Correction for the Prevalence Information Search

Indirect methods of detecting search suffer from two basic problems: some may be missed while others might be wrongly accused of searching. To worry that a method does not flag all search is to worry that under-sensitivity causes false negatives, that is, that $P(\text{flag}|\text{search}) < 1$. To worry that a method flags people who do not search is to worry that under-specificity causes false positives, that is, that $P(\text{flag}|\neg\text{search}) > 0$.

This article'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, write what researchers observe ($P(\text{flag})$) in terms of the estimand ($P(\text{search})$). By the law of total probability,

$$\begin{aligned} P(\text{flag}) &= P(\text{flag}|\text{search})P(\text{search}) \\ &\quad + P(\text{flag}|\neg\text{search})(1 - P(\text{search})), \qquad (1) \end{aligned}$$

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

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

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

In Equation 2, the terms other than $P(\text{flag})$ amount to a bias correction. When one assumes that a measure is perfectly sensitive and specific (that is, $P(\text{flag}|\text{search} = 1$ and $P(\text{flag}|\neg\text{search}) = 0$), the remaining terms disappear and Equation 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 Equation 2 into practice requires one to either estimate sensitivity and specificity or to assume that these problems can be safely ignored. The remainder of this section explains how this article addresses 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 detail (page A1).

Estimating Sensitivity

To help interpret the detection methods' failures to flag search, each study featured a "pay-to-search" task. Respondents were asked to look up the answer to a catch question in exchange for entry into a draw 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. Did they look up the answer using the same device they used to take the survey, look in some other way, or not look it up at all? 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 the following analysis.

Large majorities complied with the request to look up the answer. The percentages that either answered correctly or were flagged in the paradata were 69.3 in Study 1, 92.5 in Study 2, and 72.3 in Study 3. Among those flagged by both methods, about 90% 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 correctly answered the pay-to-search question but was not flagged in the paradata (Table 2, second group of rows). The self-reported 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 reported using 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.[4] This equals 0.78 (i.e., 78%) in Study 1, 0.84 in Study 2, and 0.70 in Study 3. The analysis assumes that this quantity is constant across knowledge questions, which is reasonable given that the causes of under-sensitivity are either constant for the duration the survey (for example, browser incompatibility) or plausibly reflective of individual-level dispositions (for example, a tendency to use a different device). This assumption does have vulnerabilities. In particular, respondents may not try as hard to avoid detection when told that search is allowed. However, relative to the prevailing practice of analyzing paradata methods as if they are error-free, the assumptions that error exists and is constant across questions relaxes stronger, less credible assumptions.[5]

The catch method's failures to detect search are represented by the "incorrect + paradata flag" category

[4] Specifically, $P(\text{paradata flag}|\text{paradata flag or answered correctly})$. For example, in Study 1, the calculation is $(532 + 52)/(532 + 166 + 52)$.

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

(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. This combination of results suggests that catch questions may be internally under-sensitive: that is, respondents who were flagged by the paradata but answer the catch question incorrectly really did try to look up the answer. Further evidence that catch questions can be internally under-sensitive emerged from an informative accident in Study 2, wherein some respondents were fooled by incorrect answers in search results (for example, 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 internal under-sensitivity, the analysis treats the catch method'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, under-sensitivity cannot exist.

As noted above, the catch method is more vulnerable to external under-sensitivity. A strategy for examining this is introduced further below.

Estimating Under-Specificity

The paradata method flags instances in which the survey becomes partially or fully obscured on the respondent's screen. This will produce false positives, and consequently be under-specific, if behavior other than search triggers the flag. To approximate $P(\text{flag}|\neg\text{search})$ for the paradata, the method was added to two sets of baseline questions where looking up the answer is unlikely to be necessary (such as age) or undefined due to the question's subjective nature (such as interest in politics). 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.

The catch method flags respondents who answer catch questions correctly. Internally, lucky guesses are the key source of under-specificity. Researchers try to maximize specificity by choosing catch questions with many plausible response options, thereby minimizing the probability of a correct guess. For example, if 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, Callaghan, and Smith 2016) is about 0.004. However, because non-searchers' guesses concentrate in recent years and at multiples of five (Figure 1), avoiding cases decided in such years further increases internal specificity. To estimate the expected rate of lucky guessing, Appendix A.2 uses local linear regression to estimate the probability distribution function 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 zero in all analysis.

As with sensitivity, the catch method is more vulnerable to external under-specificity. This is examined further below.

Detecting Information Search

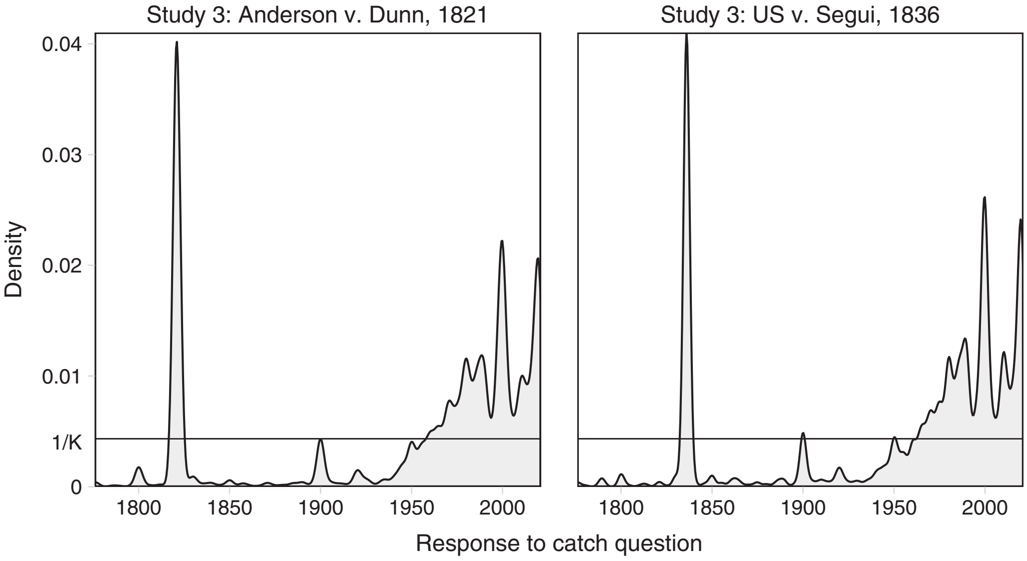
The bias correction approach is first applied to the prevalence of search in the absence of deterrence. Figure 2 presents these base-rate estimates for each question in all three studies. Light grey bars are empirical estimates of Equation 2. For catch questions, dark grey bars show the proportion of correct answers, which should equal Equation 2 if catch questions have perfect internal sensitivity and specificity.

The estimates indicate that search is common. Across all answers to knowledge questions, search occurred in 7.8% of observations in Study 1, 17.6% in Study 2, and 11.6% in Study 3.

The average rate hides substantial variability between questions. In Study 1's knowledge battery, search ranged from 2.5% (on a question about party control of the senate) to 11.5% (on a question about Chief Justice John Roberts' job or political office). In Study 2, search ranged from 8.7% (the closed-ended version of the presidential term limit question) to 29.8% (on a question about Attorney General Merrick Garland's job or political office). Study 3 exactly replicated Study 2's question battery. Here, search was most rare and most common on the same questions (1.9% and 21.3%) but fell below the Study 2 level on every question. 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%) more common on open-ended than

FIGURE 1 Distribution of Responses to Catch Questions—Study 3



Notes: Figure displays the probability density function (PDF) of responses to the catch questions in Study 3. The horizontal line at 1/K depicts the uniform PDF that would realize if all K response options were equally likely to be guessed. Equivalent figures for Studies 1 and 2 appear in Appendix A.2 (page A7).

closed-ended questions. Despite this, respondents were two to four p.p. less likely to answer the open-ended versions correctly. 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.

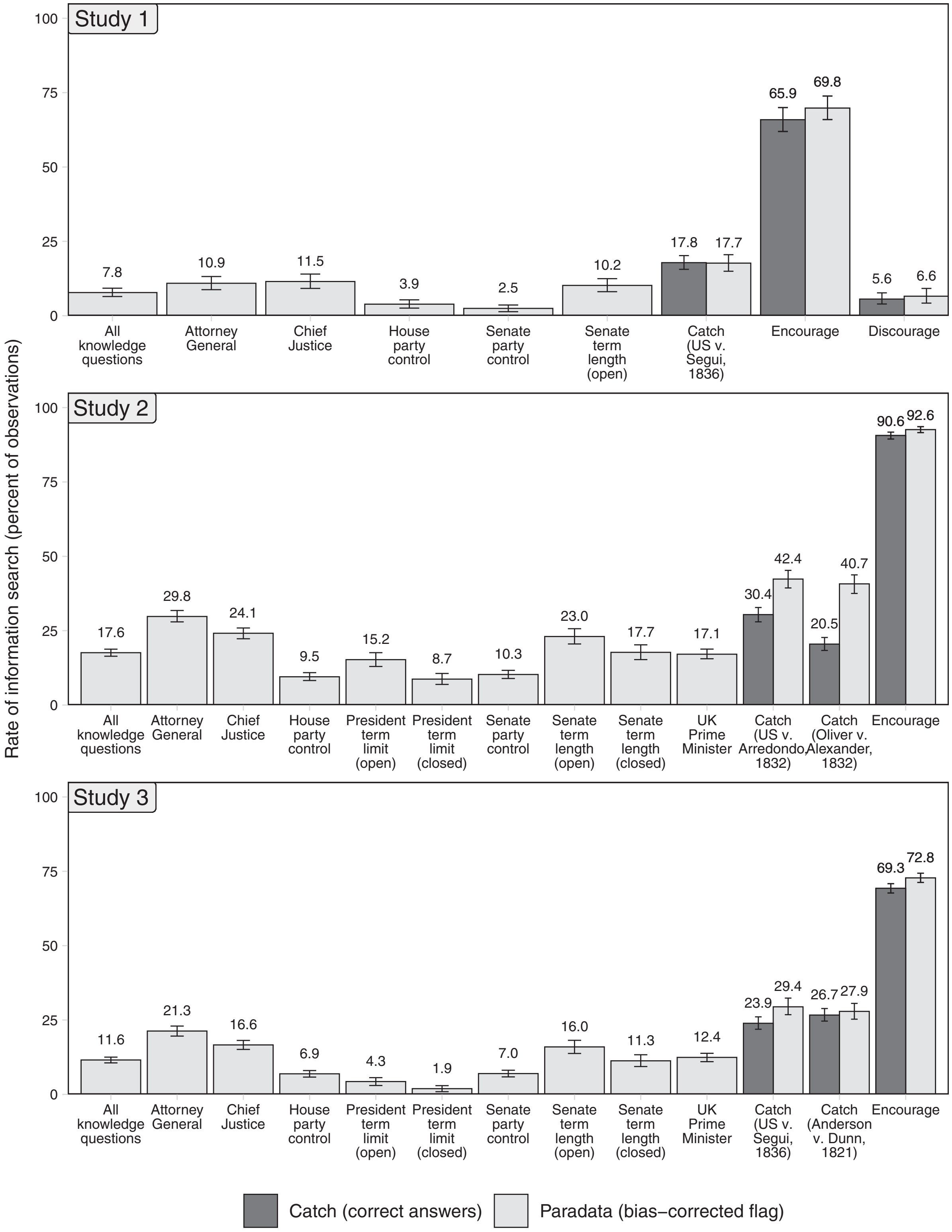
Search is much more common on catch questions than on knowledge questions (Figure 2, center-right). First, consider the bias-corrected estimates based on the paradata. In all three studies, the estimated rate of search on catch questions more than doubled the average rate for the knowledge battery: 17.7 versus 7.8 in Study 1, 42.4 and 40.7 versus 30.4 and 20.5 in Study 2, and 29.4 and 27.9 versus 23.9 and 26.7 in Study 3. When one instead uses the proportion of correct answers on catch questions as an estimate of the prevalence of search (what has been referred to as the "catch method"), results are similar in Studies 1 and 3. By contrast, in Study 2, the proportion of correct answers is considerably lower than the bias-corrected paradata estimate. Consequently, the catch method provides a somewhat more accurate estimate of the rate of search on knowledge questions.

The catch method is not much better as an approximation for the proportion of respondents 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 searching at least once. 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 respondents 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 (−5.5 and −6.1. p.p.). Though the estimate from Study 2 again stands out, this time it is for the opposite reason: the proportion of correct answers to the catch question substantially *underestimates* the proportion who actually searched.

The differences in the catch method's performance between Study 2 and the others are at least partly attributable to a source of internal under-sensitivity in the catch method: it cannot detect failed attempts to cheat. Though the Supreme Court cases for Study 2 were selected based on the false positive-minimizing principles developed in Study 1 (early 1800s, no multiples of five), they had another feature that drove up the share of false negatives: incorrect answers appearing in search queries that, to a satisficing searcher,[6] 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 year of a district court case by the name of Oliver v. Alexander County Housing Authority. Whereas the catch method's under-sensitivity in Study 2 allows it to more accurately approximate the rate of search on the average question (Figure 2),

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

FIGURE 2 Estimated Rate of Information Search



Notes: Figure displays the estimated rate of search by study, question, and detection method.

TABLE 3 Estimated Percentage of Respondents Searching at Least Once

	No Pledge			Pledge		
	Percent correctly answering catch question	Percent searching on at least one question	Difference	Percent correctly answering catch question	Percent searching on at least one 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: Table displays the estimated percentage of respondents who searched at least once by study, detection method, and presence/absence of a pledge. Block bootstrapped s.e. in parentheses.

this same property leads to the largest underestimate of the proportion who search at least once (Table 3). This highlights the scattershot nature of using the catch method to approximate the rate of search on knowledge questions.

Deterring Information Search

The framework is next applied to the efficacy of deterrence methods. This requires bias-corrected estimates of the difference in the rate of search between two groups: those who were and were not randomly assigned to the pledge. Generically, suppose that two groups are defined by $X \in 0, 1$. The resulting difference in conditional proportions,

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

can be used to compare the rate at which any two groups of respondents look up answers in a political knowledge survey.

Table 4 presents estimates of the pledge's efficacy for each knowledge question in all three studies. Overall, the pledge reduced search by about half in all three studies: 50.7% in Study 1, 47.7% in Study 2, and 56.7% 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% of responses in Study 1, 9.2% of responses in Study 2, and 5.0% of responses in Study 3.

The pledge's effect is similar from question to question. All but one estimate is both negative and statistically significant, suggesting that the pledge works for a range of questions. All but two of of the percentage reductions fall in the 35% to 65% range.[7] When the effects are expressed in percentage points the variation is somewhat more pronounced. This is a function of base rates: the more cheating there is to begin with, the more the pledge eliminates. Across all questions in all three studies, the correlation between the base rate of search and the effect of the pledge is −0.94.

The catch questions provide an opportunity to cross-check the paradata-based estimates of the pledge's efficacy (Table 5). In Study 1, the paradata-based estimate suggests that the pledge reduced search on the catch questions by 8.7 p.p. (50.9%). Similarly, the proportion of correct answers declined by 8.5 p.p. (52.3%). In Study 2, the paradata-based estimates suggest larger absolute reductions (14.9 and 22.0 p.p. 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%). In Study 3, the estimates are once again similar in both absolute and percentage terms (about a 60% 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% reduction, search remains fairly common. The next section examines the benefits and costs of going further.

[7] The two exceptions are the two questions with the lowest base rate of search, which makes the percentage reduction difficult to estimate precisely.

TABLE 4 Deterrent Effect of Pledge—Knowledge Questions

	Study 1			Study 2			Study 3		
	No pledge	Pledge	Effect	No pledge	Pledge	Effect	No pledge	Pledge	Effect
All knowledge questions	7.8	3.8	−3.9	17.6	9.2	−8.4	11.6	5.0	−6.6
	(0.7)	(0.5)	(0.9)	(0.6)	(0.5)	(0.7)	(0.5)	(0.4)	(0.6)
Attorney general	10.9	5.8	−5.1	29.8	14.7	−15.0	21.3	9.0	−12.3
	(1.1)	(0.9)	(1.4)	(1.0)	(0.8)	(1.2)	(0.9)	(0.6)	(1.0)
Chief justice	11.5	5.6	−5.9	24.1	13.4	−10.7	16.6	6.9	−9.7
	(1.2)	(0.9)	(1.5)	(0.9)	(0.7)	(1.2)	(0.8)	(0.5)	(1.0)
House party control	3.9	1.5	−2.4	9.5	3.7	−5.8	6.9	2.5	−4.4
	(0.7)	(0.5)	(0.9)	(0.7)	(0.4)	(0.8)	(0.6)	(0.4)	(0.7)
Presidential term limit (open)				15.2	9.7	−5.5	4.3	2.3	−2.0
				(1.2)	(0.9)	(1.5)	(0.7)	(0.5)	(0.9)
Presidential term limit (closed)				8.7	5.6	−3.1	1.9	2.5	0.6
				(0.9)	(0.8)	(1.2)	(0.5)	(0.6)	(0.8)
Senate party control	2.5	0.2	−2.2	10.3	3.2	−7.0	7.0	2.5	−4.5
	(0.6)	(0.3)	(0.7)	(0.7)	(0.4)	(0.8)	(0.6)	(0.4)	(0.7)
Senate term length (open)	10.2	6.1	−4.1	23.0	13.7	−9.3	16.0	6.9	−9.0
	(1.1)	(0.9)	(1.5)	(1.3)	(1.0)	(1.7)	(1.1)	(0.8)	(1.4)
Senate term length (closed)				17.7	8.2	−9.6	11.3	5.8	−5.5
				(1.2)	(0.9)	(1.5)	(1.0)	(0.8)	(1.2)
UK prime minister				17.1	10.7	−6.4	12.4	5.5	−7.0
				(0.8)	(0.7)	(1.1)	(0.7)	(0.5)	(0.9)

Note: Table displays the estimated rate of search by study, knowledge question, and presence/absence of a pledge. Block bootstrapped s.e. in parentheses.

Eliminating Information Search

Efforts to deter search are unlikely to be perfectly successful. Detection gives researchers the option to take further, *ex post* steps to eliminate search from their data, either in the main analysis or as a robustness check. In particular, researchers may treat contaminated responses as missing data, then manage the missingness by dropping observations or imputing values. The missing data constitute a cost that deterrence does not impose.

This section builds on the framework above to define and estimate quantities that capture the trade-off between eliminating search and creating missing data. The key new step is to use the paradata to evaluate the catch method's external specificity. To do so, I condition estimates of Equation 2 on whether the respondent is flagged by the catch method. Combined with further algebra, this enables one to quantify essential aspects of the catch method's performance.

The paradata method is superior to the catch method in one sense that is not captured by the quantities 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. A reasonable knowledge score can be estimated for respondents like this using a model that makes use of the other items in a principled manner, for example, 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 detects search at the respondent level only. This limits the *ex post*

TABLE 5 **Deterrent Effect of Pledge—Catch Questions**

		Paradata (Bias-Corrected)			**Catch (Percent Correct)**		
		No pledge	Pledge	Effect	No pledge	Pledge	Effect
Study 1	US v. Segui, 1836	17.7	8.7	−9.0	17.8	8.5	−9.3
		(1.4)	(1.0)	(1.8)	(1.2)	(0.8)	(1.4)
Study 2	US v. Arredondo, 1832	42.4	27.5	−14.9	30.4	20.0	−10.4
		(1.5)	(1.3)	(2.0)	(1.3)	(1.0)	(1.6)
	Oliver v. Alexander, 1832	40.7	18.7	−22.0	20.5	9.1	−11.4
		(1.6)	(1.2)	(2.1)	(1.1)	(0.8)	(1.4)
Study 3	US v. Segui, 1836	29.4	11.5	−17.9	23.9	9.0	−14.9
		(1.4)	(1.0)	(1.7)	(1.1)	(0.7)	(1.3)
	Anderson v. Dunn, 1821	27.9	11.7	−16.2	26.7	10.1	−16.6
		(1.3)	(1.0)	(1.7)	(1.1)	(0.7)	(1.3)

Note: Table displays the estimated rate of search by study, catch question, and presence/absence of a pledge. Block bootstrapped s.e. in parentheses.

solution toolkit to dropping suspected searchers from the analysis entirely or imputation based on variables that are not part of the knowledge battery.

Quantifying Trade-Offs

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

The first quantity is sensitivity, which is defined above. Sensitivity answers the question, "what proportion 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. The ability to use the paradata to estimate the rate of search for different subgroups of respondents (for example, Equation 3) provides an opportunity to examine this empirically. The catch method's external sensitivity will be estimated as

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

One of the quantities on the right-hand side is directly measured: P(flag) is the probability of answering the question correctly. The other two can be estimated using the paradata. The denominator is simply Equation 2, while the first term in the numerator is Equation 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(search|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 Equation 2 among those who answered the catch question correctly, just as it appears in the numerator of Equation 4. For the paradata, the positive predictive value will be estimated as

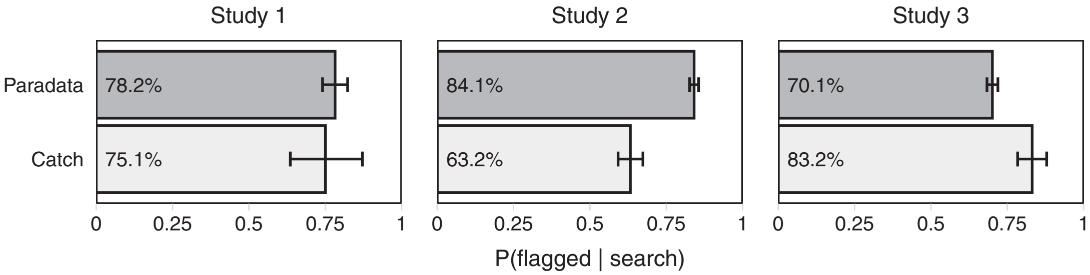
$$P(\text{search}|\text{flag}) = \frac{P(\text{flag}) - P(\text{flag}|\neg\text{search})P(\neg\text{search})}{P(\text{flag})} \quad (5)$$

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

The third quantity is the probability of being detected, P(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 proportion of respondents who trigger the flag, which is equivalent to the proportion 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, that is, the probability of search among unflagged observations (P(search|¬flag)). This quantity

FIGURE 3 Sensitivity, by Detection Method



Notes: Figure displays the estimated sensitivity of the paradata and catch methods. Error bars display 95 percent block bootstrapped confidence intervals.

could also be called the complement of the negative predictive value. It answers the question, "how much search 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 a combination of quantities that have been defined. The first term in the numerator is the complement of Equation 4. The second is Equation 2. The denominator is observed.

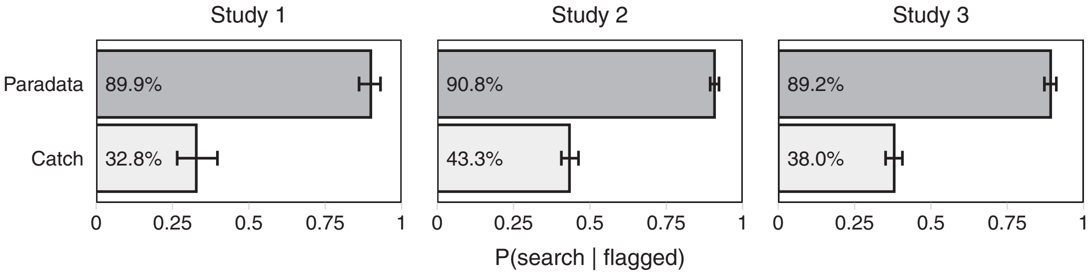
Estimates

To begin understanding the trade-off between eliminating search and creating missing data, examine the sensitivity estimates that appear in Figure 3. In all three studies, the paradata identified more than two-thirds of search: 78.2% in Study 1, 84.1% in Study 2, and 70.1% in Study 3. The catch method was about equally sensitive, flagging 75.1% of search in Study 1, 63.2% in Study 2, and 83.2% 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 about 90% of flagged observations. For every nine responses that are correctly flagged, one is incorrectly flagged. The catch method is less precise, with predictive values falling between 30% and 45% 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 also flagged. Put simply, even though the catch method flags just as many true positives, this comes at a higher cost in terms of false positives.

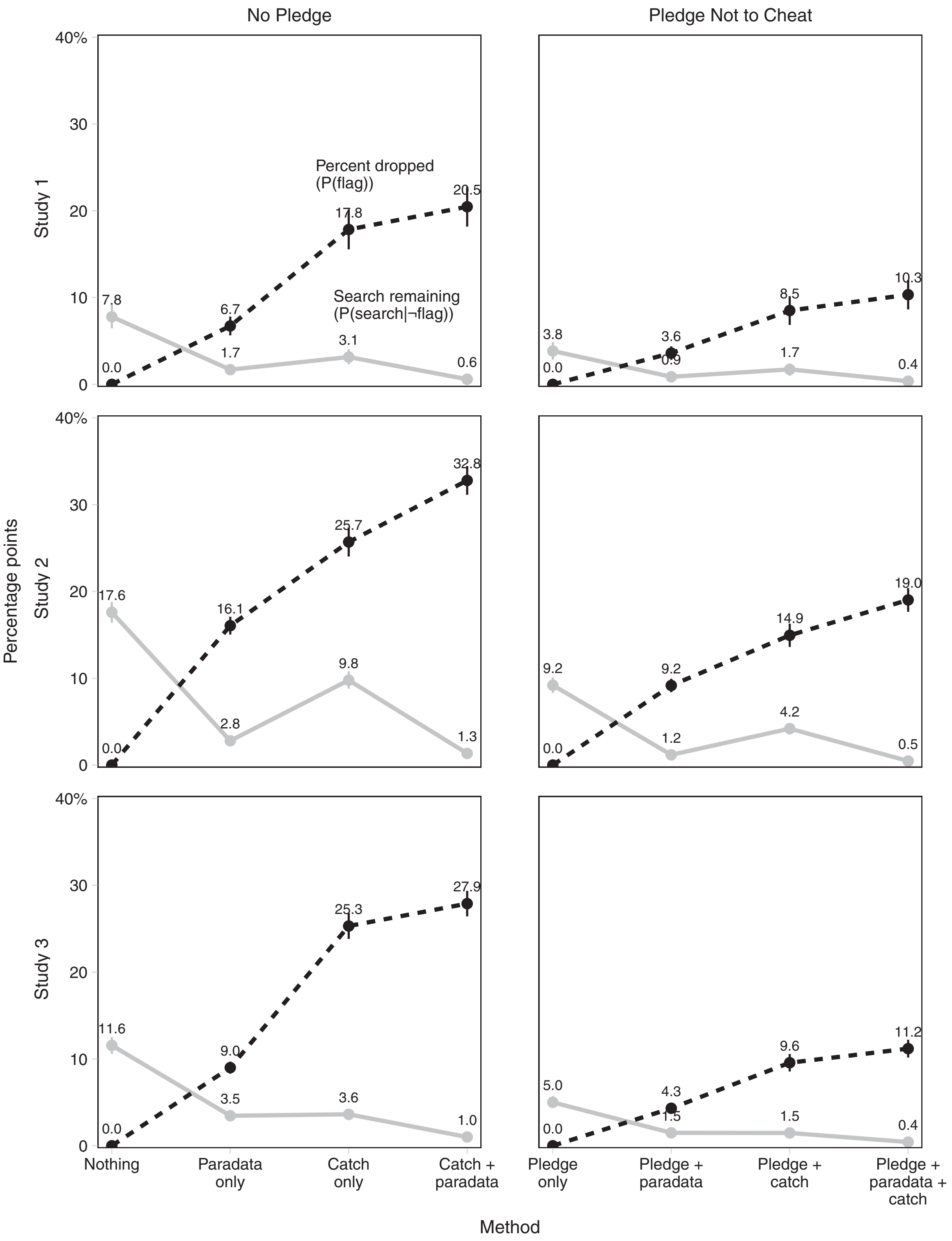
To understand the implications for practice, it is helpful to examine the trade-off between price and the bottom line. Figure 5 plots the third quantity, the percentage of data lost when instances of suspected search are treated as missing data, against the fourth quantity, the amount of search present 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%. Each point to the right represents

FIGURE 4 Positive Predictive Value, by Detection Method



Notes: Figure displays the estimated positive predictive value of the paradata and catch methods. Error bars display 95 percent block bootstrapped confidence intervals.

FIGURE 5 Trade-Off between Eliminating Search and Missing Data



Notes: Figure illustrates the tradeoff between using detection methods to eliminate search and the resultant creation of missing data. Each point on the x-axis is a strategy or combination of strategies for dealing with search. Black dots and dashed lines display the proportion of data dropped by each strategy. Grey dots and solid lines display the rate of information search in the remaining data (i.e. the data that are not dropped). Error bars display 95 percent block bootstrapped confidence intervals.

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 6.7% in order to reduce the rate of search among the remaining observations from 7.8% to 1.7%. The catch method yields less benefit at a higher cost. After dropping the 17.8% of respondents who answered the catch question correctly, search would still affect 3.1% of the remaining observations. If the paradata method is already in place, the catch method reduces search by another 1.1% of the remaining data (from 1.7% to 0.6%) at a marginal cost of 13.8% of the initial observations (from 6.7% to 20.5% missing). The results of Studies 2 and 3 are similar but shifted upward in magnitude and reflective of Study 2's relatively under-sensitive catch questions.

The right column of Figure 5 examines the same trade-off 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 cuts search by about three-quarters regardless of whether a pledge is present (from 3.8% to 0.9% with a pledge versus 7.8% to 1.7% without one).

The Study 3 results include an apparent contradiction: despite being more sensitive than the paradata (Figure 3), the catch method leaves a slightly larger percentage of search among unflagged observations (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 its impressive external sensitivity. Even when the catch method detects a larger proportion of search, using it to purge data of search leaves behind fewer observations that are at least as affected by search on a per-unit basis.

The combination of the pledge, paradata, and catch methods constitutes a remarkably efficient 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% to 0.4% of observations at a cost of converting 10.3% of observations to missing. In Study 2, reducing search from 17.6% to 0.5% costs 19.0% of the data. In Study 3, reducing search from 11.6% to 0.4% costs 11.2% 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 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% of the remaining observations (from 2.7 to 0.9) at a cost of 5.2% 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% of the data (from 0.7 to 0.1) at a cost of 8.0% 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 bit more search.

Heterogeneous Effects

Researchers deciding how to address information search must also consider how the effects of these strategies vary with respondent characteristics. This section presents an exploratory analysis of how the detection and deterrence methods shape representativeness and between-group differences in search. Each method is examined in isolation, focusing on the four scenarios labeled "nothing," "pledge 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. This quantity's implications for deterrence and detection differ based on their respective *ex ante* and *ex post* natures. Because deterrence eliminates 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, deterrence 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 entails a zero-sum trade-off between dropping observations and reducing between-group differences in 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.

To examine how these factors play out in practice, the sample was split according to 10 pretreatment variables 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 one standard deviation above and below the mean.[8]

At baseline, the data reflect several between-group differences in search (Figure 6 and difference in means tests in Appendix Table A7, page A14).[9] Several demographic differences exist: younger respondents are more likely to search than older respondents (+8.5 p.p.), men are more likely to search than women (+1.3 p.p.), and non-white and Hispanic respondents are more likely to search than their counterparts (+6.9 and +7.9 p.p.). Other attitudes and traits also predict search. Search is more common among respondents who are more educated (+4.3 p.p.), endorse more conspiratorial beliefs (+8.1 p.p.), are more interested in politics (+0.7 p.p.), prefer the Democratic party (+1.2 p.p.), and are stronger partisans (+2.4 p.p.). In some cases, these differences in search behavior reinforce established differences in measured knowledge (such as gender and interest in politics), while in other cases they counteract such differences (for example, age and conspiracy beliefs).

The detection methods (catch and paradata) have fairly even effects across groups. For every characteristic, the difference in search prevalence among unflagged respondents is smaller than the baseline difference. This means that researchers who treat detected respondents as missing data succeed to some extent in reducing between-group differences in search. However, because groups that are more likely to search at baseline are disproportionately dropped, there is a corresponding cost in terms of sample representativeness.

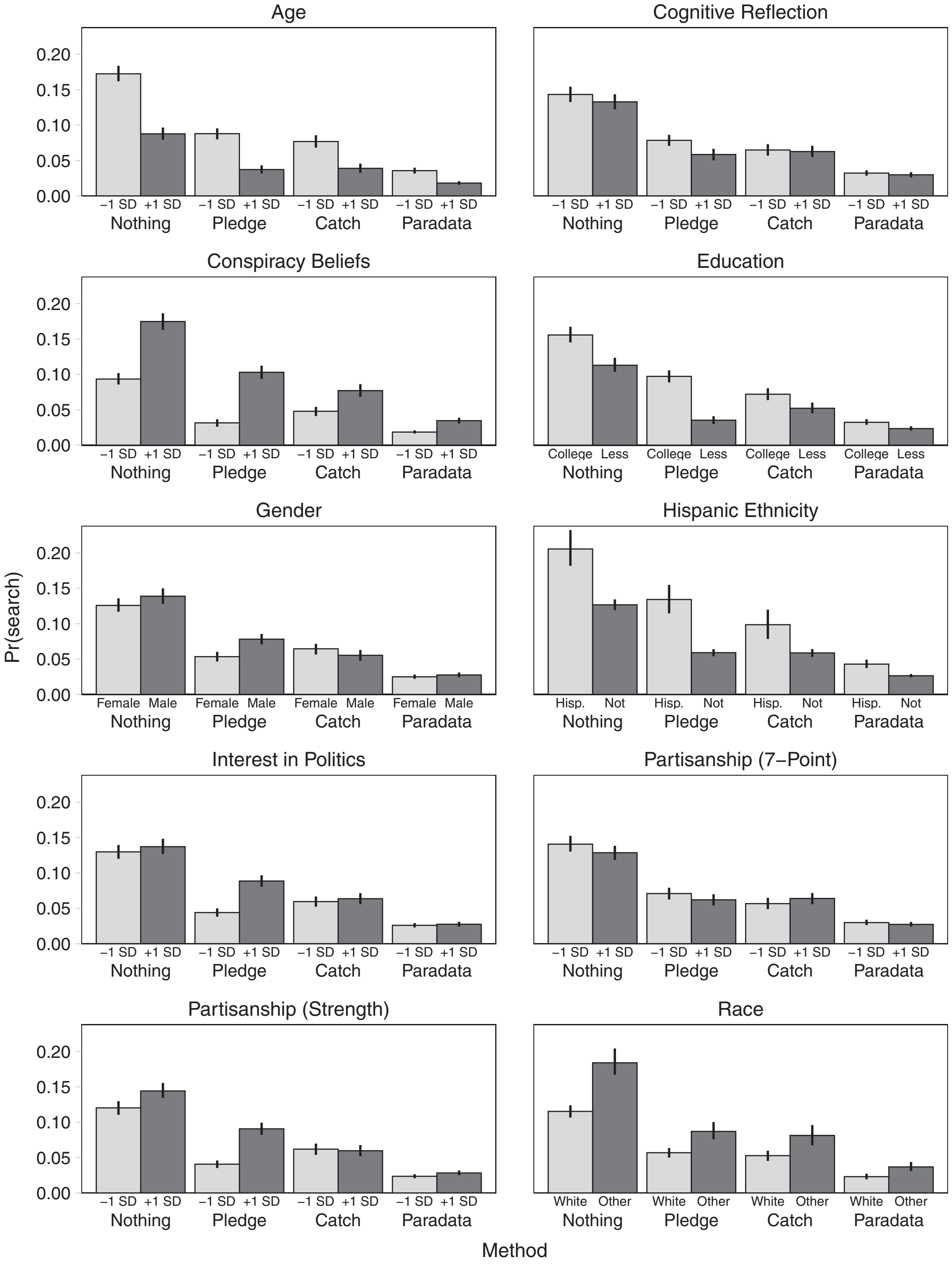
The deterrence method—the pledge—has mixed success at reducing between-group differences in the prevalence of search. In particular, respondents who are more educated, more interested in politics, and more partisan are both more likely to search *and* less deterred by the pledge than their counterparts. College graduates search 4.3 p.p. more often at baseline and 6.2 p.p. more with a pledge (difference = 1.9, s.e. = 0.9; see Appendix Table A7, page A14). High-interest respondents search just 0.7 p.p. more at baseline but 4.5 p.p. more with a pledge (difference = 3.7, s.e. = 0.8). Stronger partisans search 2.4 p.p. more at baseline and 5.0 p.p. more with a pledge (difference = 2.6, s.e. = 0.8). In two more cases (gender and cognitive reflection), statistically insignificant estimates also suggest larger differences with a 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.

The finding that deterrence can increase between-group differences in search stands out in the context of existing research. For different reasons, the tendency for the interested, the educated, the more partisan, and men to score better on political knowledge batteries are all well established. On the positive side, these subgroups' resistance to 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 quiz questions about politics. On the negative side, evaluations often treat a stronger relationship in the expected direction as evidence of improved validity (Marquis 2021; Smith, Clifford, and Jerit 2020). This is not necessarily the case. At baseline, search may either reinforce or attenuate between-group differences in knowledge, and deterrence may either counteract or reinforce the baseline

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

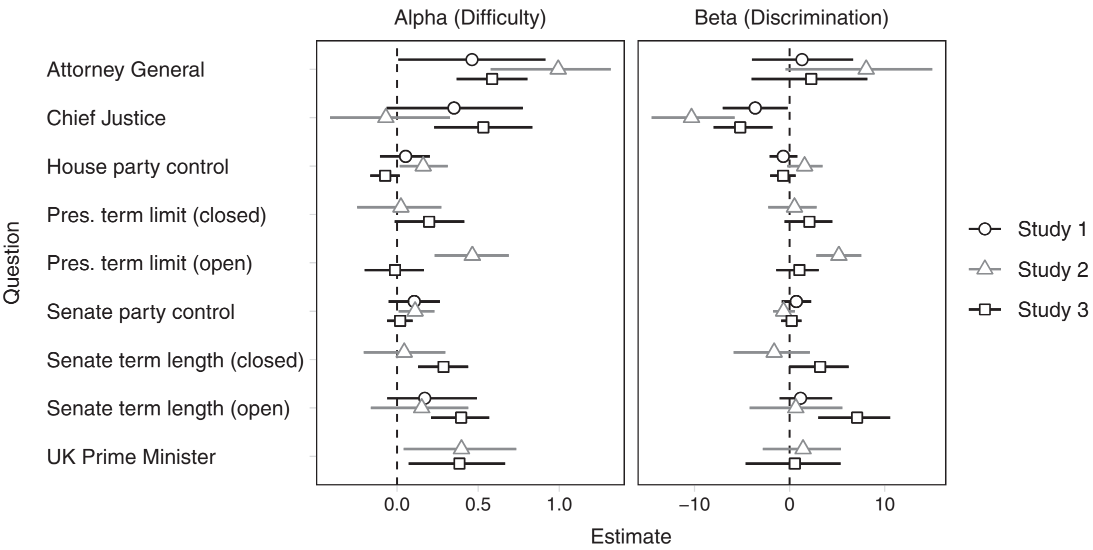
[9] As all of this section's results are consistent across studies, the data are pooled for simplicity.

FIGURE 6 Estimated Rate of Search, by Mitigation Strategy and Respondent Characteristic



Notes: Figure displays the effect of anti-search strategies on between-group differences in the rate of search. Each point on the x-axis is a strategy or combination of strategies for dealing with search. The y-axis displays the proportion of search in the remaining data (the same quantity as the grey dots and lines in Figure 5). Error bars display 95 percent confidence intervals. A table of estimates appears in the appendix (Table A.7, page A14).

FIGURE 7 Effect of Pledge on IRT Estimates



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

difference. This complicates assessments of how search affects construct validity.

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, reducing the rate of search in all subgroups to 3% or less. Catch questions add a bit of marginal value, further reducing this figure to 2% at a high cost in terms of missing data.

Implications

Information search looms as a threat to the validity of any survey measure with answers that can easily be looked up. This article shows that through a combination of detection and deterrence, researchers can manage this threat. This section highlights some lessons for practice.

As a starting point, question-level estimates of the prevalence of search are essential. 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 than indicated here or no threat at all. When search is common, researchers are likely to benefit from proven, multi-method strategies that include both deterrence and detection. But when search is rare, researchers may be satisfied with a relatively light intervention or no intervention at all.

Deterrence is a researcher's first line of defense against search. Its *ex ante* nature avoids most of the downsides of *ex post* strategies for dealing with search, for example, dropping suspected searchers as a robustness check. At the cost of one screen of survey space, the pledge tested here reduced search by 50%. Despite this, three shortcomings make deterrence an incomplete solution. First, it does not completely eliminate search. Especially for questions with high base rates of search, a substantial amount of search still occurs. Second, even though the pledge brings down the overall rate of search, it exacerbates differences in search between some subgroups. Third, on their own, deterrence methods 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 article considered the two that offer the best combination of cost and credibility: catch questions and paradata.[10] 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).

[10] 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.

Relative to paradata, 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 it *ex post*.

Evaluations of the measurement properties of knowledge scales also stand to benefit from item-level paradata detection methods.[11] To concretize these benefits, consider the effect of information search on the construction of knowledge scales. Though a full analysis is beyond this article'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 that the pledge matters 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. 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). Results that might be written off as statistical noise instead look 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 reasonably efficiently. 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 (compare Tables 4 and 5)—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.

Due to their complementary nature, combining detection and deterrence allows 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 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 article's findings suggest that reasonable solutions to the information search problem are in our grasp and that even more complete solutions are within reach. This is encouraging for online survey measures of knowledge and beliefs in all areas of research.

References

Ahler, Douglas, and Guarav Sood. 2018. "The Parties in Our Heads: Misperceptions about Party Composition and Their Consequences." *The Journal of Politics* 80(3):964–81.

Ansolabehere, Stephen, and Philip Jones. 2010. "Constituents' Responses to Congressional Roll-Call Voting." *American Journal of Political Science* 54(3):583–97.

Berinsky, Adam, Gregory Huber, and Gabriel Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk." *Political Analysis* 20(3):351–68.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics* 119(1):249–75.

Bryson, Bethany. 2020. "When Survey Respondents Cheat: Internet Exposure and Ideological Consistency in the United States." *International Journal of Communication* 14:5351–74.

Bullock, John, Alan Gerber, Seth Hill, and Gregory 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):120–31.

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):858–87.

[11] Another implication for scale validity is discussed near the end of the "Heterogeneous Effects" section.

Cooper, Emily, 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):146–53.

Diedenhofen, Birk, and Jochen Musch. 2017. "PageFocus: Using Paradata to Detect and Prevent Cheating on Online Achievement Tests." *Behavior Research Methods* 49(4):1444–59.

Domnich, Alexander, Donatella Panatto, Alessio Signori, Nicola 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.

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):143–62.

Graham, Matthew. 2020. "Self-Awareness of Political Knowledge." *Political Behavior* 42(1):305–26.

Gummer, Tobias, and Tanja Kunz. 2019. "Relying on External Information Sources When Answering Knowledge Questions in Web Surveys." *Sociological Methods and Research* 51(2):816–36.

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

Höhne, Jan Karem, Carina Cornesse, Stephan Schlosser, Mick Couper, and Annelies Blom. 2021. "Looking up Answers to Political Knowledge Questions in Web Surveys." *Public Opinion Quarterly* 84(4):986–99.

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

Jensen, Carsten, and Jens Thomsen. 2014. "Self-Reported Cheating in Web Surveys on Political Knowledge." *Quality and Quantity* 48(6):3343–54.

Krosnick, Jon A. 1991. "Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys." *Applied Cognitive Psychology* 5(3):213–36.

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

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.

Motta, Matthew, Timothy 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.

Permut, Stephanie, Matthew Fisher, and Daniel Oppenheimer. 2019. "TaskMaster: A Tool for Determining When Subjects Are on Task." *Advances in Methods and Practices in Psychological Science* 2(2):188–96.

Peyton, Kyle, Gregory Huber, and Alexander Coppock. 2022. "The Generalizability of Online Experiments Conducted During The COVID-19 Pandemic." *Journal of Experimental Political Science* 9(3):379–94.

Shulman, Hillary, and Franklin Boster. 2014. "Effect of Test-Taking Venue and Response Format on Political Knowledge Tests." *Communication Methods and Measures* 8(3):177–89.

Smith, Brianna, Scott Clifford, and Jennifer Jerit. 2020. "How Internet Search Undermines the Validity of Political Knowledge Measures." *Political Research Quarterly* 73(1):141–55.

Starratt, Gerene, 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):175–92.

Style, Hillary, and Jennifer Jerit. 2021. "Does it Matter if Respondents Look up Answers to Political Knowledge Questions?" *Public Opinion Quarterly* 84(3):760–75.

Ternovski, John, and Lilla Orr. 2022. "A Note on Increases in Inattentive Online Survey-Takers Since 2020." Journal of Quantitative Description: Digitial Media 2.

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):251–65.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Supplemental Results
Appendix B: Study Information