How to detect and influence looking up answers to political knowledge questions in web surveys

Tobias Gummer
GESIS – Leibniz Institute for the Social Sciences (Germany)

Tanja Kunz
GESIS – Leibniz Institute for the Social Sciences (Germany)

Tobias Rettig
University of Mannheim (Germany)

Jan Karem Höhne
Leibniz University Hannover (Germany)
German Centre for Higher Education Research and Science Studies (Germany)

Abstract

When answering political knowledge questions in web surveys, respondents can look up the correct answer on the Internet. This response behavior artificially inflates political knowledge scores that are supposed to measure fact-based information. In the present study, we address the existing knowledge gaps of previous research regarding looking up answers to political knowledge questions in web surveys. We conducted an experimental study based on the German Internet Panel, a large-scale population survey that uses a probability-based sample. Based on this experiment, we show that instructions help to reduce the number of lookups to knowledge questions in web surveys. We provide further evidence that looking up answers results in more correct answers to knowledge questions and, thus, in inflated political knowledge scores. Finally, our findings illustrate the challenges and benefits of using self-reported or paradata-based lookup measures as well as a combined measure that aims at utilizing both to detect lookups to political knowledge questions in web surveys.

Keywords: web surveys, political knowledge, instructions, self-reports, paradata, data quality

Introduction

Measures of political knowledge are used in several areas of the social sciences, for example, to determine and explain public and political phenomena. Many social science surveys, such as the American National Election Study (ANES), the Eurobarometer, or the German Longitudinal Election Study (GLES), regularly ask respondents questions about political knowledge. When answering political knowledge questions, researchers usually expect respondents to retrieve the information from their declarative memory that contains fact-based information (Prior & Lupia, 2008). However, if respondents cannot retrieve the relevant information from their declarative

This document is a preprint and thus it may differ from the final version: Gummer, Tobias, Kunz, Tanja, Rettig, Tobias, & Höhne, Jan K. (2023). How to detect and influence looking up answers to political knowledge questions in web surveys. Public Opinion Quarterly. DOI: 10.1093/poq/nfad027.

memory—either because they simply do not know the correct answer, or they are not sufficiently motivated to think about the issue in question—they can draw on their rule-based procedural memory. Procedural memory includes problem-solving strategies to acquire needed information from external sources (Prior & Lupia, 2008). For self-administered web surveys, this means that respondents need only switch browser windows or tabs to look up the correct answers to political knowledge questions on the Internet using search engines such as Google Search or Bing. This response behavior artificially inflates political knowledge scores that are supposed to measure fact-based information.

When asking political knowledge questions, researchers usually want to measure respondents' true fact-based knowledge, unobstructed by other sources. Thus, looking up answers to political knowledge questions in web surveys is frequently referred to as "cheating" (e.g., Clifford & Jerit, 2014; Clifford & Jerit, 2016; Diedenhofen & Musch, 2017; Jensen & Thomsen, 2014; Motta, Callaghan, & Smith, 2017; Shulman & Boster, 2014; Strabac & Aalberg, 2011). Smith, Clifford, and Jerit (2020), for example, showed that looking up answers results in biased data and concluded that this response behavior "reduces the validity of political knowledge measures and undermines the ability to replicate canonical findings in the public opinion literature" (p. 15). In line with these findings, Clifford and Jerit (2014) found that looking up answers lowers the concurrent validity of political knowledge questions (i.e., theoretically expected correlations are decreased). Clifford and Jerit (2016) also provided evidence that preventing respondents from looking up answers can improve the predictive validity of political knowledge questions. Thus, many previous methodological studies have focused on how to prevent respondents from looking up answers by utilizing question design, for example, discouraging instructions (Clifford & Jerit, 2016; Motta et al., 2017; Smith et al., 2020; Vezzoni & Ladini, 2017) or pictures of politicians instead of verbal-only answer options (Munzert & Selb, 2017). Smith et al. (2020) found that respondents who received an instruction not to look up answers were less likely to report looking up answers than respondents who received an instruction that allowed looking up answers. This finding indicates that the respondents complied with the instructions they were given. Since the detection of looking up answers was based primarily on self-reports, social desirability cannot be completely ruled out as an explanation for the difference between the discourage and allow conditions.

Most studies on looking up answers to political knowledge questions in web surveys have relied on self-reports (e.g., Clifford & Jerit, 2014; Clifford & Jerit, 2016; Jensen & Thomsen, 2014). Making an argument that self-report measures might be biased due to social desirability, other studies have used passively and unobtrusively collected client-side paradata on browser window and tab switching (Diedenhofen & Musch, 2017; Gummer & Kunz, 2022; Höhne, Cornesse, Schlosser, Couper, & Blom, 2021), browser history data (Gooch & Vavreck, 2019), or web tracking (Munzert, Ramirez-Ruiz, Barberá, Guess, & Yang, 2022). For example, Gummer and Kunz (2022) found that respondents who switched browser windows and tabs when answering political knowledge questions were more likely to give correct answers. Höhne et al. (2021) replicated this finding, although they reported mismatches between the paradata and selfreports. They found that PC respondents underreported looking up answers, which suggests a social desirability bias in self-reports, whereas smartphone respondents overreported looking up answers. The latter was explained by the fact that smartphone respondents used a different device to look up answers than to complete the web survey due to the small screen size and operating limitations. If answers are looked up on another device, the paradata

measures fail to detect the lookup behavior. In addition, a growing body of literature on multitasking behavior in web surveys (Décieux, 2022; Höhne, Schlosser, Couper, & Blom, 2020; Sendelbah, Vehovar, Slavec, & Petrovčič, 2016) has indicated that switching browser windows and tabs does not necessarily mean looking up answers to a question, since respondents may be engaging in activities unrelated to the survey, such as checking emails and engaging in social networks. In summary, both self-report and paradata measures may be underestimating lookup behavior. Furthermore, paradata measures run the risk of mistaking multitasking for lookup behavior, thereby overestimating the incidence of looking up answers. Consequently, it is unclear which measure is more valid for detecting looking up answers to political knowledge questions in web surveys.

The present study complements previous research on looking up answers to political knowledge questions in web surveys. We address existing research gaps by employing a comprehensive experimental design that includes a control group for establishing a baseline rate for looking up answers, and three treatment groups that systematically vary instructions on looking up answers to political knowledge questions in web surveys. In our analyses, we used self-report and paradata measures, as well as a combination of the two, to detect the looking up of answers. By using data from the German Internet Panel, a large-scale population survey based on a probability-based sample, our study aimed for generalizability to state-of-the-art social science surveys. We addressed the following three research questions:

(1) How do instructions affect respondents' tendencies to look up answers to political knowledge questions in web surveys?

(2) Does looking up answers to political knowledge questions in web surveys increase the likelihood of providing the correct answer?

(3) Should researchers use self-report, paradata, or a combined measure to detect looking up answers to political knowledge questions in web surveys?

In the next section, we introduce the cognitive process of answering and looking up answers to political knowledge questions in web surveys. Then, we describe our experiment, data, and methods. After reporting our results, we close with a discussion including perspectives for future research.

Answering Knowledge Questions in Web Surveys

Tourangeau, Rips, and Rasinski (2000) outlined four steps of the cognitive response process through which respondents progress when they answer a survey question. First, respondents must comprehend the meaning of the question (comprehension), second retrieve relevant information for answering the question (retrieval), third form a judgement based on the retrieved information (judgement), and fourth map their judgement on the response options provided (response). Gummer and Kunz (2022) proposed an extension of this cognitive model that incorporates the following additional steps when respondents search for additional information in external sources to answer a knowledge question in web surveys: (a) decide on an external information source, (b) specify a search query, (c) evaluate the search result, and (d) translate the search result back into the context of the knowledge question (see Figure 1).

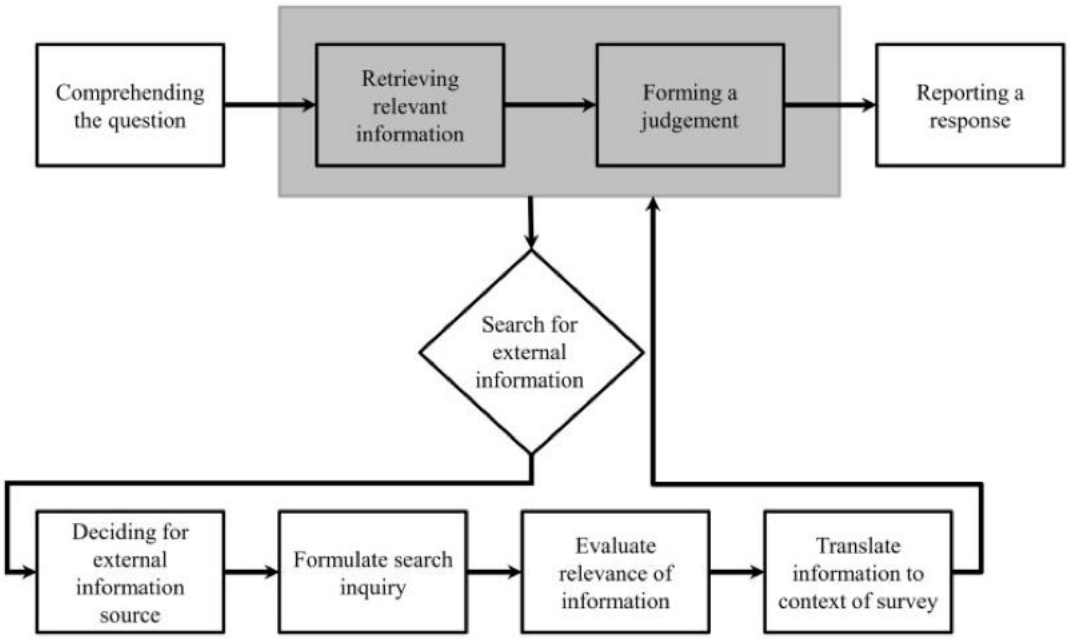

Figure 1. Extension by Gummer and Kunz (2022) to the model of the cognitive response process by Tourangeau, Rips, and Rasinski (2000).

Gummer and Kunz (2022) have referred to completing these additional steps in answering knowledge questions ("going the extra mile") as over-optimizing, as opposed to satisficing, which means the skipping of steps in the cognitive response process by respondents (Krosnick, 1991, 1999). Gummer and Kunz (2022) also found that respondents with higher formal education and a greater interest in the survey topic were more likely to look up answers than respondents with lower education and less interest in the survey topic. In addition, the more difficult knowledge questions were to answer (i.e., open-ended questions were more difficult than closed ones), the more frequently respondents looked up the answers. Style and Jerit (2020) have added to these findings by showing that respondents who looked up answers also reported a higher level of political engagement. These authors also provided further nuance to the findings regarding question design by arguing that the complexity of the answering and search task is moderated by how easily information can be obtained.

When completing web surveys, external information sources are at respondents' fingertips (Sparrow, Liu, & Wegner, 2011). If they decide to progress through additional steps in their response process by using external information sources, they rely on their rule-based procedural memory. Thus, when respondents make a decision to search for external information, progress through the additional cognitive steps of the search task, and process the additional information they obtained in their judgement step, declarative and procedural memory become intertwined, and fact-based knowledge scores become inflated.

Data and Methods

Experimental design and political knowledge questions

We used a between-subjects design and randomly assigned respondents to one of four experimental groups (see Table 1). The experimental groups differed with respect to the instructions respondents received for answering three political knowledge questions. Respondents in the control group were not given instructions on how to answer. Respondents

in the encouragement and discouragement group were shown instructions that either encouraged or discouraged looking up answers. Respondents in the commitment group were presented with a commitment question that explicitly asked them not to look up answers to the political knowledge questions. This question was asked immediately before the presentation of the three political knowledge questions.

Table 1. Experimental design

Experimental group	Instruction	Group size
1	No instruction (control group)	1,106
2	Encouragement to look up answers	1,102
3	Discouragement to look up answers	1,103
4	Commitment to not look up answers	1,102

We adopted three political knowledge questions on the European Union from Höhne et al. (2021). These questions asked for (1) the number of member states of the European Union, (2) the name of the president of the European Commission, and (3) the number of delegates of the European Parliament (see Figure 2). All three questions were open-ended with a text field for entering the answer, and we asked them in random order to minimize any question order effects.

We employed two self-report questions that were asked directly after the three political knowledge questions. The first self-report question asked whether respondents used the Internet to look up the correct answers to the three political knowledge questions. If respondents stated that they looked up the answer to at least one of the three questions, the second self-report question asked whether they looked up the answer on the same device on which they completed the survey or on a different device.

We administered all the questions in German. We employed a mobile-optimized survey layout that generally avoids horizontal scrolling to facilitate survey navigation and operation on smartphones and other small devices.

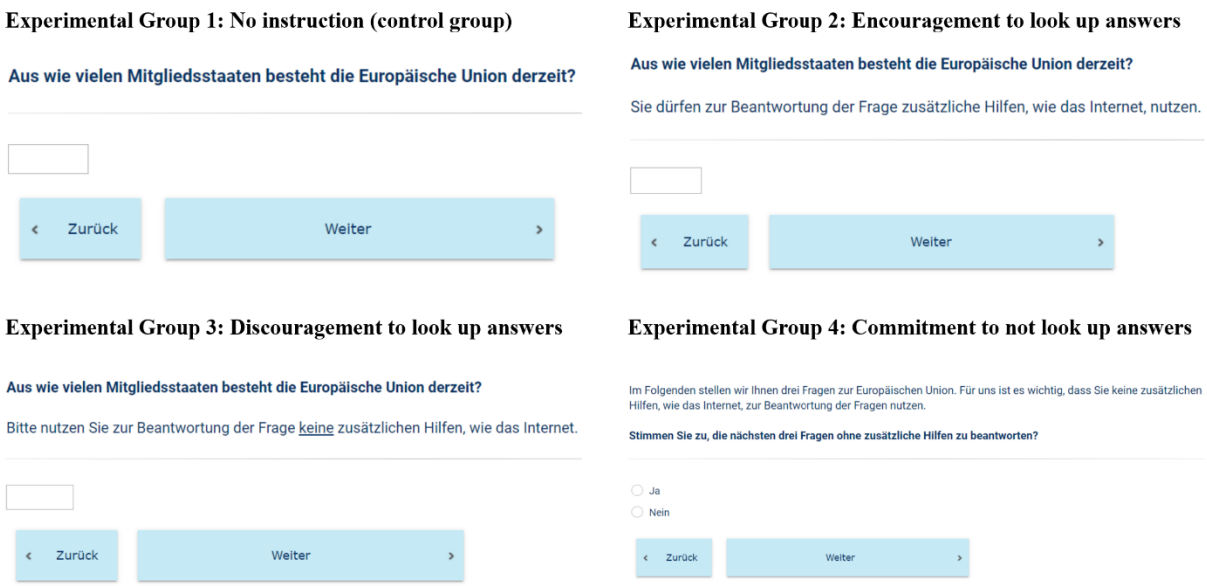

Figure 2. Design of political knowledge questions.

Survey and sample description

We implemented the survey experiment in wave 51 of the German Internet Panel (GIP; Blom et al., 2021), which was based on an initial recruitment in 2012 (cumulative response rate [CUMRR]: 18.5%) and two refreshments in 2014 (CUMRR: 21.0%) and 2018 (CUMRR: 24.1%). While the recruitments in 2012 and 2014 were based on a three-stage stratified probability sample, the recruitment in 2018 was based on a two-stage stratified probability sample of the German population aged 16 to 75 years. Participants of the 2012 and 2014 samples who did not have internet access or an internet enabled device were offered one. For a detailed methodological description of the GIP, we refer interested readers to Blom et al. (2015).

The GIP includes parts of the Embedded Client Side Paradata (ECSP) tool developed by Schlosser and Höhne (2018) that enables the collection of a variety of client-side paradata, including page-level information on browser window and tab switching. For this purpose, ECSP utilizes the JavaScript "OnBlur" function to detect whether respondents switch away from a web survey page.

The survey experiment was included in the middle of the questionnaire of wave 51, which was fielded between January 1 and January 31, 2021. A total of 6,294 respondents were invited to participate in Wave 51. Out of the 4,468 respondents who participated, 39 broke off the survey before or during our experiment, 11 did not provide self-reports, and 5 were excluded because no paradata could be collected. This left us with 4,413 respondents (98.8%) for our statistical analyses. Of these respondents, 20.3% were 35 years or younger, 24.3% were aged between 36 and 50, 24.4% were aged between 51 and 60, and 30.9% were aged 61 years or older. In total, 48.4% of the respondents were female. In terms of education, 12.9% graduated from a lower secondary school (low education level), 29.3% from an intermediate secondary school (medium education level), and 55.4% from a college preparatory secondary school or university (high education level). Furthermore, 0.9% were still attending school or had left school without a diploma, and 1.5% reported having a different degree from those already mentioned. Finally, 34.7% of respondents completed the survey on a smartphone, the other 65.3% used computers or tablets.

To evaluate the effectiveness of the random assignment and the sample composition of the four experimental groups, we conducted χ^2-tests of independence. The results showed no statistically significant differences with respect to age (χ^2=7.11, p=.625), gender (χ^2=3.02, p=.388), education (χ^2=8.19, p=.224), and the device used to answer the survey (χ^2=4.90, p=.556).

Operationalization

For all three political knowledge questions, we coded whether respondents provided the correct answer. For the first question (member states of the European Union—correct answer: 27) and the third question (delegates of the European Parliament—correct answer: 705), this was relatively simple because respondents were restricted to entering numbers in the open-ended text field. For the second question (president of the European Commission—correct answer: Ursula von der Leyen), respondents were supposed to provide a name. Some respondents provided a correct, but misspelled name (e.g., "Ursula van der Laien"). We decided to count such misspellings as correct answers because we were interested in respondents' political knowledge rather than in their writing or typing skills. We coded deviating answers (e.g., "Kim

Kardashian") (n=64) and nonresponses (n=304) as incorrect. Finally, we added up the number of correct answers for each respondent (values ranging between 0 and 3). We also created a dummy variable based on all correct answers, indicating correct but misspelled answers to the second political knowledge question (0=no, 1=yes).

Based on the first self-report question, we determined how many answers respondents looked up online (values ranging between 0 and 3). Based on the second self-report question, we determined whether respondents looked up the answers on the same device on which they completed the survey or on a different device.

We used the paradata collected by the ECSP tool to assess whether respondents left a web survey page to look up answers online. Since all three political knowledge questions were displayed on separate web survey pages, the page-level paradata enabled us to determine for which of the three political knowledge questions respondents switched away. Finally, we added up the number of switching events (or lookups; values ranging between 0 and 3).

As a combined measure of self-report and paradata, we used the higher number of lookups from either self-report or paradata. A higher number of lookups detected by paradata may be an indication that respondents underreported the number of lookups (potentially due to social desirability), whereas higher self-reports may indicate that respondents looked up answers on a different device than the one used to complete the survey, or that switching windows and tabs was due to multitasking rather than looking up answers. By design, the combined measure was correlated with the respective individual measures based on self-reports and paradata.

Analytical Strategy

In the present study, we refer to effects with p-values <.05 as statistically significant. If not stated otherwise, all tests were computed as two-tailed. Since the GIP does not provide design weights, we conducted all our analyses based on unweighted data.

RQ1: With respect to our first research question, we compared the number of lookups across experimental groups and used one-way ANOVAs to test for differences for each lookup measure. Further, we fitted negative binomial regression models with the number of lookups as the dependent variable, and dummy variables for each experimental group as independent variables, with the control group serving as reference. Following Gummer and Kunz (2022) and Höhne et al. (2021), we included the following control variables: age (1 = 35 years and younger [reference category], 2 = 36–50 years, 3 = 51–60 years, 4 = 61 years and older), gender (1 = male [reference category], 2 = female), education (1 = low [reference category], 2 = intermediate, 3 = high), and survey evaluation (rated as interesting, varied, relevant, long, difficult, too personal on a scale of 1 "not at all" to 4 "very much"). For each of the three lookup measures (self-report, paradata, and combined), we fitted a separate regression model. We chose to use negative binomial regressions instead of ordinary least squares regressions because the number of lookups was coded as a count variable. To assess whether differences between our experimental groups were statistically significant, we computed additional Wald-tests (not included in the table).

To gain further insights into how instructions affected respondents' lookup behavior, we investigated response times across our experimental groups and the number of lookups using server-side timestamp paradata. We computed the time required to answer the three political knowledge questions. Due to the susceptibility of response times for outliers, we excluded times above 1200s (20min). We conducted one-way ANOVAs to test for differences in mean response

times.

RQ2: To answer our second research question, we again fitted negative binomial regression models separately for the three lookup measures, with the number of correct answers as the dependent variable. With respect to independent variables, we included the number of lookups and dummy variables for each experimental group, with the control group serving as reference. Similar to RQ1, we included age, gender, education, and survey evaluation as control variables. In additional models, we included interaction terms between the number of lookups and the experimental groups to test whether the number of lookups had a smaller or larger effect on the number of correct answers in one experimental group compared to another. To make the regression results more accessible to readers, we also present predicted values in our description of outcomes.

We reran the regression models used for subgroups of respondents to investigate any group-specific effects of different instructions. We computed the models separately by gender (female, male) and education (low, intermediate, high), since previous research had suggested differences in political knowledge scores for these groups (e.g., Jerit, 2009; Prior, 2014). For the sake of simplicity, we created a coefficient plot visualizing 18 group-specific effects, depending on the different experimental groups.

RQ3: With respect to our third research question, we conducted different types of analyses. First, we compared the number of lookups each measure detected. Following studies such as Höhne et al. (2021), we found it reasonable to assume that if respondents tended to underreport looking up answers due to social desirability, detecting more lookups generally indicates a better measure. Second, we compared the predictive power of self-report, paradata, and the combined measure with respect to the prediction of correct answers. Thus, we relied on the regression models used to answer RQ2 and compared the model fit indicators between the three measures (Pseudo-R^2, Akaike Information Criterion [AIC], and Bayes Information Criterion [BIC]). For additional insights, we replicated the comparison of the model fit indicators separately for each experimental group. Third, we utilized the second self-report question, asking respondents on which device they looked up answers, to learn more about whether paradata measures can adequately capture lookup behavior because they are limited to search behavior on the same device. Fourth, we investigated misspelling the correct open-ended answer to our second political knowledge question to learn more about whether lookup behavior also may indicate something different than searching for an unknown answer, such as knowing the correct answer but looking up the correct spelling.

Results

RQ1: Effects of instructions on lookup behavior

Table 2 details the differences in the number of lookups to the three political knowledge questions, depending on the different experimental groups. Our findings show that instructions did affect lookup behavior (self-reports: $F(3)=962.87$, $p=.000$; paradata: $F(3)=320.53$, $p=.000$; combined: $F(3)=607.81$, $p=.000$). Most prominently, respondents who were encouraged to look up answers had the most lookups based on all three lookup measures.

Drawing on respondents who were not given any instructions that encouraged or discouraged looking up answers, we established a baseline rate for looking up answers in our web survey. The share of control group respondents who looked up answers to the three political knowledge questions were 50.5% (self-reports), 43.5% (paradata), and 59.0% (combined).

Table 2. Number of lookups for each measure across experimental groups.

	Number of lookups (%)					
	0	1	2	3	Mean no. of lookups (0 included)	Mean no. of correct answers (0 included, max. 3)
Control						1.5
Self-report	49.50	28.30	14.60	7.70	0.8	
Paradata	56.50	11.80	10.50	21.20	1.0	
Combined	41.00	19.90	15.00	24.10	1.2	
Encouragement						2.2
Self-report	16.50	22.10	31.60	29.80	1.7	
Paradata	35.00	9.50	12.40	43.00	1.6	
Combined	14.30	13.30	20.40	51.90	2.1	
Discouragement						1.0
Self-report	82.80	13.10	2.70	1.40	0.2	
Paradata	74.40	11.30	5.70	8.50	0.5	
Combined	66.90	16.40	7.20	9.50	0.6	
Commitment						1.0
Self-report	89.80	8.30	1.10	0.80	0.1	
Paradata	82.90	7.20	5.10	4.80	0.3	
Combined	78.00	11.30	5.40	5.40	0.4	

Note. Percentages do not add up to 100% due to rounding.

As Table 3 shows, the differences in the number of lookups between the experimental groups remained, even when controlling for gender, age, education, and survey evaluation. For all three lookup measures, we found that respondents who received an encouraging instruction were most likely to look up answers as indicated by the highest mean number of lookups compared to the other experimental groups (all differences significant with p=.000). In contrast, an instruction that discouraged respondents from looking up answers resulted in a lower mean number of lookups compared to an encouraging instruction or no instruction at all (both differences significant with p=.000). We further found that asking for respondents' commitment to not look up answers lowered the mean number of lookups even more than a discouraging instruction, and yielded the lowest mean number of lookups among all the experimental groups (all differences significant with p=.000).

In addition, we found that based on all three lookup measures, respondents aged 60 or older were less likely to look up answers compared to respondents aged 35 or younger. In contrast, the effects of gender and education differed depending on which lookup measure was used. We found that female respondents were more likely to self-report looking up answers compared to males. This effect was not detected when using paradata. Conversely, we found no differences between lower- and higher-educated respondents regarding how they self-reported lookups. In addition, the paradata showed that higher-educated respondents were more likely to look up answers compared to lower-educated respondents. Using the combined measure seemed to incorporate the potentials of both measures: we found that both female and higher-educated respondents were more likely to look up answers to political knowledge questions.

Table 3. Negative binomial regressions of number of lookups to three political knowledge questions.

	Self-report			Paradata			Combined		
Parameter	coef	SE	p	coef	SE	p	coef	SE	p
Experimental group (ref.: Control)									
Encouraged	0.782	0.041	0.000	0.522	0.053	0.000	0.537	0.035	0.000
Discouraged	-1.264	0.072	0.000	-0.718	0.065	0.000	-0.736	0.048	0.000
Commitment	-1.836	0.091	0.000	-1.123	0.072	0.000	-1.181	0.056	0.000
Age (ref.: 35 and under)									
36–50	-0.023	0.052	0.655	-0.114	0.060	0.057	-0.074	0.041	0.073
51–60	-0.053	0.052	0.310	-0.354	0.063	0.000	-0.159	0.042	0.000
Over 60	-0.291	0.053	0.000	-0.855	0.065	0.000	-0.478	0.044	0.000
Female	0.190	0.037	0.000	-0.018	0.044	0.678	0.090	0.030	0.003
Education (ref.: low)									
Intermediate	0.060	0.060	0.317	0.081	0.077	0.292	0.066	0.051	0.198
High	0.033	0.057	0.569	0.219	0.071	0.002	0.106	0.048	0.028
Survey Evaluation									
Interesting	-0.011	0.036	0.760	0.016	0.044	0.711	0.031	0.030	0.290
Varied	0.070	0.034	0.041	0.028	0.041	0.506	0.028	0.028	0.324
Relevant	0.012	0.032	0.710	0.089	0.038	0.020	0.044	0.026	0.089
Long	-0.002	0.024	0.920	0.017	0.029	0.554	0.020	0.020	0.304
Difficult	0.080	0.025	0.001	-0.010	0.031	0.755	0.034	0.021	0.105
Too personal	0.009	0.024	0.716	0.016	0.030	0.595	0.024	0.020	0.238
Constant	-0.649	0.149	0.000	-0.293	0.182	0.108	-0.193	0.123	0.115
Pseudo-$R^2_{\text{McFadden Adj.}}$	.207			.079			.120		
AIC	8,032			10,086			10,891		
BIC	8,140			10,194			10,999		
LR(15)	2,133		0.000	894		0.000	1,524		0.000
Observations	4,318			4,318			4,318		

Note. Coef=regression coefficient, SE=standard error, p=p-value; ref=reference category.

Table 4 shows the mean response time to the three political knowledge questions, depending on the different experimental groups and across the number of lookups. We consistently found that looking up answers resulted in higher mean response times compared to not looking up answers. Considering that our experimental groups differed in the number of lookups, it is not particularly surprising that mean response times also significantly varied between experimental groups ($F(3)=130.81$, $p=.000$). The commitment group had the lowest mean response time, the discouragement and control group took intermediate positions, and the encouragement group had the highest mean response time.

Table 4. Mean response times (in seconds) to three political knowledge questions by experimental groups and number of lookups.

Experimental group	Number of lookups				Overall mean response time
	0	1	2	3	
Self-report					
Control	81	174	186	219	133
Encouragement	90	206	206	211	188
Discouragement	104	182	166	187	119
Commitment	88	174	181	178	96
Paradata					
Control	112	130	153	179	133
Encouragement	172	213	200	193	188
Discouragement	105	143	176	170	119
Commitment	85	139	152	172	96
Combined					
Control	73	159	172	189	133
Encouragement	80	195	222	203	188
Discouragement	93	162	199	172	119
Commitment	79	155	153	172	96

Note. Response times over 1200s (20min) excluded from analysis.

RQ2: Prediction of correct answers to political knowledge questions

As previously shown in Table 2, instructions affected the number of lookups, and consequently, the number of correct answers to the political knowledge questions ($F(3)=458.14$, $p=.000$). Accordingly, respondents who were encouraged to look up answers gave the most correct answers compared to the other groups. We replicated this finding with our negative binomial regression models on correctly answering the three political knowledge questions (see Table 5). For all three lookup measures, these models revealed that the number of lookups was positively associated with the number of correct answers given by respondents. This finding indicates that looking up the answers to political knowledge questions in web surveys increases the respondents' likelihood of answering them correctly.

Based on the regression models, we predicted the number of correct answers for 0 lookups at between 1.0 (self-report, combined) and 1.1 correct answers per respondent (paradata). As the number of lookups increased, so did the predicted number of correct answers. For one lookup, we predicted between 1.3 (combined) and 1.5 correct answers (self-report); for two lookups, between 1.8 (paradata, combined) and 2.2 correct answers (self-report); and for three lookups, the models indicated between 2.3 (paradata) and 2.8 correct answers (self-report). These findings suggest that the self-report measure was more closely related to correct answers than either the paradata or the combined measure.

Even though we included the number of lookups and additional control variables in our regression models, the effects of the experimental groups on the number of correct answers remained significant in all three models. If the differences in looking up answers induced in the experimental groups had been fully captured by the lookup measures (i.e., the number of lookups variable), the experimental groups should have had no effect in the models controlling

for number of lookups. The persistence of this effect suggests that additional lookup behavior occurred that was not detected by the self-report, paradata, or combined measure, and/or that the instructions elicited different kinds of response behavior that varied in how well the lookup measures captured it.

Table 5. Negative binomial regressions of correct answers to three political knowledge questions.

	Self-report			Paradata			Combined		
	Coef	SE	p	Coef	SE	p	Coef	SE	p
No. of lookups	0.276	0.015	0.000	0.171	0.011	0.000	0.252	0.012	0.000
Experimental group (ref.: Control)									
Encouragement	0.115	0.035	0.001	0.266	0.033	0.000	0.173	0.034	0.000
Discouragement	-0.214	0.041	0.000	-0.302	0.040	0.000	-0.220	0.040	0.000
Commitment	-0.151	0.041	0.000	-0.241	0.040	0.000	-0.130	0.041	0.001
Age (ref.: 35 and under)									
36–50	0.005	0.039	0.905	0.021	0.039	0.594	0.018	0.039	0.635
51–60	0.073	0.039	0.060	0.125	0.039	0.001	0.106	0.039	0.006
Over 60	0.150	0.038	0.000	0.214	0.039	0.000	0.216	0.038	0.000
Female	-0.088	0.026	0.001	-0.042	0.026	0.105	-0.069	0.026	0.007
Education (ref.: low)									
Intermediate	0.057	0.043	0.182	0.066	0.043	0.123	0.053	0.043	0.214
High	0.136	0.040	0.001	0.116	0.040	0.004	0.115	0.040	0.004
Survey Evaluation									
Interesting	-0.003	0.026	0.907	-0.008	0.026	0.765	-0.016	0.026	0.536
Varied	0.014	0.025	0.578	0.025	0.024	0.314	0.025	0.025	0.314
Relevant	0.029	0.023	0.203	0.020	0.022	0.385	0.019	0.023	0.394
Long	0.003	0.017	0.880	0.000	0.017	0.979	-0.003	0.017	0.884
Difficult	-0.065	0.018	0.000	-0.041	0.018	0.022	-0.053	0.018	0.003
Too personal	0.027	0.017	0.120	0.026	0.017	0.128	0.021	0.017	0.226
Constant	-0.008	0.106	0.937	-0.025	0.106	0.816	-0.105	0.107	0.326
Pseudo-$R^2_{\text{McFadden Adj.}}$	.083			.073			.089		
AIC	11,253			11,374			11,179		
BIC	11,362			11,482			11,287		
LR(16)	1,053		0.000	932		0.000	1,128		0.000
Observations	4,318			4,318			4,318		

Note. Coef=regression coefficient, SE=standard error p=p-value; ref=reference category.

Therefore, we investigated whether the effects of the number of lookups differed by experimental group. For this purpose, we added interaction effects between the number of lookups and the experimental groups (see Figure 3). We found that a discouraging instruction and commitment question reduced the effect of the number of lookups on correct answers compared to the control and encouragement groups. The lookup measures differed in how well they captured these interaction effects. The paradata and combined measures reflected this

pattern most clearly, whereas the self-report measure yielded smaller differences. Thus, while the paradata measure was not closely associated with looking up the correct answer for the discouragement and commitment groups, it was more closely associated with the correct answers for the encouragement and control groups.

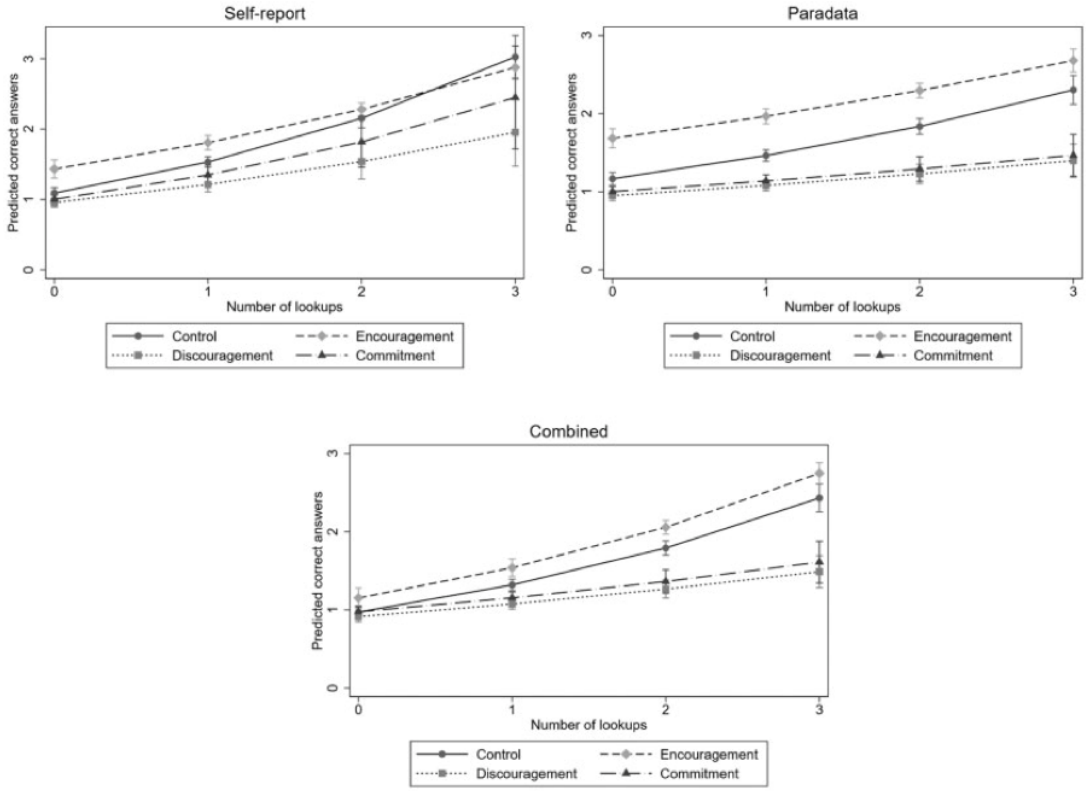

Figure 3. Interactions between number of lookups and experimental groups on number of correct answers to three political knowledge questions. Lines indicate 95 percent confidence intervals.

Our finding that an effect of the experimental groups remained even after controlling for the number of lookups (and interactions) lead us to conclude that the instructions affected lookup behavior differently. In other words, depending on the instruction, respondents engaged in different response behaviors that were not captured perfectly by our lookup measures. For example, respondents may have complied with the discouraging instruction, and a higher share of window switching occurred that was not associated with looking up answers, which negatively affected the paradata measure. This evidence adds to our initial assumption that selfreport and paradata measures both have weaknesses, and neither captures lookups perfectly.

Figure 4 details our supplemental analyses on correct answers to our three political knowledge questions, separately for groups of respondents. We found no significant differences in the effects of the instructions between female and male respondents, nor between the different levels of education. However, across all three lookup measures, encouragement and commitment had significant effects among respondents with medium and high levels of education, but not among respondents with low levels of education. With all necessary caution, this finding might be a first indication that the effects of instructions can vary between groups of respondents.

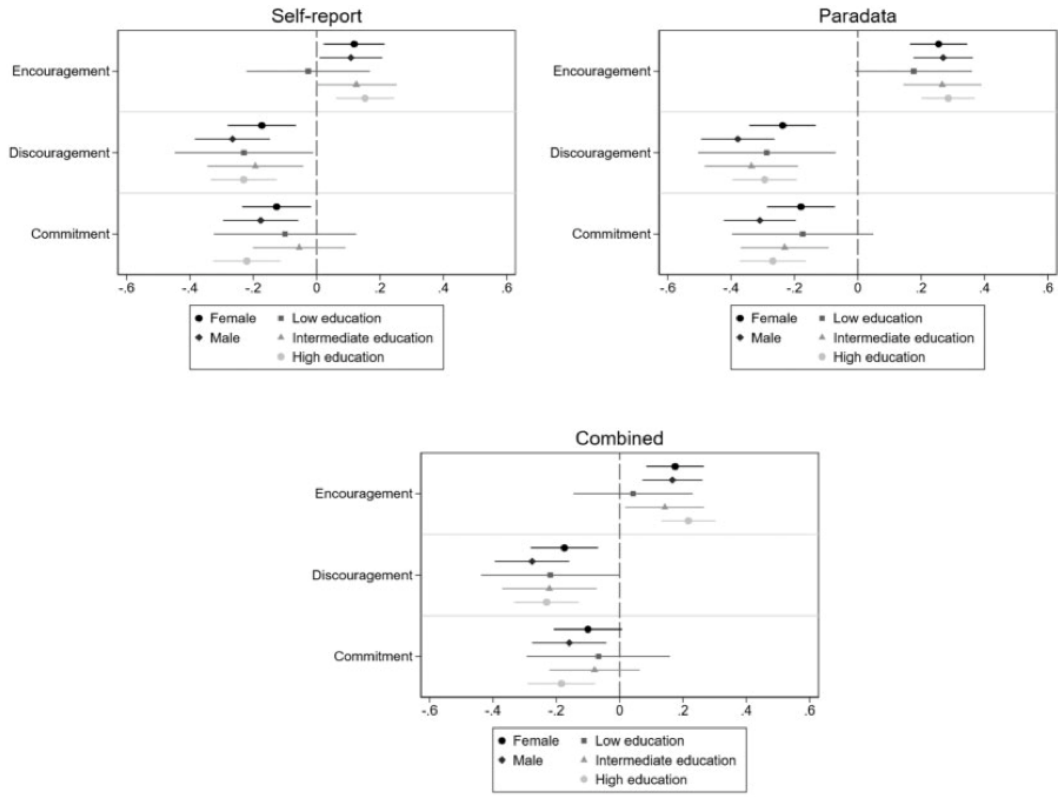

Figure 4. Plots of effects of instructions on correct answers to three political knowledge questions for subgroups (gender, education). Lines indicate 95 per- cent confidence intervals.

RQ3: Comparison of different lookup measures

Turning to our third research question, we investigated which lookup measure was most appropriate for identifying looking up answers to political knowledge questions in web surveys. We followed the reasoning that it is uncomfortable or socially undesirable to admit that one does not know the correct answer to a political knowledge question and had to look it up. Particularly in the third and fourth experimental groups (discouragement and commitment), in which respondents were instructed or asked for their commitment to not look up the answer, doing so anyway represented socially undesirable behavior. Thus, detecting more lookups was taken as an indication of an improved lookup measure.

In line with our reasoning of a downward bias in self-reports due to social desirability, we found that paradata detected a higher mean number of lookups than self-reports in the control group and the discouragement and commitment groups (see Table 2). However, the higher number of lookups measured by paradata in these experimental groups can also be due to multitasking behavior such as checking emails and engaging in social networks rather than lookup behavior. Consistent with our reasoning about social desirability are the results for the encouragement group. In this case, paradata and self-reports yielded a rather similar number of lookups, with self-reports slightly higher than paradata (see Table 2). Thus, when the encouragement group looked up the correct answers and reported this behavior when asked, they considered their behavior socially desirable or at least socially neutral. In each experimental group, the combination of self-reports and paradata resulted in detecting the highest number of lookups.

When comparing the self-report, paradata, and combined measures with respect to predicting correct answers to the three political knowledge questions (see Table 6), we found

that a combination of self-reports and paradata yielded the best model fit of the three measures with respect to Pseudo-R^2, AIC, and BIC. The paradata measure performed worst in terms of the three model fit indicators. The self-report measure performed better than the paradata measure, but still worse than the combined measure. With separate models for each experimental group, we replicated the same patterns when comparing model fits between lookup measures. Detecting more lookups with the paradata measure should have resulted in a model that better explains correct answers. In this regard, it seems that more is—in contrast to our initial assumption—not generally better. Our previous analyses on predicting correct answers (RQ2) add to these findings by showing a closer relationship between the self-report measure and correct answers compared to the paradata and combined measures. Taking both findings together, the appropriateness of the paradata and the combined measure for identifying lookup behavior can be questioned. Paradata performed worst in terms of model fits and relation to correct answers, whereas the combined measure performed best in terms of model fit but was less strongly related to correct answers than the self-report measure.

Table 6. Model fit indicators of negative binomial regressions of correct answers to three political knowledge questions.

Model	Self-report			Paradata			Combined		
	Pseudo -R^2	AIC	BIC	Pseudo -R^2	AIC	BIC	Pseudo -R^2	AIC	BIC
Full model	0.083	11,253	11,362	0.073	11,374	11,482	0.089	11,179	11,287
By experimental group									
Control	0.056	2,889	2,959	0.038	2,941	3,011	0.066	2,857	2,927
Encouragement	0.030	3,237	3,307	0.017	3,278	3,348	0.044	3,188	3,258
Discouragement	0.011	2,578	2,648	0.008	2,584	2,654	0.014	2,570	2,640
Commitment	0.010	2,583	2,653	0.006	2,594	2,664	0.011	2,582	2,652

Note. Model specifications similar to regression models presented in Table 5.

Since paradata only enables capturing looking up answers on the same device as used for survey completion, we also investigated whether respondents looked up the answers to the political knowledge questions on the same device that they used to answer the survey or on a different device. Overall, 69.0% of the respondents reported that they looked up the answers on the same device, whereas 30.9% reported they used another device (0.1% missing answers). This surprisingly high share of respondents who stated using a different device to look up answers may explain why the paradata measure performed worse in predicting correct answers than the self-report and combined measure.

Finally, we analyzed the misspelling of the correct answer to the second political knowledge question. Of all correct answers, 16.2% were misspelled (see Table 7), and this share differed significantly between experimental groups (χ^2=83.94, p=.000). Respondents who were encouraged to look up answers showed the lowest share of misspelled answers (7.9%), whereas respondents in the discouragement (21.0%) and commitment (23.1%) groups yielded the highest shares. In the discouragement and commitment groups, the share of correct answers decreased compared to the control group (by 5.9 and 4.0 percentage points, respectively), while the share of misspellings increased more strongly (by 6.7 and 8.8 percentage points,

respectively). This finding adds more evidence to the notion that a non-neglectable share of respondents looked up answers to validate prior beliefs or check spellings instead of simply looking up the correct answers.

Table 7. Number of correct and correct-but-misspelled answers to an open-ended political knowledge question.

			pairwise χ^2-test (p-value) for misspelled			
Experimental group	Correct answers	*Of these*: misspelled	Control	Encouragement	Discouragement	Commitment
Control	797 (72.1%)	114 (14.3%)	-			
Encouragement	851 (77.2%)	67 (7.9%)	0.000	-		
Discouragement	730 (66.2%)	153 (21.0%)	0.000	0.000	-	
Commitment	754 (68.1%)	174 (23.1%)	0.000	0.000	0.325	-
Overall	3,132 (71.0%)	508 (16.2%)				

Discussion

In this study, we sought to extend previous research on looking up answers to political knowledge questions in web surveys. Prior studies have shown that lookup behavior can inflate political knowledge scores that aim at measuring fact-based knowledge retrieved from declarative memory (e.g., Clifford & Jerit, 2014; Clifford & Jerit, 2016; Höhne et al., 2021). Based on data from the probability-based GIP, we found that a non-ignorable share of respondents looked up answers to political knowledge questions (ranging from 43.5% using the paradata measure to 59.0% using the combined measure). We also found that looking up answers was associated with correctly answering political knowledge questions, even when controlling for respondents' education, survey evaluation, and further background variables. If researchers intend to use knowledge questions to measure fact-based knowledge in web surveys, the knowledge scores will likely be inflated. Therefore, we experimentally tested different instructions that either encouraged or discouraged respondents from looking up answers, finding that respondents did indeed follow these types of instructions. Moreover, we found that asking respondents for their explicit commitment not to look up answers was most effective in reducing the number of lookups to political knowledge questions. Therefore, we recommend the use of discouraging instructions and commitment questions when the aim is to prevent respondents from lookups, and to reduce, at least to some extent, biases in knowledge scores.

Furthermore, our study shows that the different lookup measures that can be used to detect lookups for political knowledge questions in web surveys (i.e., self-report, paradata, or combined measures) resulted in different lookup detection rates. Especially with a discouraging instruction and an explicit commitment, self-reports detected less lookups than the paradata measure. However, self-reports performed better than the paradata measure in predicting correct answers, which suggests that self-reports are a more valid measure for determining lookup behavior in web surveys. When discouraging instructions were used, paradata were only weakly

related to correct answers—presumably because respondents complied with the instruction and paradata measured browser window and tab switching that was unrelated to lookup behavior. Our findings suggest being cautious when using a paradata measure in combination with discouraging instructions.

Self-report and paradata measures both have strengths and weaknesses. While self-report measures may underestimate the frequency of lookups due to social desirability—primarily when capturing potentially undesirable behavior—paradata measures of switching browser windows and tabs may fail to capture lookups because they only capture lookups on the same device that is used for web survey completion. In addition, switching browser windows and tabs during web survey completion may have other purposes that do not necessarily have to do with looking up the correct answer to knowledge questions, such as multitasking (Décieux, 2022; Höhne et al., 2020; Sendelbah et al., 2016).

Considering our results and the limitations associated with lookup measures based on self-reports or paradata, in general, we recommend combining them to detect looking up answers to political knowledge questions in web surveys. Their combination results in the best model fit, improved predictive power compared to the paradata-only measure, and the highest detection rate of lookups, potentially circumventing biased political knowledge scores. We further recommend considering the design of the knowledge questions to determine whether the inclusion of a paradata measure might negatively affect the combined measure (e.g., when using discouraging instructions) and to assess whether a self-report would be more appropriate.

Our study has some limitations that pose opportunities for future research. First, in our analysis of predicting correct answers based on the number of lookups, an effect remained for the experimental groups. Instructions impacted correct answers even when controlling for our lookup measures (i.e., the number of lookups). The obvious assumption is that this is due to imperfect lookup measures. However, if instructions have a direct effect on correct answers even when controlling for lookups, the use of instructions would come with unintended side effects. Since our experimental setup does not enable a more detailed analysis of possible underlying effects, further research is needed, such as lab experiments in which respondents are explicitly instructed how to look up answers and are asked about their understanding of the instructions using cognitive interviewing. A controlled environment could help to better understand what the lookup measures capture, and uncover the mechanism behind the direct relationship between instructions and correct answers.

Second, we were unable to determine what respondents did when they switched away from the survey. Concerning their answers to political knowledge questions, respondents might do different things: look up an answer they do not know, confirm an existing belief, or check the spelling of an answer they already know. Looking up answers can take different forms, and lookup measures seem to differ in what they capture. Thus, more refined knowledge on the process of looking up answers is required. We see merit in investigating this topic using lab experiments in which the browser histories of computers are used to assess search behavior (Gooch & Vavreck, 2019) or by complementing a web survey with tracking digital trace data (Munzert et al., 2022; Stier, Breuer, Siegers, & Thorson, 2020). In future research utilizing such data, it would be interesting to include (political) knowledge questions on diverse topics that require checking several websites to learn more about how respondents search for answers on the internet.

Third, our study drew on a probability-based web panel that was fielded in Germany. While probability-based samples are considered the gold standard in social science research (Cornesse et al., 2020), we acknowledge that many researchers rely on non-probability samples for which detecting undesirable response behavior may be important. Thus, we encourage further research on looking up answers to political knowledge questions in web surveys using nonprobability samples. In addition, we conducted our study in a single country. Since cultural norms and compliance with such instructions may vary across countries, a cross-cultural or cross-national comparison would be worthwhile.

Fourth, we presented first insights that instructions can elicit different lookup behaviors and that the effects of instructions might also vary across groups of respondents. We encourage future studies to investigate heterogeneous treatment effects and analyze which respondents are prone to looking up answers and to what extent they are compliant with instructions. To enable subgroup analyses, we recommend planning with large sample size.

Fifth, to actively influence respondent behavior, we employed instructions—encouraging and discouraging answer lookup, as well as a commitment question—that we did not consider invasive. Other strategies can be used to impact lookup behavior in web surveys, for example, timers can be used that force respondents to answer within a pre-determined period of time, which would limit their ability to look up answers (Clifford & Jerit, 2016). In a recent study, Kleinberg (2022) suggested to utilize question design to minimize the detrimental effects of lookups. We are convinced that additional (and even more creative) solutions exist and that it would be worthwhile to experimentally test these solutions to improve the quality of political knowledge measures.

Overall, the present study aimed to shift the discussion on looking up answers to political knowledge questions away from singular self-reported or paradata-based measures towards utilizing the potential of their combination and acknowledging the limitations of each. As our findings illustrate, it is a worthwhile challenge to further investigate how to combine different lookup measures so they can optimally complement each other. In our study, we focused on a combination that drew on the maximum of self-reports or paradata, whereas future studies might want to focus on comparing the potential of combining these lookup measures in different ways. Moreover, we have provided evidence that looking up answers is a behavior that can be influenced by question design. In other words, our findings carry a clear message for researchers who fear that lookups might bias their political knowledge scores: discouraging instructions and commitment questions are viable strategies for reducing lookups in web surveys.

References

Blom, A. G., C. Gathmann, and U. Krieger. 2015. “Setting Up an Online Panel Representative of the General Population: The German Internet Panel.” Field Methods 27:391–408.

Blom, A. G., M. Gonzalez Ocanto, M. Fikel, U. Krieger, and T. Rettig; SFB 884 “Political Economy of Reforms.” 2021. ZA7759: German Internet Panel, Wave 51 (January 2021). Cologne, Germany: GESIS Data Archive.

Clifford, S., and J. Jerit. 2014. “Is There a Cost to Convenience? An Experimental Comparison of Data Quality in Laboratory and Online Studies.” Journal of Experimental Political Science 1:120–31. https://doi.org/10.1017/xps.2014.5.

Clifford, S., and J. Jerit. 2016. “Cheating on Political Knowledge Questions in Online Surveys: An Assessment of the Problem and Solutions.” Public Opinion Quarterly 80:858–87. https://doi.org/10.1093/ poq/nfw030.

Cornesse, Carina, Annelies G. Blom, David Dutwin, Jon A. Krosnick, Edith D. De Leeuw, Ste´phane Legleye, Josh Pasek, Darren Pennay, Benjamin Phillips, Joseph W. Sakshaug, Bella Struminskaya, and Alexander Wenz. 2020. “A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research.” Journal of Survey Statistics and Methodology 8:4–36. https://doi.org/10.1093/jssam/smz041

Décieux, J. P. 2022. “Sequential On-Device Multitasking within Online Surveys: A Data Quality and Response Behavior Perspective.” Sociological Methods & Research 004912412210825.https://doi.org/10.1177/00491241221082593.

Diedenhofen, B., and J. Musch. 2017. “PageFocus: Using Paradata to Detect and Prevent Cheating on Online Achievement Tests.” Behavior Research Methods 49:1444–59. https://doi. org/10.3758/s13428-016-0800-7.

Gooch, A., and L. 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 7:143–62. https://doi.org/10.1017/psrm.2016.20.

Gummer, T., and T. Kunz. 2022. “Relying on External Information Sources When Answering Knowledge Questions in Web Surveys.” Sociological Methods & Research 51:816–36. https:// doi.org/10.1177/0049124119882470.

Höhne, J. K., C. Cornesse, S. Schlosser, M. P. Couper, and A. G. Blom. 2021. “Looking Up Answers to Political Knowledge Questions in Web Surveys.” Public Opinion Quarterly 84: 986–99. https://doi.org/10.1093/poq/nfaa049.

Höhne, J. K., S. Schlosser, M. P. Couper, and A. G. Blom. 2020. “Switching Away: On-Device Media Multitasking in Web Surveys.” Computers in Human Behavior 111:106417. https://doi. org/10.1016/j.chb.2020.106417.

Jensen, C., and J. P. F. Thomsen. 2014.“Self-Reported Cheating in Web Surveys on Political Knowledge.” Quality & Quantity 48:3343–54. https://doi.org/10.1007/s11135-013-9960-z.

Jerit, J. 2009. “Understanding the Knowledge Gap: The Role of Experts and Journalists.” Journal of Politics 71:442–56. https://doi.org/10.1017/S0022381609090380.

Kleinberg, M. S. 2022. “Measuring Political Knowledge in Online Surveys: How Question Design Can Improve Measures.” Public Opinion Quarterly 86:736–47. https://doi.org/10.1093/poq/nfac024.

Krosnick, J. A. 1991. “Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys.” Applied Cognitive Psychology 5:213–36. https://doi.org/10.1002/acp.2350050305.

Krosnick, J. A. 1999. “Survey Research.” Annual Review of Psychology 50:537–67. https://doi.org/10.1146/annurev.psych.50.1.537.

Motta, M. P., T. H. Callaghan, and B. Smith. 2017. “Looking for Answers: Identifying Search Behavior and Improving Knowledge-Based Data Quality in Online Surveys.” International Journal of Public Opinion Research 29:575–603. https://doi.org/10.1093/ijpor/edw027.

Munzert, S., S. Ramirez-Ruiz, P. Barbera´, A. M. Guess, and J. Yang. 2022. “Who’s Cheating on Your Survey? A Detection Approach with Digital Trace Data.” Political Science Research and Methods 1–9. https://doi.org/10.1017/psrm.2022.42.

Munzert, S., and P. Selb. 2017. “Measuring Political Knowledge in Web-Based Surveys: An Experimental Validation of Visual Versus Verbal Instruments.” Social Science Computer Review 35:167–83. https://doi.org/10.1177/0894439315616325.

Prior, M. 2014. “Visual Political Knowledge: A Different Road to Competence?” Journal of Politics 76:41–57. https://doi.org/10.1017/S0022381613001096.

Prior, M., and A. Lupia. 2008. “Money, Time, and Political Knowledge: Distinguishing Quick Recall and Political Learning Skills.” American Journal of Political Science 52:169–83. https://doi.org/10.1111/j.1540-5907.2007.00306.x.

Schlosser, S., and J. K. Ho¨hne. 2018. ECSP – Embedded Client Side Paradata. https://zenodo.org/record/1218941.

Sendelbah, A., V. Vehovar, A. Slavec, and A. Petrovčič. 2016. “Investigating Respondent Multitasking in Web Surveys Using Paradata.” Computers in Human Behavior 55:777–87. https://doi.org/10.1016/j.chb.2015.10.028.

Shulman, H. C., and F. J. Boster. 2014. “Effect of Test-Taking Venue and Response Format on Political Knowledge Tests.” Communication Methods and Measures 8:177–89. https://doi.org/10.1080/19312458.2014.937526.

Smith, B., S. Clifford, and J. Jerit. 2020. “TRENDS: How Internet Search Undermines the Validity of Political Knowledge Measures.” Political Research Quarterly 73:141–55. https://doi.org/10.1177/1065912919882101.

Sparrow, B., J. Liu, and D. M. Wegner. 2011. “Google Effects on Memory: Cognitive Consequences of Having Information at Our Finger Prints.” Science (New York, N.Y.) 333: 776–78. https://doi.org/10.1126/science.1207745.

Stier, S., J. Breuer, P. Siegers, and K. Thorson. 2020. “Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field.” Social Science Computer Review 38: 503–16. https://doi.org/10.1177/0894439319843669.

Strabac, Z., and T. Aalberg. 2011.“Measuring Political Knowledge in Telephone and Web Surveys: A Cross-National Comparison.” Social Science Computer Review 29:175–92. https:// doi.org/10.1177/0894439310371340.

Style, H., and J. Jerit. 2020. “Does It Matter if Respondents Look Up Answers to Political Knowledge Questions?” Public Opinion Quarterly 84:760–75. https://doi.org/10.1093/poq/nfaa038.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. The Psychology of Survey Response. Cambridge: Cambridge University Press.

Vezzoni, C., and R. Ladini. 2017. “Thou Shalt Not Cheat: How to Reduce Internet Use in Web Surveys on Political Knowledge.” Italian Political Science Review/Rivista Italiana Di Scienza Politica 47:251–65. https://doi.org/10.1017/ipo.2016.25.