Erschienen in: Social Science Computer Review ; 35 (2017), 2. - S. 167-183
https://dx.doi.org/10.1177/0894439315616325

Article

Social Science Computer Review
2017, Vol. 35(2) 167-183

DOI: 10.1177/0894439315616325
journals.sagepub.com/home/ssc
SAGE

Measuring Political Knowledge in Web-Based Surveys: An Experimental Validation of Visual Versus Verbal Instruments

Simon Munzert[1] and Peter Selb[1]

Abstract

Does the opportunity to deliver visual instead of verbal stimuli of political knowledge to respondents in web-based surveys make a difference in terms of data quality? For instance, does the presentation of visual knowledge items reduce cheating, that is, looking up the answer via the Web? And do visual and verbal stimuli capture the same underlying construct? To test whether the use of visuals to measure political knowledge effectively makes a difference, we administer a question form experiment in an online survey of the German Longitudinal Election Study. Respondents are randomly assigned to one of two question formats—visual or verbal—and are asked to solve a set of eight questions on political leaders and their offices. The instruments are validated based on nonparametric item response theory and analyses of response latency. While there is no clear evidence for cheating behavior under either of the conditions, both instruments form strong knowledge scales. Results from a regression analysis indicate that the scales provide measures of closely related but not identical concepts.

Keywords

political knowledge, web surveys, cheating, experiment, visuals

> The Web offers the opportunity to enrich the survey experience with images, sound, and video. Perhaps the most promising of these visual enhancements is the ability to deliver color photographs or other images to respondents. (. . .) The chance to extend survey measurement beyond the words of a question is one of the most exciting, yet least explored, aspects of Web surveys.
>
> Couper, Tourangeau, and Kenyon (2004, p. 255)

Introduction

"Political knowledge is to democratic politics what money is to economics: it is the currency of citizenship" (Delli Carpini & Keeter, 1996, p. 8).[1] For that reason, a whole bulk of electoral research

[1] University of Konstanz, Konstanz, Germany

Corresponding Author:
Simon Munzert, University of Konstanz, Universitaetsstr. 10, Konstanz, 78457, Germany.
Email: simon.munzert@uni.kn

investigates individual-level determinants (Leighley, 1991; Luskin, 1990), institutional antecedents (Benz & Stutzer, 2004; Gordon & Segura, 1997), and consequences of political knowledge[2] for individual voting behavior, public opinion, and election outcomes (Alvarez, 1998; Bartels, 1996; Ferejohn & Kuklinski, 1990; Luskin, 2002; Macdonald, Rabinowitz, & Listhaug, 1995; Selb & Lachat, 2009; Sniderman, Brody, & Tetlock, 1993). All such studies require measures of political knowledge, most of which are based on survey items about political facts.[3] Such approaches do not easily lend themselves to being administered in online surveys, since respondents may be tempted to cheat, that is, to look up the correct answers via the Web (Elo, 2009; Strabac & Aalberg, 2011).

Recent studies have diagnosed that a substantive share of respondents seize the chance to cheat in a web setting (Clifford & Jerit, 2015; Jensen & Thomsen, 2014), but a solution to this problem is less obvious. Consequently, Jensen and Thomsen (2014, p. 3353) call for future research that "should focus on how to eliminate (or control for) cheat-oriented Googling activities in web surveys on factual political knowledge." Our aim is to construct a measure that serves this purpose.

We suggest a question format that hampers cheating in a web setting. We push on with Couper et al.'s (2004) idea to extend survey measurement beyond the words of a question (see this article's opening quote) and implement visuals as a task, not only as stylistic element. Additionally, we follow recent research on paradata use to improve measurement (Callegaro, Yang, Bhola, Dillman, & Chin, 2009; Couper & Bosnjak, 2010; Kreuter, Couper, & Lyberg, 2010) by exploiting response latency measures to identify cheating behavior.

Using visual information to measure political knowledge raises questions about the comparability with classical—usually verbal—measures. According to Prior (2014), the reliance on visual-only or verbal-only treatments can bias the estimates of political knowledge, as certain subsets of the population tend to store political information verbally, whereas others primarily consume visual information. Consequently, while the use of visuals might help reduce cheating, it could in turn introduce measurement bias due to differential item functioning.

To assess whether a visual instrument (1) reduces cheating activity in a web-based setting as well as (2) measures the same underlying construct as a comparable verbal instrument, we present results from an experiment that was embedded in a general population survey as part of the German Longitudinal Election Study. Respondents were assigned to a knowledge question format based on either visual or verbal information. Additionally, collected response latencies are utilized in an attempt to ex post identify cheating activity and satisficing behavior. Based on a nonparametric item response analysis, we demonstrate that each of the batteries captures a single latent trait. Regression analyses that use scale values as dependent variable indicate that the underlying constructs' relationship to other theoretically relevant variables seems to be similar but not identical. In that sense, the use of visuals to measure political knowledge might not be essential to prevent respondents from cheating if the task itself impedes cheating by providing hardly usable information for a search engine request. It might, however, help obtain a more complete measure of political knowledge.

Is Cheating a Problem? Current Evidence and Existing Counterstrategies

Why Cheating?

While the myth of the cheating respondent has been circulating for a while (see Vavreck, 2012; Warren, 2012), theoretically justifying the expectation of cheating respondents is not straightforward. The fact that web surveys provide the opportunity to cheat does not imply a motive why to do so. In fact, based on the satisficing paradigm (Krosnick & Alwin, 1987), we would expect respondents not to cheat because the act of cheating bears a larger effort to answer the question than just guessing or refusing to answer. Instead, it has been argued that the more likely source for

measurement error in questions on political knowledge are satisficers who refuse to think about the correct answer and choose "don't know" or a random answer instead. However, this reasoning is in tension with growing evidence on cheating in web surveys, as described below. A potential explanation of the phenomenon could base on the delicateness of the item itself: Questions on political knowledge can be sensitive, specifically to those who fear that their political ignorance is revealed. If cheating occurs mainly because of social desirability, we would expect those to cheat who are reportedly politically interested or think that their status requires being knowledgeable about politics. Jensen and Thomsen (2014) provide ambiguous evidence on this claim: They find that reported cheating activity is related to the level of education and age (younger and highly educated respondents report less cheating). However, as the question on cheating in this study is sensitive itself, this result has to be taken with a grain of salt. In our study, we test this expectation with a direct measure of political interest and a behavioral indicator for cheating activity.

Recent studies suggest that self-deceptive enhancement, a variant of social desirability bias and the "tendency to give honestly believed but overly positive reports about oneself" (Booth-Kewley, Larson, & Miyoshi, 2007, p. 464), is key to explaining cheating behavior in online tests (Clifford & Jerit, 2015; Shulman & Boster, 2014). Further, they argue that in particular groups that value the topic of research (e.g., student samples) and do not face substantive opportunity costs (like, e.g., Amazon Mechanical Turk [MTurk] panelists who want to complete as many tasks as possible for their earnings) may be predisposed to self-enhancement (Brown, 2012; Clifford & Jerit, 2015). Due to a lack of variance in sampling design, we cannot test this claim directly, but we offer detailed sampling information on our as well as related studies to put the results into context.

Current Evidence for Cheating on Political Knowledge Questions in Web Surveys

While previous studies have found higher levels of political knowledge in online settings in comparison to classroom settings, telephone, or face-to-face samples (Ansolabehere & Schaffner, 2014; Elo, 2009; Shulman & Boster, 2014; Strabac & Aalberg, 2011), it remained unclear whether these differences stem from sampling error or are indeed due to cheating activities. Therefore, using an opt-in panel of Danish Internet users, Jensen and Thomsen (2014) introduce a measure of self-reported cheating in a web survey on political knowledge and report a stunning 22% of self-reported cheaters.[4] As a self-reported measure of this kind can be subject to social desirability itself, one would consider this a rather conservative estimate. In another study, Clifford and Jerit (2014) randomly assign a sample of undergraduate students enrolled in political science classes to a laboratory or online mode and find that about 10% of the online respondents report to have browsed the Internet during the survey. While this is not necessarily due to cheating, participants in the online condition performed significantly better than laboratory participants did. Oversampling of more knowledgeable respondents in the online mode can be ruled out as an alternative explanation in this design, since the mode assignment was random. In subsequent work, Clifford and Jerit (2015) present a set of studies and find varying rates of self-reported cheating, ostensibly depending on sample characteristics (from 4% in an MTurk sample to 41% in a student sample). To sum up, there is growing evidence for cheating on political knowledge questions in web surveys, although causes and extent require further exploration.

Existing Counterstrategies

Two main strategies have been suggested to cope with the cheating problem by design: time constraints and the use of visuals. Strabac and Aalberg (2011) report an experiment in which presidents' names are presented to one subsample, presidents' photos to the other, and respondents are given a 30-s time limit to answer the question. In another study, Prior and Lupia (2008) employ a 1-min time

limit. Similarly, Prior (2014) introduces a set of different time restrictions but does not find any substantive differences in the results when excluding respondents who do not meet certain time limits. However, this strategy does not take into account that people may have heterogeneous reaction times and deprives the self-administered setting of its comfortable feature that respondents carry the survey's progress in their own hands. Additionally, Jensen and Thomsen (2014) find an *inverse* relationship between self-reported cheating activity and actual response time, questioning that the use of time limits is an effective way to dam cheating at all, at least if the correct solution can be looked up very quickly.

Another strategy is the use of visuals. Strabac and Aalberg (2011) provide respondents with photographs as answers to questions on politicians (e.g., replacing the question "Who is Nicolas Sarkozy?" with "Who is this person?" and a picture). Yet, it is questionable if this way of using visuals impedes cheating a lot. Tricking respondents can still use the information from the answer categories to search for matching pictures. Therefore, we suggest a modification of the use-of-visuals approach and extend it so that visuals provide the only information available to solve the task.

A New Question Design for Web-Based Surveys

Our approach to measuring political knowledge is classical as to content but formally novel. In terms of content, it requires respondents to assign politicians to their office, but it does so by relying on visual stimuli alone. We illustrate the setup in Figure 1. Specifically, our suggestion is to prompt respondents to identify the pair of politicians on the photographs who are (or were) immediate successors in office. For example, the first task screen features pictures of Colin Powell, Ronald Reagan, George W. Bush, and Barack Obama. The correct solution would obviously be George W. Bush and Barack Obama. Certainly, Ronald Reagan had been president of the United States, too, but neither George W. Bush nor Barack Obama was his direct successors. This task is repeated with several samples of politicians, two of which (and only two) meet the selection criterion. As is the case with other multiitem knowledge measures, the items are supposed to vary in difficulty to allow for precise estimates of knowledge over the whole range of the latent construct.

We are aware of the fact that, even in this setup, it is still possible to look up answers if the respondent holds partial information on some of the displayed politicians. It is, however, a more complex task, as the respondent has to transfer visual into verbal information (e.g., for a query on a search engine) and then reconvert it into visual information.[5] We would suppose that this should reflect in the time it takes the respondent to answer the question. Response latency measures for each question captured as a by-product of the survey completion process allow us to correct for errors induced by cheating.

Irrespective of whether the stimuli are presented visually or verbally, the described approach also helps reduce the problem of guessing when compared with standard closed items (see Note 2): Selecting two of the four images can be done in six ways which, statistically, leaves the nescient respondent an approximate 17% chance of guessing the right combination.[6] Yet another strategy to both prevent respondents from cheating and reduce guessing bias would be to offer open-ended questions about a photograph, asking the name, or office of the depicted person. We refrain from this strategy for two reasons: First, this format tends to induce higher item-nonresponse levels than closed items (Millar & Dillman, 2012). Second, coding open-ended answers is a complex task, which has raised concerns about reliability and validity (DeBell, 2013). There has been a debate whether the respondent should be encouraged to guess or not (Mondak, 2000, 2001; Sturgis, Allum, & Smith, 2008). We decided not to encourage the respondent to guess and offered a don't know category, signalizing that lack of knowledge is acceptable and cheating not necessary. Obviously, if this strategy to reduce cheating works, we would observe less cheating under either of the conditions, which again would make our comparisons more conservative.

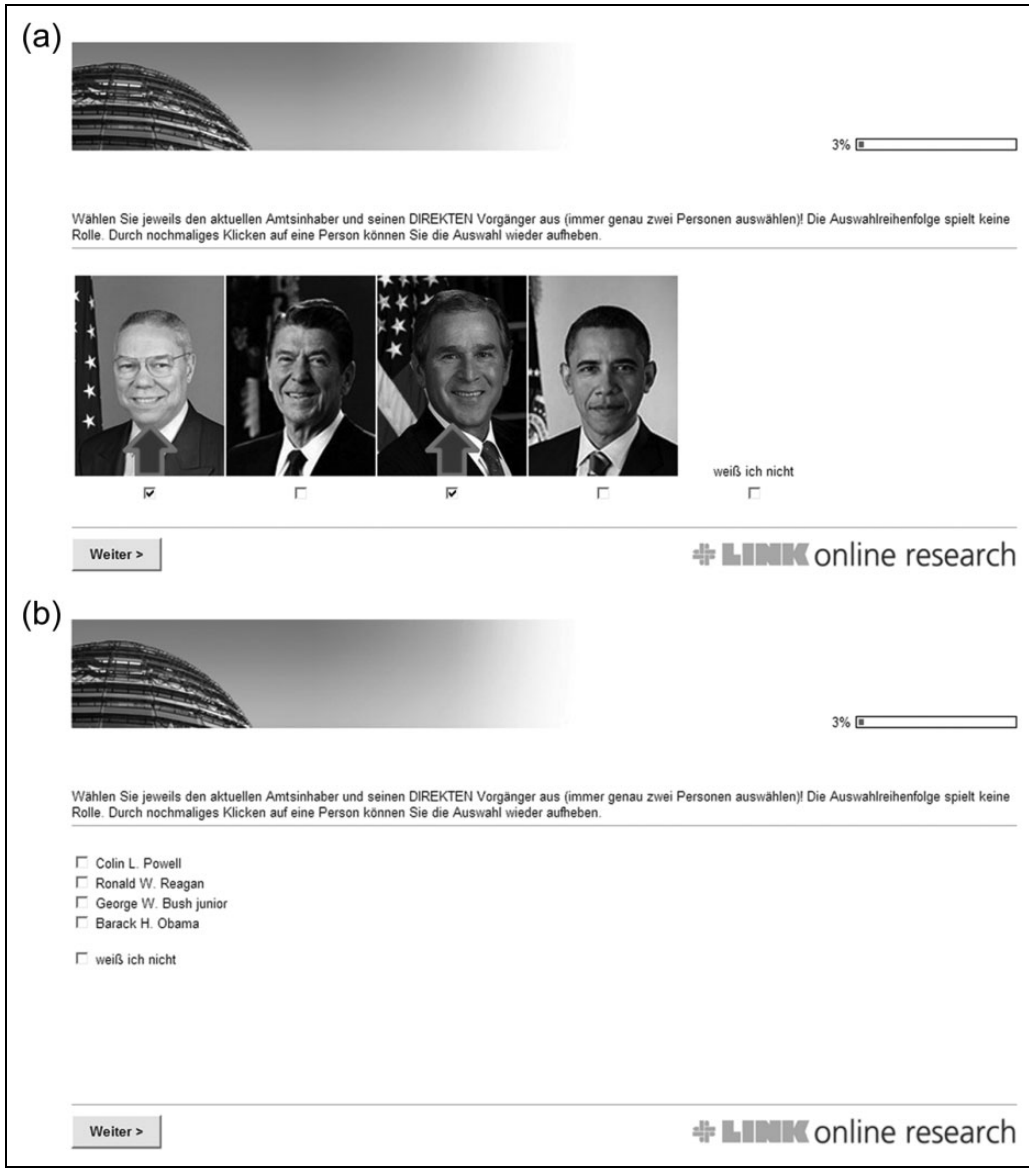

Figure 1. Screenshot of the first question of the political knowledge battery; visual (a) and verbal condition (b). The portraits were originally presented in color.

Data and Results

Experimental and Survey Design

In order to assess the quality of our proposed instrument, we set up a split-sample experiment. Half of the respondents were randomly assigned to the visual condition; the other half had to solve the same task but was given a list of names instead of pictures. This way we are able to compare item performance and response latencies and have a benchmark for measures of internal and external validity for the visual instrument.

Table 1. Descriptive Statistics of Respondent Characteristics.

Respondent Characteristic	Group		Microcensus	(N)Onliner Atlas
	Verbal	Visual		
Gender				
Male	253	272	48.5%	52.4%
	50.4%	54.4%		
Age				
18–29	105	109	16.9%	22.6%
	20.9%	21.8%		
30–44	174	161	24.1%	28.9%
	34.7%	32.2%		
45–59	153	138	26.9%	30.9%
	30.5%	27.6%		
60+	70	92	32.1%	17.6%
	13.9%	18.4%		
Formal education				
Lower secondary school (Volksschule, Hauptschule)	148	154	43.7%	33.7%
	29.5%	30.8%		
Middle school (Realschule)	232	211	30.1%	32.5%
	46.2%	42.2%		
High school (Gymnasium)	122	135	26.2%	33.8%
	24.3%	27.0%		
Total	502	500		
	50.1%	49.9%		

Note. Figures for the Microcensus and the (N)Onliner Atlas are taken from Rattinger, Roßteutscher, Schmitt-Beck, Weßels, and Wolf (2014).

Based on previous evidence on cheating activities, we expect to find (1) a higher share of correct answers under the verbal treatment, (2) a higher share of suspicious response times (our indicator for cheating activity) under the verbal treatment, and (3) people with high reported levels of political interest especially prone to cheating. In order to check whether the use of visuals induces measurement bias due to differential item functioning (Prior, 2014), we estimate treatment effects on determinants of verbal and visual political knowledge that are potentially linked to a certain cognitive style, such as education, age, and media consumption.

The experiment was embedded in the Long-Term Online Tracking survey, a survey project that is part of the German Longitudinal Election Study (Rattinger, Roßteutscher, Schmitt-Beck, Weßels, & Wolf, 2014). Participants were recruited via quota sampling from an off-line opt-in panel and had to be German citizens, 18 or older, and to reportedly use the Internet for private purposes at least once a week. The survey was launched on February 21, 2014, and closed on March 7, 2014. The response rate (RR) was 23.8% (AAPOR RR2; see The American Association for Public Opinion Research, 2015), the completion rate 83.3% (RR6; see Statistisches Bundesamt, 2010; Initiative D21, 2012). Table 1 provides a comparison of respondent characteristics in both groups as well as distributional information from the German microcensus 2009 and the (N)Onliner Atlas 2012.[7] The groups are balanced reasonably well on this set of observables. After excluding 21 speeders from the analysis, we were left with a sample of 1,002 respondents.[8]

The battery of political knowledge items consisted of 8 items, plus 1 burn-in item to explain the task. For each of them, the respondents were presented with a sample of four politicians, two of whom (and only two) met the selection criterion "one of them is the *direct* successor in office of the other." The offices were (in that order) German Federal President, U.S. President, Russian

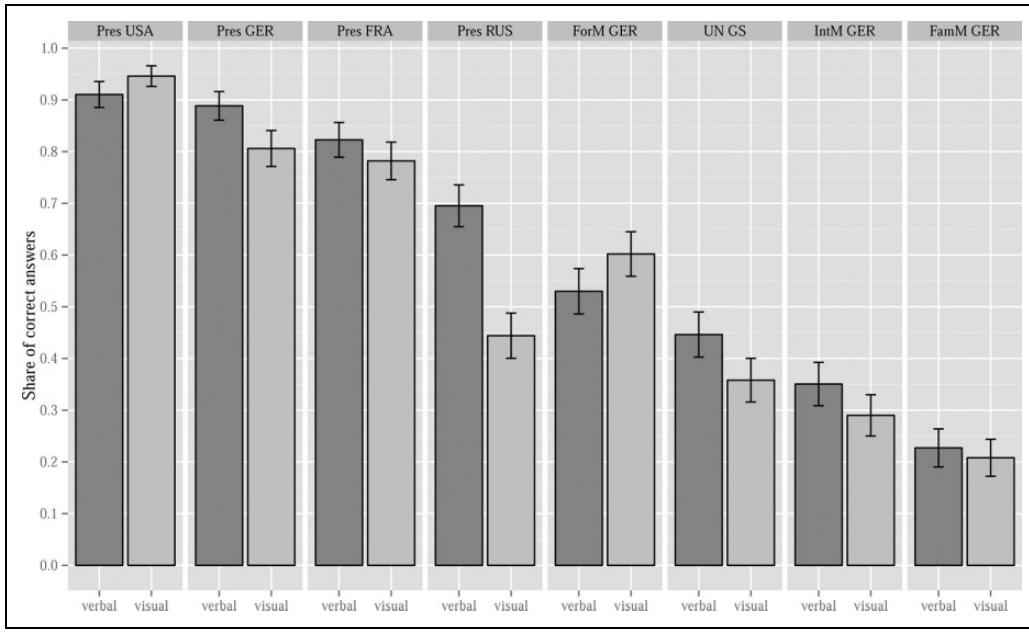

Figure 2. Shares of correct responses on political knowledge items, by treatment. Vertical lines indicate 95% confidence intervals.

President, French President, German Foreign Minister, German Interior Minister, United Nation General Secretary, and German Minister for Family Affairs. The detailed list of items as well as the introductory text are presented in Figure A1 and Table A1 in the Online Appendix. In order to answer a question, the respondents had to mark two of the four politicians by clicking on them. They were able to undo a choice by clicking on an image for a second time. Further, the current selection was clearly marked with a red arrow (see Figure 1).

Using Response Latencies to Identify Cheating Behavior

First, we ask whether our visual question format actually helped prevent respondents from cheating. Our study does not allow us to detect cheating directly because respondents fill out the survey in a private setting.[9] Therefore, we propose a detection strategy based on response latencies. We assume that cheating under the verbal condition is easier because search engines provide quick results on verbal input but are less informative for visual queries. Thus, while cheating under the visual condition cannot be ruled out entirely, we would expect cheating to be more prevalent in the verbal group. The higher average number of solved items among respondents who received the verbal treatment—4.99 versus 4.52 of the 8 among those who received the visual treatment ($p < .01$)—may be considered an initial indication of cheating. Figure 2 disentangles the performance at the item level. For both groups, the items seem to cover a broad range of difficulty levels (about 20–90% shares of correct answers). On 6 of the 8 items, respondents provided with the verbal treatment performed better than the visual group. With over 20%, the difference is most significant with the Russian President item.[10]

In the next step, we assume that cheating, in particular looking up information in the Web while completing the questionnaire, will take some time, and thus will reflect in markedly longer response latencies (see also Shulman & Boster, 2014, p. 187, for a similar conjecture). While Jensen and

Thomsen (2014) find an *inverse* relationship between response time and reported cheating activity, we argue that cheating should indeed reflect in longer response times in our setting because the research for the correct answer is more complex. Respondents would have to search for the politicians and compare their tenure. Therefore, we use outlying item-level response latencies as an indicator for potential cheating. The time it takes to answer an item is likely to be both item- and person-specific, that is, dependent on features like item difficulty or a person's general reaction times.[11] Further, the question format itself may affect latencies. Consequently, an absolute latency criterion might not quite hit the target. Instead, we try to identify the cheating attempts as follows: To isolate the portion of "suspicious latency" from latency that can be explained by systematic as well as item- and person-specific factors, we model (log) response latencies as a function of person-specific random intercepts and further include fixed effects for the items, treatment condition (visual/verbal), and accuracy of an answer (wrong/right). Further, we allowed the residual errors to vary by treatment group because another observable implication of larger cheating prevalence among those who received the verbal treatment would be a larger variance in their response latencies. Detailed results of this model are reported in Table A2 of the Online Appendix. We find substantial variance for person-specific random effects, substantive question effects, and moderate differences in response latencies between treatment groups and right versus wrong answers. Respondents who received the visual treatment tended to take slightly longer to answer the items, but the difference was not significant. The residual variance is higher in the verbal group but to a negligible extent.

Next, we extracted the residuals, classified the top 2% of observations as potential cheaters, and recoded their answer (if correct) to 0.[12] Fifty-four percent of the suspicious observations responded to the verbal treatment, which is only weak evidence for our prior belief that cheating is more pronounced among the verbal group. On the other hand, 63% of the suspicious outliers had the answer right, which points toward potential cheating behavior (especially since suspicious answer times were more prevalent with more difficult items). Therefore, we continue to use our latency-based correction scheme for the following analyses.[13]

In general, the fact that we fail to identify strong differences in cheating activity in the response time patterns may also be because both treatments exacerbate cheating. The effort to search for (up to) four politicians and compare their political career might have been an obstacle too big for most of the respondents in the verbal treatment group, too.

One more general question about cheating remains—does political interest predict cheating? Our expectation was that cheating because of social desirability primarily affects people who want to be seen as politically interested. This finding is confirmed when we regress the indicator of suspicious response latency on a measure of political interest.[14] Controlling for the treatment assignment, age, and education, the probability of suspicious response time increases significantly with the reported level of political interest. The results are reported in Table A3 in the Online Appendix. This underlines the relevance of instruments that exacerbate cheating—if specific groups are more prone to cheating, this does not only lead to overestimation of general levels of political knowledge but also affects inference on the relationship between personal characteristics and political knowledge.

Assessing Measurement Validity Using a Nonparametric Item Response Theory (IRT) Model

Are item responses actually driven by the respondents' political knowledge? Most previous studies take the number of correct responses as their measure of political knowledge and thus simply assume the existence of the latent trait. Lately, this practice has come under fire on test theoretical grounds (Barabas, 2002; Delli, Carpini, & Keeter, 1993, 1996; Levendusky & Jackman, 2003; Luskin, 1987; Mondak, 2000, 2001; Prior, 2014; Sturgis et al., 2008). We will employ a model from the IRT paradigm that is more explicit about the relationship between the latent characteristics to be measured

Table 2. Distribution of Individual Response Patterns Across 8 Knowledge Items.

Response Patterns								Verbal, Frequencies		Visual, Frequencies	
								Absolute	Relative	Absolute	Relative
0	0	0	0	0	0	0	0	9	0.018	6	0.012
1	0	0	0	0	0	0	0	5	0.010	32	0.064
1	1	0	0	0	0	0	0	20	0.040	15	0.030
1	1	1	0	0	0	0	0	21	0.042	30	0.060
1	1	1	1	0	0	0	0	46	0.092	56	0.112
1	1	1	1	1	0	0	0	35	0.070	24	0.048
1	1	1	1	1	1	0	0	26	0.052	18	0.036
1	1	1	1	1	1	1	0	37	0.074	21	0.042
1	1	1	1	1	1	1	1	42	0.084	34	0.068
								241	0.482	236	0.472

Note. Items are in ascending order according to their difficulty, that is, the proportion of respondents providing incorrect item responses. The item order is given in Table 3. Only model-consistent response patterns are reported.

and the individuals' responses to the items. In particular, we utilize the Mokken scale, a nonparametric variety of IRT models (van Schuur, 2003). Like its deterministic precursor, the Guttman scale, Mokken's model considers individual item responses as a manifestation of a single underlying personal trait (*unidimensionality*) to the extent that items can be arranged in an order so that an individual who "dominates" (i.e., solves) a particular item also dominates items of lower rank order. In other words, response patterns that are consistent with this model are fully described by the most difficult question that the respondent was able to solve, because all easier questions are expected to be solved as well.

As an initial test of the item responses' unidimensionality, we may simply see how many of each group's respondents exhibit model-consistent response patterns. For the verbal sample, we find 241 or 48% of the observations with a consistent pattern (see Table 2). Rates are virtually the same for the visual sample (236 observations or 47%). Given the number of items and the fact that some of the items are very similar in terms of difficulty (see again Figure 2), these figures are rather impressive and support the assumption of unidimensionality of both scales.

A second assumption, *monotonicity*, requires that the probability of a correct response to any item is a nondecreasing function of the individuals' locations on the latent trait or score, which is given by the number of correct responses across items. Obviously, monotonicity is an important feature that allows for the interpretation of the score as an ordinal measure of the latent trait. The monotonicity assumption can be checked by visual inspection of the item trace lines which display the proportions of correct responses across the latent trait (see Figure 3). With the exception of the German Foreign Minister item, all the item trace lines exhibit virtually monotonic increases over the range of the scores. We offer an explanation for the suspicious pattern of the German Foreign Minister item below.

As opposed to the Guttman scale, the Mokken scale is probabilistic in the sense that it provides clear criteria against which to value empirical departures from model-consistent response patterns as either random or systematic. In particular, Loevinger's coefficient of homogeneity H is used to denote (1 minus) the ratio of observed Guttman errors—individual patterns that include, for any pair of ordered items, correct responses to the more difficult and incorrect responses to the easier item—versus expected Guttman errors assuming that item responses were statistically independent. As a rule of thumb, H should be greater than 0.3 for each item involved (and for the scale as well), which

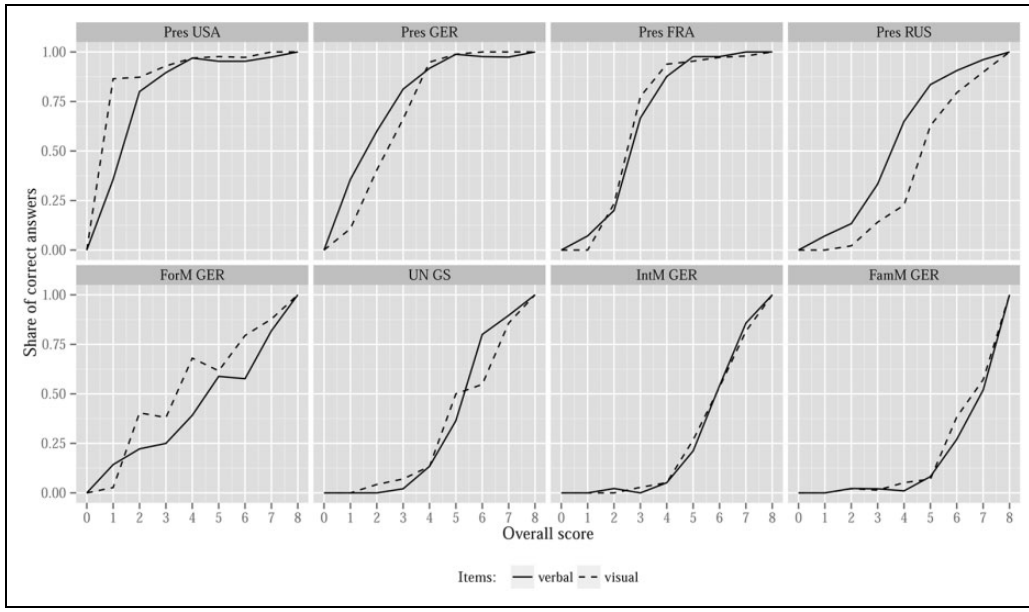

Figure 3. Trace lines of the 8 items as a function of the scores, separated by treatment.

Table 3. Results From a Mokken Scaling Analysis of 8 Knowledge Items: Item- and Scale-level Values of Loevinger's H.

	Easiness Rank		Loevinger's H	
Item	Verbal	Visual	Verbal	Visual
Pres USA	1	1	0.37	0.35
Pres GER	2	2	0.44	0.58
Pres FRA	3	3	0.56	0.57
Pres RUS	4	5	0.47	0.48
ForM GER	5	4	0.26	0.29
UN GS	6	6	0.54	0.45
IntM GER	7	7	0.53	0.52
FamM GER	8	8	0.55	0.54
Scale			0.46	0.48

Note. All H values are significantly greater than 0 at $p < .01$. Pres USA = U.S. President; Pres GER = German Federal President; Pres FRA = French President; Pres RUS = Russian President; ForM GER = German Foreign Minister; UN GS = United Nation General Secretary; IntM GER = German Interior Minister; FamM GER = German Minister for Family Affairs.

holds true for the set of items under consideration (see Table 3) under both treatments, except for the German Foreign Minister item. At the scale level, H values of 0.46 (verbal group) and 0.48 (visual group) indicate that both item batteries form quite strong one-dimensional scales and that the scale built from visual items is slightly more homogeneous.

In sum, both of our item batteries seem to capture a single latent trait, and for each value of the latent trait, the item difficulty order is approximately the same. Based on the results of the Mokken analysis, however, we drop the German Foreign Minister item from the scale, as it violates several of the model's assumptions in both treatment groups. This is also what the automated item selection

procedure as proposed by Mokken (1971) and Sijtsma and Molenaar (2002) suggests. A closer inspection of the answer patterns on this item gives a hint to why this item does not work very well. The wrong answers are not distributed equally over the five wrong combinations of politicians but peak on the "Westerwelle/Lindner" pair (see again Table A1 for all possible answers). While neither of them was a direct successor of the other in a political office, they are indeed closely related as prominent members of the Free Democratic Party in Germany. At the time when the survey was conducted, Lindner just had become the party's new leader, with Westerwelle as his pre-predecessor. Obviously, many respondents were misguided by this item, and perhaps particularly those with higher levels of political knowledge. This finding also has implications for the construction of political knowledge measures. Wrong answer options that lead respondents astray might not only make a task more difficult to solve (which is sometimes desirable to capture the population's full variance on the latent trait) but can also render this item useless as a measurement instrument. The choice of appropriate wrong answer options has not gained much attention so far in the research on measurement of political knowledge. The use of IRT methods can help researchers identify such flaws in an instrument. After excluding this item, we gain Loevinger's H values for the scales of 0.57 (verbal group) and 0.56 (visual group), which imply strong scales (Mokken, 1971; van Schuur, 2003). For the time being, the test scores, that is, the sum of correctly answered items, may be interpreted as an ordinal measure of political knowledge.

Assessing Treatment Effects on Determinants of Verbal and Visual Political Knowledge

In a last step, we investigate whether the choice of question mode matters for the study of political knowledge as a dependent variable. Prior (2014) argues that differences in levels of political knowledge by personal attributes such as gender or socioeconomic status can be partially ascribed to the question format. People who predominantly process information visually are at a disadvantage if knowledge is only tested in the traditional verbal way (Graber, 1990, 2001). In order to find out whether the relationship between political knowledge and personal characteristics is substantively shaped by the question format, we proceed as follows: As we have demonstrated in the previous section, it is legitimate to use the (7-item) score on each of the scales as an ordinal measure. Therefore, we estimate interaction effects between treatment type and classical determinants of knowledge in an ordered probit model to determine whether the effects of the selected covariates on knowledge levels vary by treatment. Similar to Prior (2014), we use standard measures for gender, age, and income. Further, we construct a variable indicating whether the respondent mainly relies on visual media (TV) to get informed about politics. The estimation results are reported in Table 4. We ran the model with both corrected (Model 1) and uncorrected data (Model 2) for alleged cheating behavior to assess whether the corrections affect our substantive conclusions. As can be seen from the table, the results were virtually the same, which is why we do not discuss them separately in the following.

Two main insights can be gained from this analysis. First, in line with past research (e.g., Delli Carpini & Keeter, 1996), most of the socioeconomic variables are substantively related to the measured levels of political knowledge. Male, highly educated, older, and economically well-equipped people tend to score higher on the scale. For example, the odds for male respondents to score highest (7 points) compared to scoring less than highest (0–6 points) are about 2 times greater than for women ($\exp(0.72) = 2.05$).[15] We also included a variable indicating whether the respondent mainly relies on visual media (which is, by our definition, TV). The negative coefficient implies that these respondents tend to score lower on the verbal scale, which seems plausible.

Second, while most of the estimated interaction effects are of rather small size, some of the coefficients indicate that the relationship between respondent characteristic and knowledge score indeed depends upon the question format. Highly educated persons seem to perform relatively worse under the visual treatment, while respondents that mainly rely on visual media for political information

Table 4. Estimates From Ordinal Probit Models of Political Knowledge.

	Model 1	Model 2
Coefficient	Estimates (*SE*)	
Formal education (baseline: lower secondary school)		
Middle school	**0.65 (0.11)**	**0.63 (0.11)**
High school	**1.10 (0.14)**	**1.12 (0.14)**
Age (baseline: 18–29 years)		
30–44 years	**0.40 (0.13)**	**0.40 (0.13)**
45–59 years	**0.46 (0.14)**	**0.47 (0.14)**
60 Years or more	**1.09 (0.17)**	**1.05 (0.17)**
Income (baseline: less than 1,000 EUR)		
1,000–1,999 EUR	0.04 (0.18)	0.09 (0.18)
2,000–2,999 EUR	0.35 (0.19)	**0.37 (0.18)**
3,000 EUR and more	**0.47 (0.19)**	**0.55 (0.19)**
Male	**0.72 (0.10)**	**0.69 (0.10)**
Main information source: TV	−0.16 (0.10)	**−0.21 (0.10)**
Visual treatment	−0.19 (0.29)	−0.21 (0.29)
X Male	−0.05 (0.14)	−0.02 (0.14)
X Middle school	**−0.34 (0.16)**	−0.32 (0.16)
X High school	**−0.52 (0.19)**	**−0.51 (0.19)**
X 30–44 years	−0.21 (0.19)	−0.19 (0.19)
X 45–59 years	0.00 (0.20)	−0.01 (0.20)
X 60 years or more	−0.39 (0.23)	−0.32 (0.23)
X 1,000–1,999 EUR	0.34 (0.25)	0.28 (0.25)
X 2,000–2,999 EUR	−0.06 (0.25)	−0.10 (0.25)
X 3,000 EUR and more	0.08 (0.25)	0.02 (0.25)
X Main information source: TV	**0.29 (0.14)**	**0.33 (0.14)**
κ_1 (0|1)	**−1.02 (0.23)**	**−1.11 (0.24)**
κ_2 (1|2)	−0.17 (0.22)	−0.20 (0.22)
κ_3 (2|3)	0.36 (0.22)	0.32 (0.22)
κ_4 (3|4)	**1.00 (0.22)**	**0.96 (0.22)**
κ_5 (4|5)	**1.59 (0.22)**	**1.55 (0.22)**
κ_6 (5|6)	**2.16 (0.23)**	**2.08 (0.23)**
κ_7 (6|7)	**2.94 (0.23)**	**2.77 (0.23)**
Log likelihood	−1,710.54	−1,710.28
AIC	3,477.08	3,476.57
Condition Hessian	1,100	1,100
N	951	951

Note. The dependent variable is constructed as an additive score from 7 knowledge items, ranking from 0 to 7 (see items in Table 3, ForM GER excluded). Model 1 uses corrections for alleged cheating behaviour, Model 2 uses uncorrected data. Bolded coefficients are significant at $p < .05$. AIC = Akaike information criterion.

score better on the picture-only scale.[16] Our findings are mostly in line with past research: Prior (2014) finds highly educated respondents underperform under the visual condition, while TV news consumers overperform in his study when asked for visual political knowledge. Similar to Prior, we get inconclusive effects for our income and age measures. We cannot replicate Prior's finding of female respondents significantly overperforming under the visual treatment, although the estimated effects point to the same direction. Furthermore, the substantive general conclusions drawn from the model do not differ, as the directions of effects remain the same under both treatment conditions. Thus, while we have shown that both instruments provide strong scales of a one-dimensional trait

each of which we label "political knowledge," this indicates that the scales are measures of closely related but not identical concepts.

Discussion and Conclusion

Web-based surveys offer new opportunities in terms of extended measurement but also may render existing instruments useless if they are sabotaged in the self-administered setting. We argued that using classical, that is, verbal-only tasks to measure political knowledge in a web-based survey turns out to be problematic if respondents start to cheat by looking up the correct answers on the web. To solve this problem, we suggested using an alternative instrument that is purely based on visual information. It consists of a battery of 8 items on which respondents have to match faces of politicians according to their office and can be easily implemented in an online survey while it is hard to be outwitted.

Using response latencies allowed identifying alleged cheating behavior, and a nonparametric item analysis helped assess the quality of the scale. Comparisons with a verbal-only treatment group provided only weak evidence that the latter is predominantly affected by cheating and that both instruments performed remarkably similar. We hypothesized that this is because both treatments aggravate cheating compared to traditional knowledge items. Additionally, the uncertainty associated with our behavioral indicator for cheating—residuals of models of response latency—may cover existing but weak differences.

Both instruments formed a strong Mokken scale of the latent trait. According to these findings, we conclude that both instruments can be used in web-administered settings without much loss. This, again, may also bear a remarkable potential, as it encourages researchers to vary the question format by subpopulations for which one of the treatment conditions is not an option.[17]

We are aware of the fact that an exclusively image-based approach does not come without certain limitations. Items that address offices of politicians are always at risk to run out of fashion quickly, which calls for item analyses (such as the conducted Mokken analysis) to assure the validity of the scale. We are less skeptical regarding the content of our questions. Asking for politicians and their office is consistent with previous measures of political knowledge. It is by no means exhaustive of what a person *should* know when he or she is attested profound political knowledge. This may also include an understanding of the work of major political institutions as well as the ideological positions of political actors, and possibly also historical knowledge with respect to political processes and events. The suggested method is in no way limited to the matching of pairs of politicians, although it might be the most straightforward task. Using pictures of events or buildings or linking pictures of politicians with other textual information that cannot be related to the picture via a search engine request (e.g., politicians and their party affiliations) might help to capture a more general measure of political knowledge. However, it might be advisable to stick to one set of tasks in order to minimize the fatiguing or overstraining of the respondents. Finally, we acknowledge that our specific scale hardly provides a suitable measure in other political contexts. Based on our findings, we therefore encourage other scholars of political knowledge to cast an eye on the following guideline for the construction of new instruments of political knowledge to be implemented in a web-based survey:

1. Develop an instrument based on visual or verbal cues. While our results suggest that the question mode choice may not severely affect inference, a mixture may help to gather more valid estimates of the latent trait in certain subpopulations that are at a disadvantage under either of the variants (see also Prior, 2014).
2. Conduct a pretest of the instruments to identify flawed items as well as a selection of items that covers the full variance of the latent trait in the population of interest. Prior to our experiment, we conducted a first test in a student sample and discovered considerable differential item functioning on some of the items, which we then excluded from the battery.

3. If available, use response latencies to correct for malfeasant answering behavior. As we failed to demonstrate that this step produces more homogenous scales, one could also build upon the rationale that visual cues reduce the potential for cheating a priori, or argue that the consequences of cheating are expected to be negligible.
4. Conduct item analyses to assess the validity of the scale as well as to identify and exclude potentially unsuitable items.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Supplementary and replication materials for this article can be accessed at https://github.com/simonmunzert/sscr-political-knowledge.
2. Political knowledge has been conceptualized under a variety of labels, such as political sophistication, political expertise, structural knowledge, political awareness, or factual knowledge. We refer to factual political knowledge, which can be defined as knowledge on generally accepted facts of the political system, including parties, politicians, or rules.
3. For example, standard questions asked in the American National Election Studies (ANES) are "Do you happen to know which party had the most members in the House of Representatives in Washington BEFORE the election (this/last) month?" and "Which party is more conservative?" Additionally, since 1986, many of the ANES surveys include open-ended questions which ask for offices certain politicians hold.
4. After answering a set of four knowledge questions, the respondents were asked: "The Internet has made it much easier for ordinary people to get access to information about important questions. Many use the Internet on a regular basis. Therefore, we would like to know if you used the Internet when answering one or more of the previous four questions."
5. In fact, Google provides a service (images.google.com) that allows the user to look up an image on the Web by pasting it into the input field. However, this tool is not known very well and does not always deliver valid feedback.
6. Selecting two of the four images is possible in $\binom{4}{2} = \frac{4!}{2!(4-2)!} = 6$ ways, the inverse of which is 16.7%.
7. Quota sampling was based on gender, education, and age characteristics. Completing the questionnaire was incentivized with an Amazon voucher of EUR 3.50. Two reminders were sent during the field period. Respondents that dropped out during the survey were not recorded in the data set, which makes it impossible to identify dropout figures by treatment, another potentially interesting characteristic of measurement instruments.
8. Following common practice, speeders are excluded from analysis because they give no substantive answers. In this case, it is inappropriate to merely recode them as wrong answers for two reasons. First, we cannot know if the person actually did not know the correct solution or just refused to solve the task properly (satisficing). Second, if speeders' answers are coded as wrong but are not excluded from the sample, the homogeneity of the scale is artificially increased, because the response pattern "all answers wrong" is, by definition, consistent with the item response theory framework we apply later.
9. Although technically possible, keeping a log of the respondents' browser history without their consent seems ethically dubious if not illegal.
10. In the Online Appendix, we provide further descriptive evidence on item and scale performance.

11. Figure A2 in the Online Appendix provides visual evidence that response latency is, in part, item specific. Regardless of the treatment, we see that more difficult items tend to take more time to be answered, as well as items that are asked at the very beginning (i.e., when respondents still have to get used to the task).
12. We wanted to keep corrections at a minimum, as we can only provide indirect evidence for cheating behavior and do not want our validation results to be driven by our recoding efforts. We compared the proportion of correct answers within this group across different specifications (0.5–5% of observations) and found the differences to be increasingly blurred with lower thresholds.
13. Note that we found that using the uncorrected answers actually led to slightly *higher* homogeneity scores on the Mokken scales (0.50 vs. 0.46 for the verbal and 0.49 vs. 0.48 for the visual scale). This could indicate that people tend to cheat especially when they feel they should know the answer, that is, on items whose difficulty is within their range of knowledge (for a similar argument about the sociopsychological mechanisms behind vote overreporting, see Bernstein, Chadha, & Montjoy, 2001). In fact, there is a significant positive relationship between a respondent's overall score and the difficulty (as identified in the Mokken scale) of the suspicious item. Consequently, we would expect cheating to artificially increase the scale's homogeneity at the cost of an upward bias of the knowledge level estimate.
14. The corresponding question in the survey was "In general, how interested are you in politics?," with an answer scale ranging from *not interested at all* (1) to *very interested* (5).
15. Under the proportional odds assumption, this applies to any other comparison at all thresholds.
16. We arrive at the same conclusions if we estimate separate ordinal probit models for each treatment group.
17. As an example, consider the following: According to Grüter, Grüter, and Carbon (2008), up to 2.5% of the population suffer from prosopagnosia or face blindness, whereas common estimates for the prevalence of dyslexia, which is sometimes associated with name blindness, range from 5% to 17% (McCandliss & Noble, 2003).

Supplemental Material

The online appendices are available at http://journals.sagepub.com/doi/suppl/10.1177/0894439315616325

References

Alvarez, R. M. (1998). *Information and Elections. Revised to Include the 1996 Presidential Election*. Ann Arbor: University of Michigan Press.

Ansolabehere, S., & Schaffer, B. F. (2014). Does survey mode still matter? Findings from a 2010 multi-mode comparison. *Political Analysis*, *22*, 285–303.

Barabas, J. (2002). Another look at the measurement of political knowledge. *Political Analysis*, *10*, 209–223.

Bartels, L. M. (1996). Uninformed votes: Information effects in presidential elections. *American Journal of Political Science*, *40*, 194–230.

Benz, M., & Stutzer, A. (2004). Are voters better informed when they have a larger say in politics?—Evidence for the European Union and Switzerland. *Public Choice*, *119*, 31–59.

Bernstein, R., Chadha, A., & Montjoy, R. (2001). Overreporting voting: Why it happens and why it matters. *Public Opinion Quarterly*, *65*, 22–44.

Booth-Kewley, S., Larson, G. E., & Miyoshi, D. K. (2007). Social desirability effects on computerized and paper-and-pencil questionnaires. *Computers in Human Behavior*, *23*, 463–477.

Brown, J. D. (2012). Understanding the better than average effect: Motives (still) matter. *Personality and Social Psychology Bulletin*, *38*, 209–219.

Callegaro, M., Yang, Y., Bhola, D. S., Dillman, D. A., & Chin, T.-Y. (2009). Response latency as an indicator of optimizing in online questionnaires. *Bulletin of Sociological Methodology*, *103*, 5–25.

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*, 120–131.

Clifford, S., & Jerit, J. (2015). *Cheating on knowledge questions in online surveys: An assessment of the problem and solutions* (Unpublished Manuscript).

Couper, M. P., & Bosnjak, M. (2010). Internet surveys. In P. Marsden & J. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 527–550). Bingley, England: Emerald Group Publishing Limited.

Couper, M. P., Tourangeau, R., & Kenyon, K. (2004). Picture this! Exploring visual effects in web surveys. *Public Opinion Quarterly*, *68*, 255–266.

DeBell, M. (2013). Harder than it looks: Coding political knowledge on the ANES. *Political Analysis*, *21*, 393–406.

Delli Carpini, M. X., & Keeter, S. (1993). Measuring political knowledge: Putting first things first. *American Journal of Political Science*, *37*, 1179–1206.

Delli Carpini, M. X., & Keeter, S. (1996). *What Americans know about politics and why it matters*. New Haven, CT: Yale University Press.

Elo, K. (2009). Asking factual knowledge questions. Reliability in web-based, passive sampling surveys. *Social Science Computer Review*, *27*, 1–16.

Ferejohn, J. A., & Kuklinski, J. H. (1990). *Information and democratic processes*. Chicago: University of Illinois Press.

Gordon, S. B., & Segura, G. M. (1997). Cross-national variation in the political sophistication of individuals: Capability or choice? *Journal of Politics*, *59*, 126–147.

Graber, D. A. (1990). Seeing I remembering. How visuals contribute to learning from television news. *Journal of Communication*, *40*, 134–155.

Graber, D. A. (2001). *Processing politics: Learning from television in the internet age*. Chicago: University of Chicago Press.

Grüter, T., Grüter, M., & Carbon, C.-C. (2008). Neural and genetic foundations of face recognition and prosopagnosia. *Journal of Neuropsychology*, *2*, 79–97.

Initiative D21. (2012). *(N)Onliner atlas 2012*. Berlin, Germany: Basiszahlen für Deutschland.

Jensen, C., & Thomsen, J. P. F. (2014). Self-reported cheating in web surveys on political knowledge. *Quality & Quantity*, *48*, 3343–3354.

Kreuter, F., Couper, M. P., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. In *Proceedings of the American Statistical Association's Survey Research Methods Section* (pp. 282–296). Alexandria, VA: American Statistical Association.

Krosnick, J. A., & Alwin, D. F. (1987). *Satisficing: A strategy for dealing with the demands of survey questions*. Columbus: Ohio State University.

Leighley, J. (1991). Participation as a stimulus of political conceptualization. *Journal of Politics*, *53*, 198–211.

Levendusky, M. S., & Jackman, S. D. (2003). *Reconsidering the measurement of political knowledge*. Working paper. Retrieved from http://www.stanford.edu/class/polisci353/2004spring/reading/levendusky_final.pdf

Luskin, R. C. (1987). Measuring political sophistication. *American Journal of Political Science*, *31*, 856–899.

Luskin, R. C. (1990). Explaining political sophistication. *Political Behavior*, *12*, 331–361.

Luskin, R. C. (2002). From denial to extenuation (and finally beyond): Political sophistication and citizen performance. In J. H. Kuklinski (Ed.), *Thinking about political psychology* (pp. 281–301). Cambridge, England: Cambridge University Press.

Macdonald, S. E., Rabinowitz, G., & Listhaug, O. (1995). Political sophistication and models of issue voting. *British Journal of Political Science*, *25*, 453–483.

McCandliss, B. D., & Noble, K. G. (2003). The development of reading impairment: a cognitive neuroscience model. *Mental retardation and developmental disabilities research reviews*, *9*, 196–205.

Millar, M., & Dillman, D. A. (2012). Do mail and internet surveys produce different item nonresponse rates? An experiment using random mode assignment. *Survey Practice*, *5, 1–6*.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin, Germany: De Gruyter.

Mondak, J. J. (2000). Reconsidering the measurement of political knowledge. *Political Analysis*, *8*, 57–82.

Mondak, J. J. (2001). Developing valid knowledge scales. *American Journal of Political Science*, *45*, 224–238.

Prior, M. (2014). Visual political knowledge: A different road to competence? *The Journal of Politics*, *76*, 41–57.

Prior, M., & Lupia, A. (2008). Money, time, and political knowledge: Distinguishing quick recall and political learning skills. *American Journal of Political Science*, *52*, 169–183.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., & Wolf, C. (2014). *Long-term online tracking, T23 (GLES)*. ZA5723 Data file Version 1.0.0, doi:10.4232/1.11880. Cologne, Germany: GESIS Data Archive.

Selb, P., & Lachat, R. (2009). The more, the better? Counterfactual evidence on the effect of compulsory voting on the consistency of party choice. *European Journal of Political Research*, *48*, 573–597.

Shulman, H. C., & Boster, F. J. (2014). Effect of test-taking venue and response format on political knowledge tests. *Communication Methods and Measures*, *8*, 177–189.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sniderman, P. M., Brody, R. A., & Tetlock, P. E. (1993). *Reasoning and choice. Explorations in political psychology*. Cambridge, England: Cambridge University Press.

Statistisches Bundesamt. (2010). *Mikrozensus 2010. Qualitätsbericht*. Wiesbaden, Germany: Author.

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

Sturgis, P., Allum, N., & Smith, P. (2008). An experiment on the measurement of political knowledge in surveys. *Public Opinion Quarterly*, *85*, 90–102.

The American Association for Public Opinion Research. (2015). *Standard definitions: Final dispositions of case codes and outcome rates for surveys* (8th ed.). AAPOR. Retrieved from http://www.aapor.org/AAPORKentico/AAPOR_Main/media/publications/Standard-Definitions2015_8theditionwithchanges_April2015_logo.pdf

van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, *11*, 139–163.

Vavreck, L. (2012). *The myth of cheating on self-completed surveys*. Retrieved February 2, 2015, from http://today.yougov.com/news/2012/04/17/myth-cheating-self-completed-surveys/

Warren, J. (2012). Fake orgasms and the tea party: Just another political science convention. *The Atlantic Online*. Retrieved February 2, 2015, from http://www.theatlantic.com/politics/archive/2012/04/fake-orgasms-and-the-tea-party-just-another-political-science-convention/255909/

Author Biographies

Simon Munzert is a postdoctoral research and teaching fellow at the Chair for Survey Research, University of Konstanz, Germany. His research interests include methods of election forecasting, public opinion measurement, and techniques for web scraping and Big Data management. Email: simon.munzert@uni.kn

Peter Selb is a professor of survey research at the University of Konstanz. His research interests include survey methods as well as political behavior and public opinion, with a particular emphasis on elections and political representation. Email: peter.selb@uni.kn