Survey Research Methods (2021)
Vol. 15, No. 1, pp. 79-100
doi:10.18148/srm/2021.v15i1.7594
European Survey Research Association

Using Response Times to Enhance the Reliability of Political Knowledge Items: An Application to the 2015 Swiss Post-Election Survey

Lionel Marquis
Faculty of Social and Political Sciences
University of Lausanne, Switzerland

In this article, I consider the problem of "cheating" in political knowledge tests. This problem has been made more pressing by the transition of many surveys to online interviewing, opening up the possibility of looking up the correct answers on the internet. Several methods have been proposed to deal with cheating ex-ante, including self-reports of cheating, control for internet browsing, or time limits. Against this background, "response times" (RTs, i.e., the time taken by respondents to answer a survey question) suggest themselves as a post-hoc, unobtrusive means of detecting cheating. In this paper, I propose a procedure for measuring individual-specific and item-specific RTs, which are then used to identify unusually long but correct answers to knowledge questions as potential cases of cheating. I apply this procedure to the post-electoral survey for the 2015 Swiss national elections. My analysis suggests that extremely slow responses to two out of four questions are definitely suspicious. Accordingly, I propose a method for "correcting" individual knowledge scores and examine its convergent and predictive validity. Based on the finding that a simple revised scale of political knowledge has greater validity than the original additive scale, I conclude that the problem of cheating can be alleviated by using the RT method, which is again summarized in the conclusion to ensure its applicability in empirical research.

Keywords: political knowledge; cheating; response times

1 Introduction

In this contribution, I aim to demonstrate that response times (i.e., the time taken by respondents to answer a survey question) can be a useful adjunct to the investigation of political knowledge in mass publics. More specifically, I argue that response times can be helpful for detecting potentially "dishonest" behavior in responding to political knowledge questions. The issue is all the more pressing as the large-scale transition from traditional to online surveys has widened the possibilities of "cheating" by looking up answers to knowledge questions on the internet. When answering a questionnaire online, respondents are inevitably tempted to browse the internet for additional information about the issues at stake, in order to improve the quality of their answers. In the case of political knowledge questions, this concern for "quality" obviously includes the search for a "correct" answer. Insofar as web searching is a time-consuming activity, unusually long response times may indicate the occurrence of cheating. More generally, response times can provide supplementary evidence about the quality of knowledge questions in the first place. For example, if respondents are bewildered or misled by a question, this should be reflected not only in higher "don't know" rates, but also in extended response times, in comparison with items of similar length or complexity (e.g., in terms of the number of response options). Thus, response times can allow researchers to assess the inherent difficulty, clarity, and potentially misleading aspects of items they use to measure individuals' degree of knowledge about politics.

In the second section, I will discuss in more detail the problems we are facing when trying to measure political knowledge with online surveys, and I present some solutions tested in recent research. In the third section, I will spell out the theoretical and methodological assumptions of my approach, showing how response times can help alleviate some difficulties involved in the "cheating" issue. In Section 4, I present the measurements used in the empirical analysis presented in Section 5. Focusing on the 2015 Swiss national elections, where a post-electoral survey ($N = 5337$) was conducted online and by telephone, my analysis suggests that two of the four knowledge questions are biased by cheating, and that a third one is misleading. In Section 6, a method of

Contact information: Lionel Marquis, Institut d'études politiques (IEP), Quartier UNIL-Mouline, Bâtiment Géopolis, CH-1015 Lausanne, Switzerland (E-Mail: Lionel.Marquis@unil.ch)

correcting for these biases is proposed and tested in terms of convergent and predictive validity. A final section concludes with a summary of results and some guidelines for future research.

2 The nature of the problem

In this section, I focus on the issue of "cheating behavior" in responses to knowledge questions. As we go along, however, other measurement problems will emerge, to which response times may also offer practical solutions. First, I review the literature specifying the importance and possible causes of cheating behavior. Second, I show how this problem has been addressed in recent survey research.

Before turning to these questions, it is necessary to spell out the definition of "cheating" as used in this contribution. By cheating, I mean the use of any outside source of information (other than an individual's own preexisting knowledge stored in her long-term memory) to answer a survey question. This definition does not take into account whether the questionnaire provides explicit instructions not to look up answers or whether respondents are actually feeling that they are "cheating". Similarly, it is blind to empirical and normative considerations such as why individuals cheat or whether cheating may be justified in given circumstances. Importantly, my definition puts a premium on declarative knowledge (from explicit memory systems) rather than on procedural knowledge (from implicit memory systems) such as knowing where and how to find information (Prior & Lupia, 2008), even though both types of knowledge are arguably necessary for an informed and responsible citizenry to exist. I come back to this issue in the conclusion.

2.1 How reliable are political knowledge tests in the internet age?

With the large-scale development of online surveys in recent years (Evans & Mathur, 2018; T. P. Johnson, Basic, & Joscelyn, 2016), new practical and methodological questions have arisen for all profit and nonprofit sectors which have turned to this interviewing mode. In political science, scholars have been summoned to deal with the challenge of cheating behavior in political knowledge tests—indeed, web surveys may increase the risk of "cheating" by enabling respondents to quietly look up answers on the internet. In this regard, survey research has reached two main conclusions. First, the speculation that online interviewees can easily search for the correct answers to factual knowledge questions on the internet is founded. For example, a 2012 web survey of Danish residents included four political knowledge questions. In response to a follow-up question, no less than 22% of respondents admitted having searched for the correct answer on the internet for at least one of the knowledge questions they had to answer (Jensen & Thomsen, 2014). The authors concluded that "cheating in terms of Internet browsing is a systematic feature of Web surveys of political knowledge" (2014, p. 3348). As it seems, this claim is probably not excessive—according to recent research, the proportion of self-reported cheaters mostly fall in the 10–20% range (e.g. Clifford and Jerit, 2014, 2016; but see Gooch, 2015 for a strikingly different conclusion). Interestingly, the share of "cheaters" may be highest (up to 40%) in student samples (Clifford & Jerit, 2016, pp. 864–865). This squares well with the results of further studies suggesting that cheating is consistently related to higher test performance, and that performance is driven by extrinsic motivation such as the prospect of reward for good test achievement (Diedenhofen & Musch, 2017) and intrinsic motivation such as political interest (Munzert & Selb, 2017).[1]

A second conclusion from recent research is that respondents who give untruthful answers tend to display "aberrant" response times (hereafter RTs). In a nutshell, aberrant RTs are "outliers" defined in a relative sense, that is, as values which deviate from an individual's expected RTs, in other words from predictions based on both respondents' and items' characteristics. We come to terms with this issue in greater detail below.

2.2 Practical solutions to the cheating problem.

To date, there have been at least eight different proposed "solutions" to the problem of cheating in knowledge tests. Although this list probably does not exhaust the possibilities, I believe that it includes most of the current proposals to deal with cheating in factual knowledge survey tests.

1. *Allow self-reports of cheating.* As sketched out above, it is not unusual for 10–20 percent of respondents to admit cheating when asked shortly after they completed a political knowledge test. It is certainly possible to include such a question in any questionnaire, but its real relevance and the risk of deteriorating the normal rapport between interviewer and interviewee (in CATI surveys) are issues worth discussing.[2]

[1] Indeed, students may be expected to have greater political interest than other groups; and they are likely to internalize their desire to succeed in academic tests as a general disposition toward testing. Hence students (who are also members of the "net generation") may feel special pressure for achievement and could be especially prone to cheating (Davis, Drinan, & Bertram Gallant, 2009; Yu, Glanzer, & Johnson, 2017, chap. 3).

[2] More than thirty years ago, Luskin (1987, p. 892) warned against the risk that including "(too) many questions aimed at cognition may make the interview seem too much like a test". Besides, one major drawback of the self-report question is that it does not (and probably cannot) distinguish between those respondents who had no idea of the correct answer and those who knew the answer (or would have made an educated guess) but simply wanted to check whether they were right.

2. *Control for internet browsing.* Applications and scripts such as PageFocus (Diedenhofen & Musch, 2017) can be integrated in online questionnaire pages to detect when respondents "defocus" and "refocus" on the questionnaire, i.e., when they possibly leave the questionnaire to look up the answers on various internet sites. In turn, these devices can generate a popup message that appears whenever respondents abandon the questionnaire page and asks respondents not to look up answers. This procedure has been shown to be successful in reducing cheating behavior (Diedenhofen & Musch, 2017).

3. *Give explicit instructions.* Instead of asking afterwards whether respondents looked up solutions, an online questionnaire—and a factual knowledge test in particular—can encourage honest responding by making clear from the start that searching for answers from outside sources is not permitted (e.g. Motta, Callaghan, & Smith, 2017). In the Selects 2015 online questionnaire that will be used in the empirical section, the political knowledge questions were preceded by a "soft" instruction not to look up answers: "Please simply choose the answers which you consider correct (without checking up on them)". Such instructions could certainly be made more explicit, and hopefully more effective, by briefly explaining the purpose of the knowledge test. The explanation should make clear that cheating behavior is not prohibited because it is morally bad or socially unacceptable, but because it compromises the scientific validity of the test itself.

4. *Impose time limits.* Some authors have argued that limiting the time available to answer knowledge questions (for example to 30, 45 or 60 seconds) will prevent respondents from consulting outside sources (e.g. Marshall, 2019; Prior, 2014; Prior & Lupia, 2008; Prior, Sood, & Khanna, 2015; Strabac & Aalberg, 2010). When such time limits are used, the ability to provide correct responses is significantly reduced in comparison to conditions where time is virtually unlimited (e.g., 24 hours to answer 14 questions as in Prior & Lupia, 2008) and consulting "references" it not openly discouraged. This suggests that the very possibility of cheating is cut down by time limitation.

5. *Use of pictures.* In recent times, a growing number of studies have tested political knowledge through the use of visual items, reasoning that knowledge of politics is not reducible to verbal or semantic information alone (e.g. Prior, 2007, 2014). For example, some questions ask respondents to identify a political leader represented in a picture or to indicate which of several pictures corresponds to a named leader. Interestingly, some of these efforts have explicitly aimed to reduce cheating, based on the premise that online search is inherently (and technically) more difficult for visual information than for textual information. However, the evidence that visual knowledge questions prevent cheating is mixed at best (Munzert & Selb, 2017; Prior, 2014; Strabac & Aalberg, 2010).

6. *Use of obscure questions.* Motta et al. (2017) have demonstrated that posing "impossible" questions (e.g., knowing in which year an obscure Supreme Court judgment was rendered or a shadowy treaty was signed) can serve to identify cheaters (see also B. Smith, Clifford, & Jerit, 2020; but see Bullock, Gerber, Hill, & Huber, 2015). Because these "catch" questions are virtually impossible to answer correctly, all respondents who do provide a correct response are highly suspected of cheating. However, it can be argued that a catch question should be embedded among several other items of varying difficulty and that the strategy is inappropriate for a questionnaire comprising few items.

7. *Measurement of knowledge as a latent trait.* Latent trait theories such as Item-Response Theory (IRT) have proven increasingly relevant for the measurement of factual knowledge. The recent integration of RTs in IRT research has extended the three-parameter logistic (3PL) model to a four-parameter model which may be applied to the problem of test cheating (e.g. van der Linden, 2011; Wang & Hanson, 2005; Wise & DeMars, 2006). However, for such models to be fully effective in measuring knowledge as a latent trait, it is probably necessary to have at least 8 or 10 test items, which could be interspersed in the questionnaire to prevent measurement of political knowledge from sounding like a "school test".

8. *Detect and correct cheating through an analysis of RTs.* A final strategy is to analyze RTs data to identify the most suspicious cases. The basic principle of this strategy is to build a predictive model of RTs with sufficient explanatory power. Deviations from predictions of this model (i.e., residuals) will then indicate to what extent observed RTs are different from expected RTs (see Section 3.3). Finally, an examination of relationships between residuals and item knowledge will ascertain whether there exist patterns of (correct vs. incorrect) answers, or whether the answers are distributed essentially at random, i.e., independently of RT residuals. This will provide a basis for the identification of suspected "cheaters" and the rescaling of knowledge scores.

The eight solutions above can be distinguished according to whether they are implemented "ex-ante" (i.e., in the interviewing process) or "ex-post" (i.e., at the data analysis stage). As can be seen, most of them (solutions 1–6) are of the ex-ante type. Only the last two (solutions 7–8) allow researchers to deal with cheating post factum. As no time machine has been invented yet, there is no way to change questions that have already been asked. However, the solution (#8) proposed in this article allows researchers to identify cheating on a question-by-question basis and provides them with a method for correcting knowledge scores which have likely been inflated by cheating. This double function may represent a significant innovation in the methodology of political knowledge—in comparison with previous efforts

which have focused on making cheating more difficult or on identifying individual cheaters. Accordingly, I argue in the next section that the analysis of RTs is a promising strategy for alleviating the problem of cheating behavior in knowledge tests.

3 Response times: Theoretical and methodological aspects

3.1 What are response times, and how useful are they?

Time is an important ingredient of survey responses. Admittedly, this ingredient is still uncommon in survey research and, more generally, it has long been overlooked by social science scholars. The main reason for this neglect is that RTs have been largely unavailable as operational measures until the advent of computer assisted interviewing and online surveys (Bassili, 2000). Nowadays, RTs are produced automatically in most computerized surveying systems, using internal clocks to record the time elapsed between two keystrokes, i.e., the procedure required to switch from one question to the next (hence the name "keystroke data" sometimes given to such information). Interestingly, the speed (or slowness) with which respondents answer survey questions can reveal a number of important features of questions, respondents, or both. First, the time necessary to provide an answer tells us something about the underlying concepts and conceptual associations which are retrieved to produce a response. In particular, response times (or "response latencies") are used to measure the accessibility or strength of attitudes and other mental concepts tapped by a survey question (Fazio, 2001; Meyer & Schoen, 2014; Mulligan, Grant, Mockabee, & Monson, 2003). They may also indicate whether respondents hold ambivalent or unstable attitudes and beliefs (Heerwegh, 2003; Van Harreveld, van der Pligt, de Vries, Wenneker, & Verhue, 2004). In turn, stronger, more accessible and more stable attitudes have been shown to be more resistant to persuasion attempts and more predictive of future behavior (e.g. Dalege et al., 2016; Erber, Hodges, & Wilson, 1995; Fabrigar, MacDonald, & Wegner, 2005; Fazio, 2001; Fazio & Williams, 1986; Glasman & Albarracín, 2006). In short, more accessible constructs are also "stronger" in one sense or another.[3] Second, RTs can reveal some characteristics of the questions themselves, for example whether they are badly worded or too difficult to answer (Bassili & Scott, 1996; Yan & Tourangeau, 2008), whether they address issues about which people have little or no preexisting knowledge, or whether they focus on "sensitive issues" eliciting social desirability biases (Holtgraves, 2004; M. Johnson, 2004). In this respect, RTs offer auxiliary information to assess and improve the quality of survey questions. Third, RTs inform us about the survey respondents, inasmuch as they display a general tendency to provide fast or slow answers. Such as tendency may be related to interviewees' propensity to "satisfice" in their survey answers (Krosnick, 1991; Turner, Sturgis, & Martin, 2015), to their degree of experience or interest toward the topics of the survey, but also to socio-demographic characteristics such as age (Harms, Jackel, & Montag, 2017; Marquis, 2014; Yan & Tourangeau, 2008). Finally, the time taken to answer survey questions may point to features of the survey context, such as the presence of distractions in the respondent's environment or properties of the interviewing mode. Some authors (e.g. Couper & Kreuter, 2013) show that response latencies can be used to check whether interviewer characteristics affect the time to answer questions and to identify potential problems with specific interviewers. More importantly, research suggests that the average interview length may be shorter for web surveys than for telephone or face-to-face surveys (e.g. Kinsey, Iannacchione, Shook-Sa, Peytcheva, & Triplett, 2013, p. 36; Ansolabehere & Schaffner, 2014, p. 286; Watson, Porteous, Bolt, & Ryan, 2019, p. 837; but see Fricker, Galesic, Tourangeau, & Yan, 2005, pp. 385–387). Against this background, out-of-range (usually very long) answers may serve as evidence of "cheating behavior" on the part of web respondents who look for answers to factual knowledge questions on the web (Burnett, 2016; Munzert & Selb, 2017).

3.2 Timepieces

Needless to say, RTs are an imperfect indicator of construct accessibility. The main reason is that, depending on situation, there are other factors that may affect RTs above and beyond the impact of accessibility per se. For illustration purposes, let us take the case of attitudes.[4] As a starting point, I should make clear that, in my view, attitudes are not fixed and transparent entities, but "temporary constructions" (Lord & Lepper, 1999; Schwarz, 2007; E. R. Smith & Conrey, 2007; Wilson & Hodges, 1992; Zaller & Feldman, 1992).[5] To be sure, some essential part of many attitudes is made up of rather immutable things—beliefs, affects, or preferences which hardly ever change. However, I assume that attitude reports (i.e., that which is expressed in opinion surveys to "reflect" underlying attitudes) are based on a limited sample of all available considerations. Importantly, "true enduring evaluations" may be just one element in that sample,

[3] As a matter of fact, accessibility is related to most other dimensions of attitudes strength such as importance, centrality, connectivity, stability, certainty, and extremity (see Bizer & Krosnick, 2001; Dalege et al., 2016; Holbrook & Krosnick, 2010; Krosnick, 1989; Mulligan et al., 2003; Prislin, 1996; Van Harreveld & van der Pligt, 2004; Wegener, Downing, Krosnick, & Petty, 1995).

[4] I use "attitudes" here in a generic sense for any mental concept which is about an identifiable object and has both a cognitive and an affective (evaluative) component.

[5] This view holds that all attitudes are (at least partially) constructed; therefore, it is not equivalent to Converse (1964) "black-and-white" conceptualization of attitudes.

along with more ephemeral elements of the response context (Schwarz, 2007). In addition, elements which have been made more accessible by recent or frequent activation have a higher probability of being sampled into the set of considerations used to produce an attitude report (Wyer, 2008; Zaller & Feldman, 1992).

From the perspective of response latencies, it is thus important to specify the process by which attitude reports are created. As some scholars (e.g. M. Johnson, 2004; Tourangeau, Rips, & Rasinski, 2000; Van Harreveld et al., 2004) have argued, answering survey questions is a multistep and multifactorial process. In that process, one may single out the following constituents:

- Interpretation: For a number of reasons, it may be more or less difficult (or simply time-consuming) for a given respondent to understand some question and to ascertain how her existing beliefs and attitudes bear on that question.
- Retrieval: After making sense of a question, respondents initiate the response process by retrieving relevant information from memory. This is where accessibility is expected to make a difference, since the retrieval of more accessible memory contents occurs faster. However, as the time pressure of the interview is a further incentive to satisfice rather than optimize (see Krosnick, 1991; Turner et al., 2015; Vannette & Krosnick, 2014), only the very first elements that "come to mind" are likely to be taken into account for building a response.
- Editing: Once memory contents have been retrieved, they must be assembled into an overall judgment or response.[6] Likewise, it may require additional time to fit the response to the answer categories provided in the survey, and to adjust it to perceived demands of the response context. For example, RTs are usually longer on sensitive social issues such as race, when normative pressures elicit social desirability concerns (Holtgraves, 2004; M. Johnson, 2004). It is also at this step that cheating is expected to occur (for example when respondents ask another person or look on the internet for the correct answer). In the case of political knowledge questions, cheating is formally equivalent to editing a "don't know" answer.

Each of these "processes" accounts for some part of the total latency i measured for the answer given by individual j to question k.[7] In other words, a latency can be conceptually decomposed into several "timepieces", as M. Johnson (2004) called them. Empirically, it might prove impossible for nonexperimental research to separate these RTs bits according to process. Instead, the best one can do is to control for factors which are assumed to influence the time required to complete each of the interpretation, retrieval, and editing steps. To simplify, one may distinguish between item factors (e.g., item length), personal factors (e.g., education), and contextual factors (e.g., whether the survey was completed online or by telephone). Some factors are typically related to one particular step in the survey response process. Thus, for example, the formulation of a question can possibly affect RTs only through the interpretation mechanism. But think of a personal factor such as age; in this case, older respondents can be expected to take more time to proceed with each step—from understanding to retrieving to editing. Therefore, it is important to control for all types of factors—personal, item, and contextual. Although it is difficult to determine how these factors are related to particular "timepieces", it seems safe to assume that item and contextual factors will tend to have special influence on the interpretation and editing steps. It is thus advisable to examine RTs on a wide range of items and across different types of context.

3.3 Response times and political knowledge tests

One important consequence of the development of web surveys in recent years is that respondents are given a greater control over the whole questionnaire answering process and over the quality of responses they provide. Among the unique features of web surveys are the possibilities for respondents to determine the pace of the interview, to change their answers, or to switch back and forth between questions or questionnaire sections (though this option is not always available). In this context, RTs have been used, for example, to monitor how responses are affected by the opportunity of changing one's answers, either as a result of additional information (e.g. Heerwegh, 2003) or as attempts to correct for manipulation errors or confusion caused by unclear questions (e.g. Stern, 2008). Likewise, RTs are helpful to determine the relevance and influence of pictures on performance in factual knowledge tasks (Munzert & Selb, 2017; Prior, 2014; Sass, Wittwer, Senkbeil, & Köller, 2012). Only recently, however, have political scientists become fully aware that RTs can serve as a means for detecting cheating behavior, in particular in factual political knowledge tests. Building on recent developments in Item Response Theory (e.g. Fox, Klein Entink, & van der Linden, 2007; Marianti, Fox, Avetisyan, Veldkamp, & Tijmstra, 2014; van der Linden, 2011; van der Linden & Guo, 2008) and on earlier attempts to detect cheating in educational and professional settings (for a review, see Cizek & Wollack, 2017), researchers have begun to investigate the phenomenon of cheating in online political surveys by relying on response times.[8] One notable approach derived

[6] Thus it is that ambivalent attitudes are associated with longer RTs, because the integration of conflicting considerations takes more time (Van Harreveld et al., 2004).

[7] To be sure, some part of latency i is also determined by systematic and nonsystematic measurement error (sampling error, stammering by the interviewer or interviewee, presence of distractions, bad telephone or internet connection, and the like).

[8] Despite (or because of) their high sophistication, it is doubtful that IRT models can be applied to a political knowledge test relying on as few items as the 2015 Selects survey does (i.e., four). This

from this research is based on three main assumptions. First, answers which are reported faster stem from mental objects which are more accessible in memory. Second, to express the "true" accessibility of underlying mental concepts, RTs must be compared to a baseline of RTs indicating how fast or slow a respondent tends to answer questions in general. Put another way, it is possible to model an expected response time, by combining information from individuals, items, and the response context. We thus need a comprehensive, multifactorial model to predict RTs. Third, observed ("real") RTs can in turn be compared to predicted RTs to provide an estimation of how much actual RTs deviate from their expected values. In technical terms, we may speak here of "residuals", provided that expected values were generated by a predictive empirical model. When residuals are negative, actual RTs are "faster than expected"; when they are positive, actual RTs are "slower than expected". Assuming that cheating to knowledge questions (for example by looking up answers on the internet) takes some time, residuals with particularly high positive values can be indicative of cheating behavior, especially when combined with corroborating evidence (see section 6).

4 Measurements

4.1 Empirical data

My case study will focus on the Selects post-electoral survey conducted shortly after the Swiss federal elections of October 2015 (Selects, 2015). The total sample (N = 5337) is made up of Swiss citizens who were interviewed in the six weeks following the election, either by telephone (computer-assisted telephone interviewing, CATI, 18%) or online (computer-assisted web interviewing, CAWI, 82%). The final response rate attained 44 percent (AAPOR RR#2). To ensure representativity, stratified random sampling was used, i.e., the sample was stratified by the 26 cantons. Whenever possible, weights were applied in the following analyses to correct for oversampling in some cantons.[9]

All respondents selected in the initial sampling frame were contacted by postal mail one week before the elections to inform them about the survey. The day after the elections, they received a participation letter with a specific web address and personal login information, kindly requesting them to take part in the online survey. The CATI survey started two weeks later, as sample members who had not yet participated in the online survey were contacted by phone. The survey ended six weeks after the elections.

Provided that the online survey was presented as the standard survey mode, it is probably excessive to say that sample members self-selected into the CAWI group. However, since a fifth of respondents answered the survey by phone because they did not want to, forgot to, or were unable to participate in the online survey, it is to be expected that the CAWI and CATI samples are structurally different. As a matter of fact, in comparison to CAWI, the CATI sample comprises a significantly higher proportion of females (58% vs 49%), of older people (M = 60.2 vs. 47.5 years), of people with lower education (62% vs. 39%), and of people who have no internet access or never use it (34% vs. 1%). In contrast, differences in terms of political interest are statistically significant, but substantially trivial, judging for example by the share of politically uninterested people (7% vs. 4%) or very interested people (20% vs. 20%).[10]

4.2 Item selection

To explore the structure of RTs in the 2015 election, I selected 20 items in addition to the four knowledge items. These items will serve as "filler latencies" to estimate the influence of individual, contextual, and item-specific factors on the duration of RTs, using the procedure sketched out in Section 3.3. The main criterion for item selection was that the number of missing data (don't know and refusal answers) should not exceed 5 percent of all cases.[11] Another criterion was to choose different types of questions, that is, questions with varying degrees of inherent "difficulty" and requiring different kinds of mental processes to answer (see Mayerl & Urban, 2008, pp. 63–88). As it turns out, the item-specific amount of time necessary to yield a response depends on the average accessibility of relevant attitudes and beliefs, but also

is because estimating a latent trait (in most cases, the test takers' "ability" in some domain) is uneasy with so few independent observations (DeMars, 2010, pp. 34–37; van der Linden & Pashley, 2010).

[9] Simple random sampling was carried out within each of the 26 Swiss cantons, based on the sampling frame of the Swiss Federal Statistical Office (see Lutz, 2016, pp. 73–74). The sample was constructed to include an oversampling of the least populated cantons and to ensure that at least 80 people were interviewed in each canton; likewise, samples were augmented to 800–1000 interviewees in three cantons (Zurich, Geneva, Tessin), with the additional costs being covered by the respective cantons. The survey length was about half an hour (M = 32.5 min.; Median = 26.9 min.; SD = 46.6 min.), with shorter and more variable interview durations for CAWI respondents (M = 32.0; Median = 25.0; SD = 51.2) than for CATI respondents (M = 34.9; Median = 32.9; SD = 9.6).

[10] Gender: $\chi^2(1) = 23.0$, $p < 0.001$; Age: $F(1, 5335) = 432.0$, $p < 0.001$; Education (3 categories): $\chi^2(2) = 187.2$, $p < 0.001$; Internet use (4 categories): $\chi^2(3) = 1370.4$, $p < 0.001$, Political interest (4 categories): $\chi^2(3) = 13.5$, $p < 0.01$, Cramer's $V = 0.05$.

[11] The choice of a DK response is usually related to RTs, but the direction of the relationship is debated. While the literature on satisficing in surveys (Krosnick, 1991) seems to imply that RTs are reduced on DK responses, other research suggests that DK responses should be related to longer latencies (Bassili, 1995; Turner et al., 2015, pp. 341–342). Because it is uncertain how DK answers will bias RT measures, I decided to limit their amount to a small proportion.

on which "timepieces" these RTs are primarily involved in (see Section 3.2).

Following these guidelines, ten "attitudinal" items and ten "behavioral/factual" items were retained along with the four knowledge items. As can be seen from Table 1 (column 5), raw RTs for these items show great variation: mean values are comprised between 4 and 45 seconds.[12] Much of this variation is understandable enough—it takes less time to report whether one has a mobile phone (B6) than to give one's opinion on social expenses (A1) or to indicate the number of governing parties (K1). However, as indicated by the standard deviations (column 6), part of this variation is probably artificial. Six items have a SD longer than one minute, suggesting the presence of strong outliers, i.e., respondents with very long RTs. Figure A3 in the Online Appendix confirms this speculation.

4.3 Response time trimming

Unlike other studies, I decided not to resort to the log method to minimize the influence of outliers. Instead, I used a procedure of data trimming to remove extremely long RTs. In part, extremely long durations occur for artificial reasons, due to the presence of distractions in the respondent's environment (e.g., answering a mobile phone call, surfing on the internet, cooking, looking after children; see Clifford & Jerit, 2014). However, the increasing use of web surveys in the last two decades has had several consequences as far as outliers are concerned. For one thing, the CAWI method allows respondents to interrupt (practically as many times and as long as they wish) the process of questionnaire completion. In other words, respondents can "make a pause"—to have a snack or a nap, chat with friends, watch a football match, or even go to work or go to sleep. All this implies that very long RTs (say, more than one minute for a standard opinion question) are likely to stem from the interference of "external" causes (pausing, distractions, etc.) rather than from "internal" causes such as ambivalence, goodwill (optimizing), social pressure, or sheer confusion.

It is thus necessary to trim the RTs data in order to correct for upper-bound outliers. The trimming occurs in two stages. First, in a "winnowing" step, threshold values are computed for each item as the bottom 2nd and the top 98th percentiles, respectively. In order to control for differences between interviewing methods (CAWI vs. CATI), as well as differences between linguistic regions, these thresholds are computed within each method × region group. Values lower or higher than these values are set to the respective thresholds. From these adjusted RTs one gets three statistics for each item: its mean, median, and standard deviation; these statistics are also specific to each interviewing method and linguistic region.

In a second stage, these statistics are used to run the final trimming operation: For all RTs greater than the mean plus 3 standard deviations, the corrected RT is set to the median.[13] This M3SD rule entails setting to the median about 3% of cases, with slight differences from one variable to the next. In contrast to previous outlier treatments based on the M3SD or similar rules, outliers are not coded as missing observations; nor are they trimmed in the proper sense, since they are assigned the median value. This is because very long latencies are not considered a fixed attribute of some individuals, but are most likely due to extraneous factors that extend RTs above and beyond the time required to mull over a difficult question and even to "cheat" to factual knowledge questions. The last two columns of Table 1 give a short description of the trimmed RTs, showing their obvious differences with the raw data.

To get an overall picture of the corrected measures, I examined the pattern of correlations among them (all $p < 0.001$). Correlations within filler (attitudinal/behavioral) items (mean $r = 0.36$) or within knowledge items (mean $r = 0.24$) are noticeably higher than between filler and knowledge items (mean $r = 0.17$). In addition, a principal component analysis of RTs to all 24 items suggests that all RTs in the CATI sample have the same dimensionality (i.e., knowledge items do not differ from filler items), whereas a strong difference between knowledge and filler items emerges in the CAWI sample (see Online Appendix A.2). This suggests that the mode of interviewing plays a crucial role not only in the overall duration of RTs (see Figure A.3), but also in their general interrelationships. Again, this emphasizes the need of a general predictive model including individual, item-specific, as well as contextual factors.

4.4 Knowledge questions in the Selects survey

Turning now to political knowledge questions, I review the four items used in the 2015 Selects survey. Here are the questions (answer options in parentheses, correct response italicized):

- K1: Can you tell me how many parties are represented in the Federal Council (Swiss government)? Are there… (2 parties; 3 parties; 4 parties; *5 parties*; 6 parties; Don't know)
- K2: Who is the president of the Confederation this year? Is it… (*Simonetta Sommaruga*; Johann Schneider-Ammann; Didier Burkhalter; Doris Leuthard; Ueli Maurer;

[12] Let us note that B1 is actually a multiple-question item made up of 5 questions for which no separate RTs were measured. Accordingly, RTs are notably longer for this item.

[13] More sophisticated truncation methods are available, but the "M3SD" rule has been recommended as an effective outlier-removing method (Heerwegh, 2003; Huckfeldt, Levine, Morgan, & Sprague, 1999; Mulligan et al., 2003). A similar, "M2SD" rule has also been proposed (e.g. Bassili & Scott, 1996; Couper & Kreuter, 2013; Meyer & Schoen, 2014), along with truncation below and above some percentile (Harms et al., 2017; Yan & Tourangeau, 2008) and still other specifications (see Yan & Olson, 2013, p. 79).

Table 1
Selection and description of items for the overall analysis of response times (in seconds).

Item	Var. Name	Description of item	% Missing	Raw RTs		Trimmed RTs	
				Mean	SD	Mean	SD
A1	F15420	Opinion on social expenses	4.8	32.7	346.9	23.4	12.2
A2	F15430	Opinion on EU membership	4.9	15.2	15.1	13.6	7.9
A3	F15440	Opinion on equal chances for foreigners and Swiss	2.9	21.5	24.4	18.9	11.0
A4	F15470	Opinion on protection of the environment versus economic growth	4.0	21.3	74.5	17.5	9.6
A5	F15480	Opinion on taxes on high incomes	4.1	19.2	63.4	15.8	9.3
A6	F15490	Opinion on nuclear energy	3.9	13.1	23.1	11.4	6.6
A7	F15760	A child should have respect for elders or be independent	3.9	15.5	44.7	11.8	5.9
A8	F10100	Interest in politics	0.6	12.3	36.6	10.2	6.1
A9	F13700	Satisfaction with functioning of democracy	2.0	13.4	18.9	11.6	6.3
A10	F14600	Evaluation of the state of the economy	1.7	13.3	52.8	11.2	6.8
B1	F13401–05 (5 items)	Attention to news: radio/TV newspapers free newspapers (e.g., 20 Minuten) websites/ online newspapers social media (e.g., Facebook, Twitter)	4.2	36.1	42.7	33.0	18.3
B2	F13300	Discussion about elections with family, friends, colleagues	0.6	11.3	16.0	9.8	4.3
B3	F11100	Participation in current national elections (2015)	0.8	21.1	29.2	19.2	8.8
B4	F21310	Respondent's highest educational attainment	1.6	27.6	35.4	24.7	13.8
B5	F20600	Does respondent have a landline phone?	1.0	8.4	11.5	7.5	3.1
B6	F20601	Does respondent have a mobile phone?	0.8	4.2	8.4	3.7	1.5
B7	F20602	How often does respondent use the internet?	0.7	12.6	427.6	8.5	5.0
B8	F20300	Year since respondent lives in current canton	1.0	24.3	33.8	21.2	11.1
B9	F20500	Household size	0.8	6.5	8.1	5.7	3.0
B10	F20210	Citizenship which respondent held at birth	0.8	10.2	18.0	9.0	3.4
K1[a]	F15900	Knowledge Q1: Number of parties in the Federal Council	8.0	44.5	638.6	30.1	24.6
K2	F16000	Knowledge Q2: Name of the president of the Confederation	8.5	22.3	39.0	17.9	15.6
K3	F16100	Knowledge Q3: Required number of signatures for a federal initiative	11.2	25.7	1037.7	11.9	10.5
K4	F16300	Knowledge Q4: Party which had most seats in the National Council before the election	17.1	17.6	34.6	13.3	10.0

[a] Missing observations for knowledge items (DK/NA) are added to incorrect answers in the final measurement of these variables.

Don't know)

• K3: How many signatures are required to launch a popular initiative? (50,000; *100,000*; 150,000; 200,000; 250,000, Don't know)

• K4: Which party had the most seats in the National Council (lower chamber of the Parliament) before the elections? (*SVP*; CVP; FDP; SP; BDP; Don't know)

All questions focus on party politics and institutional matters at the national level. The proportion of correct answers varies between 43% (K1) and 69% (K2 and K3); hence, all questions are of medium difficulty.[14] To be sure, the rather narrow range of topics and difficulty found in the knowledge questions of the Selects survey is suboptimal, especially in comparison with other surveys measuring political knowledge through a higher number of items, covering a wider range of subjects and with more diverse levels of difficulty (e.g. Delli Carpini & Keeter, 1996; Prior & Lupia, 2008; Rapeli, 2014; Reichert, 2010). This may limit the generalizability of the conclusions I will draw from my analysis of the Selects survey.

However, when combined, the four items are discriminant enough to yield a roughly normally distributed scale ranging from very low to very high political knowledge. On the basis of a simple additive scale, the number of correct answers is distributed as follows: 0 = 5%; 1 = 15%; 2 = 27%; 3 = 33%; 4 = 19% (percentages do not add to 100 because of rounding). Admittedly, the reliability of the scale is poor (Cronbach's $\alpha = 0.37$, $\lambda_2 = 0.37$). While it is not unusual for scales based on a small number of dichotomous items to have low reliabilities, it is perhaps more significant to note that, after slightly declining between the 1995 and 2011 election surveys ($\Delta\alpha = -0.07$), the reliability coefficient falls sharply between the 2011 and 2015 elections ($\Delta\alpha = -0.12$). Likewise, the ratios of correct responses are consistent with those obtained for the previous five elections (1995–2011), but with an important exception. As a matter of fact, the share of 69% of correct responses to K3 exhibits an inflation of 30 percent from the average of the years 1995–2011, when online interviews were not in use. Explaining this unusual rate of success will be one of the primary goals of my empirical analysis.

Figure 1 shows the distribution of RTs for the four items. It is immediately apparent that RTs are shortest on K3 and K4. In contrast, RTs for K2 are longer and more dispersed. As it seems, some people really struggled to recall (or guess) who the Swiss president was in 2015. In this case, however, this cognitive effort tended to produce at least as many correct answers (69%) than K3 and K4 (69% and 65%, respectively). Finally, the case of K1 (number of parties represented in government) stands out in two respects. On the one hand, this item displays by far the longest RTs of all items; on the other hand, it is also the item on which correct answers were decidedly the least frequent (43%). Thus, by all standards K1 can be considered the "most difficult" item.

5 Cheating in political knowledge tests?

5.1 Overall procedure

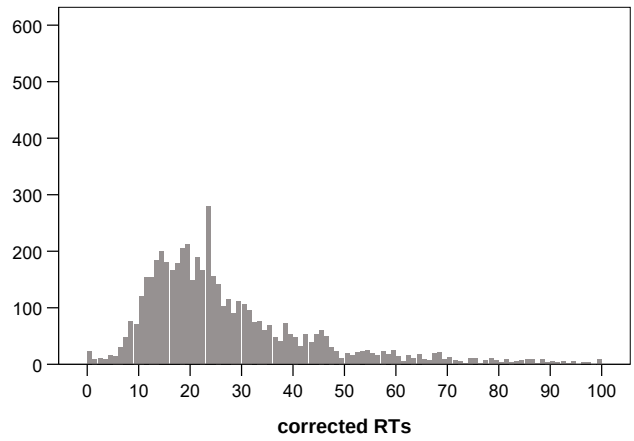
Before I proceed with the substantial part of my empirical analysis, let me emphasize very clearly that, in the context of this study, there is no irrefutable evidence that someone has cheated in answering a survey question. Such evidence simply does not exist! In other words, cheating behavior cannot be formally *proved* using the RT method, in particular because unexpectedly long RTs can stem from other causes than cheating. Instead, one can scrutinize information to substantiate a *suspicion* that cheating has occurred in some cases; however, this evidence will be hardly compelling enough to warrant a conviction. Against this background, my first strategy is to explore the relationship between RTs and knowledge of items (correct vs. wrong answers). In particular, do respondents who provide late answers "get it right too often"? Are these correct answers somehow "too good to be true" in consideration of the expected pattern of responses?

To some extent, this expectation is borne out by simply inspecting the overall relationship between RTs and the share of correct answers to the four knowledge questions. This analysis (shown in Online Appendix A.1) enables to identify a small fraction of respondents who might have engaged in cheating behavior—i.e., who could have got the correct answer from the internet or some other outside source. In a nutshell, the share of correct answers to two knowledge questions (K2 and K3), while generally decreasing as a function of RTs, suddenly peaks in the higher percentiles of the RT distributions. It may thus be tempting to jump to the conclusion that responses to these two items are plagued by cheating.

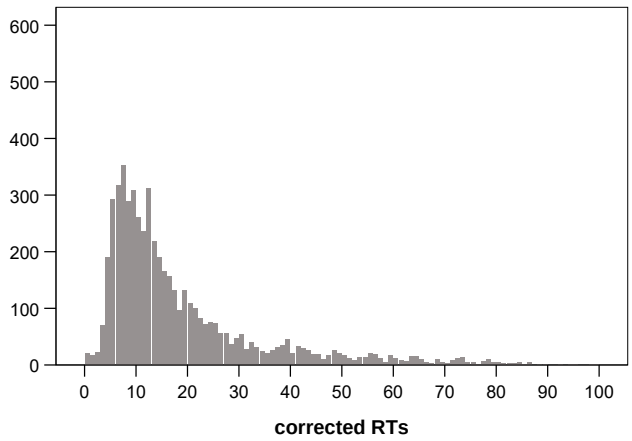
To see why such a conclusion may be unwarranted or simply false, let us take the example of age. Older respondents are expected to perform better in political knowledge tests than their younger counterparts (Prior & Lupia, 2008; Rapeli, 2014), but they also tend to take more time to answer survey questions (Fricker et al., 2005; Marquis, 2014; Wingfield, 1998; Yan & Tourangeau, 2008). Accordingly,

[14] The difficulty parameters from an IRT analysis (3-parameter logistic model) suggest a similar conclusion: Difficulty is highest for K1 ($b_1 = 0.65$) and lowest for K2 ($b_2 = -1.10$), with the other items falling in between but closer to K2 ($b_3 = -0.71$; $b_4 = -0.37$). Accordingly, it takes more than average knowledge (i.e., the assumed latent trait) to reach a 50% ratio of correct answers to K1, and less than average knowledge for the other items (see DeMars, 2010, p. 5). As it turns out, a knowledge scale measured from factor scores of the IRT model is highly correlated ($r = 0.986$) with the simple additive scale described below. In the perspective of building a corrected knowledge scale (see Section 6), it is more convenient to keep working with the additive scale.

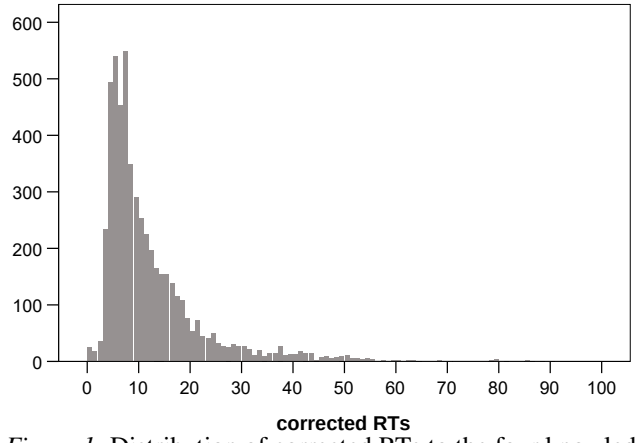
(a) K1: Number of parties in the Federal Council (M = 24.0; SD = 24.8; % correct = 42.8)



(b) K2: Name of the president of the Confederation (M = 14.4; SD = 15.6; % correct = 69.0)



(c) K3: Required number of signatures for a federal initiative (M = 8.2; SD = 10.2; % correct = 68.9)



(d) K4: Party with most seats in the National Council before election (M = 9.1; SD = 9.7; % correct = 65.4)

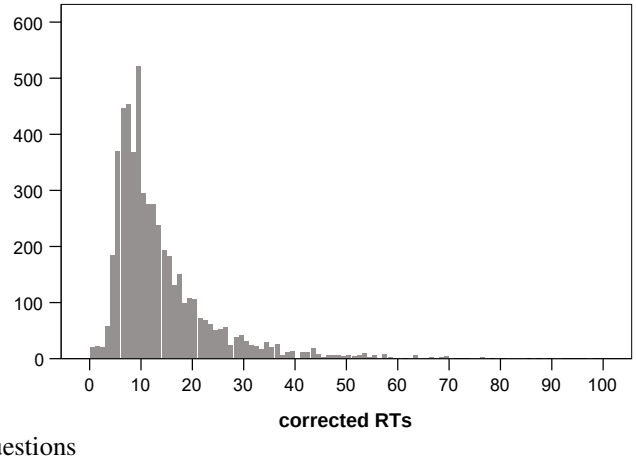


Figure 1. Distribution of corrected RTs to the four knowledge questions

if older respondents are indeed overrepresented among response laggards, they will contribute to the "bump" in correct answers observed in the last percentiles of the RT distribution for some items. As a result, they will be conflated with real "cheaters" and misidentified as potential "cheaters". In other words, the fact that older people are both slow and successful at answering knowledge questions cannot be interpreted as evidence of cheating on their part. This conclusion seems crystal clear in the simple example of age. However, there is a whole range of other individual and contextual variables which can also affect the tendency to provide quick or slow answers. The problem is that, as a rule, everyone will tend to answer all kinds of questions in their own time. What we need, then, is to compute a "baseline" of response times, tailored to each individual according to idiosyncratic and social characteristics, and to each survey question according to its internal properties (e.g., length, position in the questionnaire). In turn, this baseline will allow us to ascertain to what extend the responses times related to each individual and each survey question deviate from an "usual" pattern of responding.[15]

To make things more comprehensible, let us focus first on

[15]Originally, at a time where RTs enjoyed the exclusive attention of social and cognitive psychologists, special emphasis was put on interindividual differences. For practical purposes, it was suggested to compute an individual baseline speed on the basis of filler latencies, using as many of them as possible and computing their mean value (Fazio, 1990, pp. 78–79, 88–89), and then to compute a difference score index, namely to subtract the baseline from the actual RTs (e.g. Marquis, 2014). Alternative specifications have been proposed, such as using residuals from a regression of RTs on baseline speed (Mayerl & Sellke, 2005; Mayerl & Urban, 2008, pp. 71–81; Meyer & Schoen, 2014). However, these alternative measures do not change anything substantial to the difference score index, and they tend to have both positive and negative consequences. In the

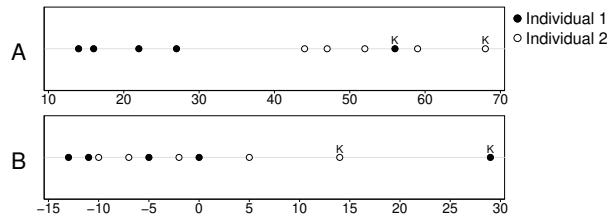


Figure 2. Raw RTs (Panel A) and RTs corrected for individual baseline speed (Panel B)

interindividual differences. Figure 2 (Panel A) displays the response times of two individuals on a series of five items. In both cases, the knowledge item (K) has a longer response latency than the other (filler) items, but individual 2 has taken an additional 12 seconds to answer it, compared to individual 1. Next, one can compute an individual mean of all RTs to get an approximate "baseline" of response times. Importantly, the deviations from this mean (Panel B of Figure 2) tell a different story. Now, in relation to their respective baselines, it appears that individual 1 has been about twice slower than individual 2 to answer the knowledge item.

This leads us to speculate about the reasons of this interindividual difference. At this stage, "looking up the answer on the internet" is only one potential cause among a myriad of other possibilities. A multivariate analysis is therefore required to ascertain which individual characteristics underly differences in response speed of the kind sketched out in Figure 2. Individuals 1 and 2 are certainly not isolated cases; rather, their RTs are probably representative of the effect of larger social, psychological and biological mechanisms. For example, individual 2 may be older and less educated, and she may share with individuals from the same stratum the propensity to provide slower responses. Likewise, as a younger and more educated respondent, individual 1 may be quick to answer filler (e.g., attitudinal) items, while at the same time struggling to answer the knowledge question (for example because it requires long-term experience with practical institutional issues).

In sum, we need a model capable to account for both individual and item-specific characteristics. In the example above, the knowledge item yields the longer RTs for both individuals. In this regard, item properties such as question length (i.e., the mere time it takes to read or hear the question) may contribute to explain overall differences between items. It takes just a little imagination to invert the perspective of Figure 2—with value points now representing the RTs of five individuals for two items (● =item 1, ○ = item 2). Here also, multivariate analysis should determine whether items 1 and 2 are exceptional cases or (more probably) representative of certain "configurations" of items properties such as length, uncertainty, position in the questionnaire, etc.

From these considerations, we draw two main conclusions. First, we need a multivariate model where RTs, our dependent variable, are *simultaneously* nested in individuals and in items. This is made possible by a particular kind of multilevel models known as a "cross-classified models", which we describe in full detail in Online Appendix A.4. Second, the predictions of this model will serve to establish a "baseline" of RTs tailored to every combination of individual and item characteristics. In turn, unexplained deviations from expected RT (i.e., residuals from the multivariate analysis) will lay the ground for a detection of potential cheating behavior. Of course, not all residuals should be considered as "suspicious cases". Rather, I will consider whether discontinuities in the relationship between RT residuals and item performance can be interpreted as evidence of potential cheating behavior. If test cheating is a sizeable phenomenon, as found in studies cited above, we should observe sharp discontinuities in the top deciles of the RT residuals distribution. In sum, my approach proceeds in three steps: (1) data preparation and model specification; (2) estimation of a cross-classified model; (3) residual analysis and identification of cheating behavior. Steps 1 and 2 are explained in Online Appendix A.4, while Step 3 is taken up in the next section.

5.2 Residual analysis and detection of cheating behavior

The results of the cross-classified model presented in Table A.2 may look unimpressive by some standards, such as the amount of variance explained (53% for CAWI and 27% for CATI). Yet they imply a large-scale "reclassification" of RTs assigned to respondents.[16] In fact, a substantial share of RTs has been adjusted to a baseline accounting for both individual and item characteristics. Referring to Figure 2, we have moved from Panel A to Panel B.

To avoid misunderstandings, let me repeat that the cross-classified models tested in Online Appendix A.4 aim to provide a refined measure of the accessibility of the pieces of knowledge targeted by knowledge questions. In themselves, these models do not purport to predict whether responses to

present case, the main advantage of using residuals (increasing statistical power) is wiped out by a major drawback, namely the impossibility of comparing baseline-corrected RTs across items since the distribution of residuals for all items have the same mean of 0 (Mayerl & Sellke, 2005, pp. 6–7). The cross-classified model presented below is an attempt to solve this problem.

[16]The mean percentile shift between the original RT distribution and the residuals' distribution is 7% for K1, 9% for K2, 17% for K3, and 16% for K4. In fact, the estimated shift exceeds a full quartile for more than 25% respondents on K3 and K4. These shifts are expressed in absolute values, summing adjustments in both directions (from lower actual RTs to higher expected RTs and vice-versa). I use percentiles because raw RTs and unstandardized residuals have the same metric but different ranges.

these questions are right or wrong, let alone whether cheating is a serious issue. To be sure, only a tiny fraction of the large amount of unexplained variance in RTs may be due to cheating. However, my argument is that cheating is not randomly or uniformly distributed over the range of RT residuals for the knowledge items. Rather, the occurrence of cheating should be evidenced by a sudden increase of correct answers among slower-than-expected respondents.

To explore this hypothesis, the third step of my analysis now relates residuals from the cross-classified models (estimated separately for CATI and CAWI respondents) to *hit rates* (share of correct answers) for knowledge items. The relationship is plotted in Figure 3, where lower values of unstandardized RT residuals (RTRs) are to be interpreted as faster-than-expected RTs and higher values as slower-than-expected RTs. As can be seen from the graph, two items, K2 and K3, might be considered "suspicious" because they exhibit an intriguing increase in item knowledge in the last deciles of RTRs values, at least for CAWI respondents. That is, on-line respondents who answered much more slowly than expected were particularly likely to provide a correct answer. Importantly, this increase represents a sharp discontinuity in the overall RTR–knowledge relationship; it contradicts the generally decreasing pattern of hit rates, which reaches its lowest level in the 8th to 9th decile range. Arguably, K2 and K3 are also the two items for which a quick internet search was most likely to be effective.

In sum, the discontinuity in hit rates already apparent on the basis of raw RTs (see Online Appendix A.1) has not been smoothed out by the residual analysis—if anything, it is all the more conspicuous. Importantly, it cannot be argued anymore that other factors (e.g., age or education; see above) account for this "bump" in knowledge. The suspicion of cheating among respondents in the last RTRs decile is, again, not proved, but it seems to stand on rather firm grounds.

A shown in Figure 3, the relationship between RTRs and knowledge follows a more conventional form for the two remaining items. The probability of a correct answer increases as a function of RTRs for item K1 (number of parties in government) and decreases for item K4 (party with most seats in the lower chamber). Most of the time, then, respondents who answered faster than expected got it wrong to K1 and right to K4 (although very quick answers to K4 tended to be less correct than slightly slower ones in the online mode). Conversely, giving the question more thought tended to improve responses to K1 quite significantly but, if anything, it worsened responses to K4.[17] More importantly, we do not observe a sudden increase in item knowledge in the last deciles of RTRs—thus there is no compelling evidence of cheating on these items.

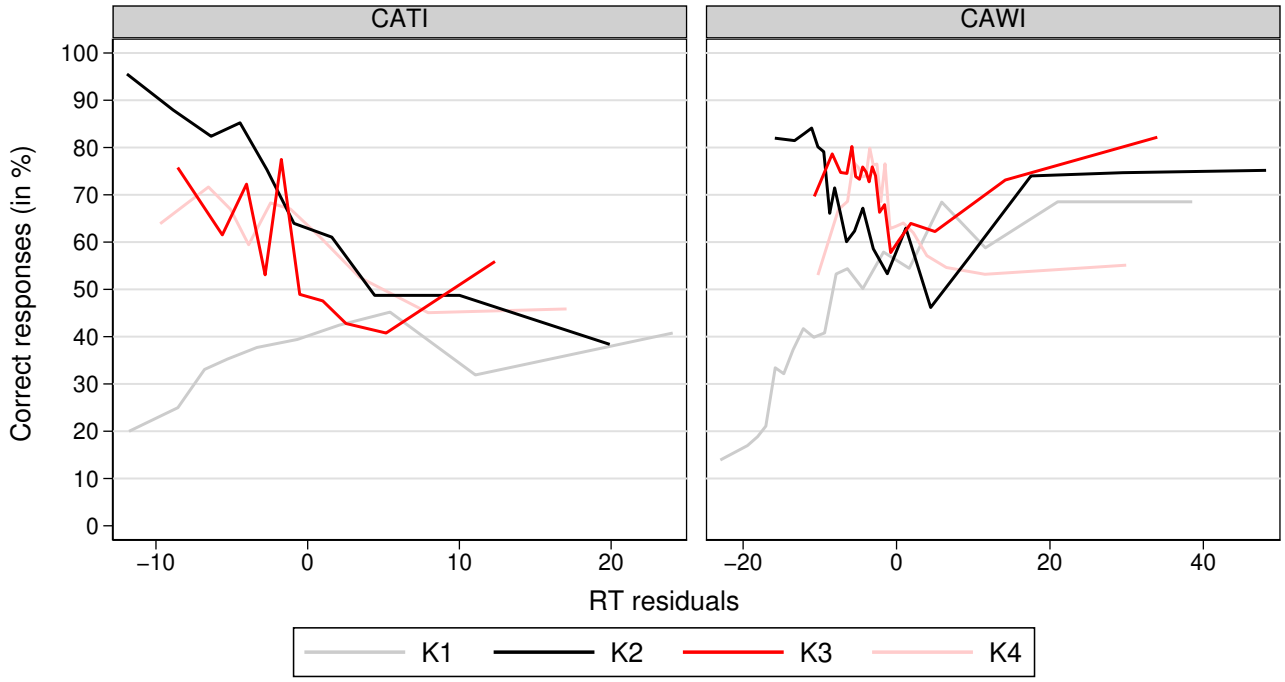
By and large, the residual analysis confirms the results obtained with simpler and more direct means, while allowing to rule out alternative explanations for the observed RT patterns (e.g., an effect of age). Controlling for these potential factors, our evidence suggests that the measurement of political knowledge is most probably plagued by "cheating", at least on the two items which provide the greatest incentive to look up answers on the internet. This interpretation is corroborated when considering how the relationship between RTs and hit rates is moderated by the interviewing mode (see Figure 3). As a matter of fact, online and telephone respondents are about as likely to provide correct answers. The only significant exception is found at longer response latencies (i.e., much longer than expected RTs) for items K2 and K3, where CAWI respondents increasingly outperform their telephone counterparts. This is at least consistent with the cheating hypothesis.[18]

In this regard, Burnett's (2016) claim that cheating is most prevalent on the "most difficult" questions may not apply to all situations. For one thing, my analysis suggests that the assessment of item difficulty can lead to different conclusions, depending on whether one emphasizes *uncertainty* (i.e., "perceived" difficulty) or "inherent" *difficulty*.[19] Thus, K1 (number of parties in government) was perceived as relatively "easy" judging from the small share of respondents volunteering a 'don't know' answer (8%, the lowest rate of all knowledge items). Yet, this item has by far the lowest hit rate (only 43% of correct responses) and longest RTs. Conversely, K4 (party with most parliamentary seats) was deemed difficult (17% of 'don't know' answers), yet it was second in terms of mean response speed and just a few percent below the highest hit rate attained by K2 and K3. On the other hand, K2 and K3 are perhaps less remarkable in terms of assigned difficulty; but they stand out in terms of how easy it was to find the correct answers from outside sources—"opportunity makes the thief".

[17] Interestingly, the relationship between RTRs and substantive response categories shows that the bulk of incorrect answers to K1 centers on the "four parties" answer. Indeed, for nearly 50 years until 2008, the government coalition had comprised four parties. Thus, when provided quickly, the "four parties" answer can be considered a "conditioned response" for many respondents to whom the question was somehow a "tricky question" (see Section 7).

[18] A different explanation may be that respondents knew the correct answers to these two "easy" questions all along, but that they wanted to "optimize" their answers by mulling over alternatives. To be sure, having response options on the screen (CAWI) may help respondents make an educated guess when they do not know the correct answer or are unsure about it. But this does not fully explain why the hit rate of CAWI respondents sharply grows in the last RTR deciles in the case of the "easy" items K2 and K3 (see Section 7).

[19] Item difficulty, which is usually assessed by the percentage of "correct" answers, only makes sense for knowledge items, while uncertainty extends to all items (for a similar distinction between difficulty and uncertainty, see Lowenstein, Richards, Leal, & et al., 2016; Radosevich, Partin, Nugent, & et al., 2004).



Note: Values on the horizontal axis are means for 10% intervals (CATI) or 5% intervals (CAWI)

Figure 3. Relationship between RT residuals from cross-classified models and percentage of correct answers to the four knowledge questions (N_{CATI} = 954; N_{CAWI} = 4301 (weighted data). The 5-percentile intervals displayed for CAWI respondents are inappropriate for CATI respondents, because they comprise less than 50 individuals; therefore, CATI respondent are regrouped by deciles (i.e., about 95 respondents). The values on the X-axis are means of RTRs for the respective intervals. For presentation reasons, the graph does not display the real data for the last interval of the K1 residuals distribution in the CAWI subsample (mean RTR=83.0; hit rate=64.1%). The maximum margin of error (for a hit rate of 50%, 95% C.I.) approximates ±7% for CAWI and ±10% for CATI.

6 Improving the measurement of political knowledge?

To sum up, I propose that cheating should be considered as a credible interpretation of the data when there is a substantial body of corroborating evidence. First, initial suspicion is raised when "laggards" (as determined by the residual analysis) have greater success in answering knowledge questions than most other respondents. However, as the example of item K1 shows, higher hit rates among laggards are not, per se, sufficient evidence for potential cheating, because they can be reflective of an overall trend toward better responding as response time (and thus thinking effort) increases. Arguably, a second condition is that higher hit rates among slower-than-expected respondents mark a discontinuity in the general pattern for a given item. Finally, when relevant, a further hint is a higher hit rate among laggards interviewed online, compared to their CATI counterparts. However, when cheating behavior can be suspected for CAWI respondents, it cannot be ruled out that slower-than-expected telephone respondents also look up answers on the internet—item K3 is a case in point, because there is a similar bump in correct answers in the last decile of RTRs for CATI (see Figure 3).

What should be done with "suspicious outliers"? In this regard, several solutions have been proposed, including removing or recoding individual cases with residuals smaller and/or larger than some value (e.g. van der Linden & Guo, 2008, p. 378; Munzert & Selb, 2017, p. 174). In this section, I propose that "outliers" be recoded rather than simply excluded. This proposal stems from an analysis of the empirical validity of the political knowledge measure provided in the Selects survey. Inspired by research on the convergent and/or predictive validity of political knowledge scales in the presence of bias and/or substantial cheating behavior among survey respondents (e.g. Clifford & Jerit, 2014, 2016; Delli Carpini & Keeter, 1993; Jensen & Thomsen, 2014; Montgomery & Cutler, 2013; B. Smith et al., 2020), I ask whether and how cheating compromises the validity of the knowledge scale as measured in the 2015 Selects survey. In terms of convergent validity, I test the hypothesis that cheating is an alternative means of knowledge acquisition that reduces

the impact of the usual predictors of knowledge. In terms of predictive validity, I examine whether correcting for cheating enables to improve the capacity of knowledge in predicting behavioral variables, such as political participation, which are known to depend on knowledge.

To facilitate comparisons across the forthcoming analyses of empirical validity, the four political knowledge items (0: incorrect answer; 1: correct answer) were averaged, resulting in a five-point scale ranging from 0 to 1 (M = 0.62; SD = 0.28). Cheating behavior was estimated on the basis of the corroborating criteria summed up above. Any respondent who fell in the highest two deciles of the residuals' distribution for K2 *or* in the highest decile for K3 was considered as a potential cheater. According to this rather inclusive measure, about 26% of respondents are considered as potential cheaters.[20]

First, I address the question of convergent validity. When the cheating variable is introduced in a prediction model of knowledge scores on the initial scale, it is shown to have both a direct effect and an indirect effect (see Table 2). First, individuals belonging to the cheating group are predicted to stand about 0.1 points higher on the 0–1 knowledge scale—compared to individuals not suspected of cheating. Second, and more importantly, cheating moderates the impact of several variables on political knowledge. To begin with, it depresses the positive impact of age ($p < 0.001$) and political interest ($p < 0.05$). Whereas knowledge scores are predicted to increase by 0.05 points between 18-year and 88-year old respondents, they are predicted to decrease by 0.08 points in the same age interval among respondents belonging to the potential cheater group. Likewise, political interest (moving from least to most interested respondents) contributes to a 0.32-point increase in knowledge scores, but this increase is reduced to a 0.25-point difference in the cheater group. Concerning gender, its interaction with potential cheating is also significant ($p < 0.03$): While the knowledge score of women is expected to be lower by 0.06 point compared to men, the corresponding difference is almost 0.1 point in the cheater group. In other words, cheating tends to widen the gender gap in political knowledge. Finally, both small-town dwellers ($p < 0.02$) and city dwellers ($p < 0.01$) are predicted to compare differently to countryside inhabitants in the non-cheating and cheating groups. While knowledge is usually highest among country dwellers and lowest among city dwellers, it is lowest for country dwellers and highest for small-town inhabitants in the cheating group.

In sum, these results suggest that being a potential cheater has a buffering effect on the usual predictors of knowledge acquisition such as older age, higher political interest, and living in a small community. Although the results reported here are less clear than those derived from experimental studies (e.g. B. Smith et al., 2020), they are nonetheless strikingly similar.

In terms of predictive validity, I tested two methods for recoding suspect cases, in order to build a corrected knowledge scale having the same (0–1) range as the initial scale (or Scale #0).[21]

- Scale #1: For suspected "cheaters", i.e., respondents in the highest two deciles of the residuals' distribution for K2 and in the highest decile for K3, correct responses are cut down to a half point. Accordingly, the overall scale is a nine-point scale with 0.125 intervals (M = 0.59; SD = 0.27).
- Scale #2: The same treatment as Scale #1 is applied, except that a full penalty (0 point) is given to suspected "cheaters"; therefore, the entire initial distribution is slightly shifted toward lower values (M = 0.56; SD = 0.28).

All scales are highly correlated with one another ($r \geq 0.92$). In order to assess the predictive validity of the initial and corrected knowledge scales, I observe whether they relate to other variables of interest in the same way as previous research has indicated. In the present case, past research has established that political knowledge has robust and positive relationships with various variables such as education, campaign activity, political interest, political engagement, and political participation (e.g. Barabas, Jerit, Pollock, & Rainey, 2014; Jacobs, Lomax Cook, & Delli Carpini, 2009; Jerit, Barabas, & Bolsen, 2006; Zukin, Keeter, Andolina, Jenkins, & Delli Carpini, 2006). Here, I focus on one well-established empirical regularity, namely the relationship between political knowledge and political participation. Because political knowledge is also correlated to some other recognized causes of participation (e.g., age, education), it makes sense to examine the effect of political knowledge within a more general predictive model of participation. This model allows me to pit the different knowledge scales against one another. I include in the model the same predictors as for the convergent analysis (i.e., age, sex, education, household income, linguistic region, dwelling area, and political interest), to which I add a measure of overall exposure to information (see Online Appendix A.3). The dependent variable, political participation, comes in two blends: direct democratic participation (the rough percentage of federal votes in which the respondent indicates she usually takes part) and electoral participation in the 2015 national election.

The results are displayed in Table 3. Because all scales have the same range (0–1), coefficients are broadly compa-

[20] For the sake of simplicity, only this measure of cheating will be retained in the upcoming analyses of convergent and predictive validity. However, I tested other measures varying along a dimension of inclusivity of the "cheating group". All analyses are presented in Online Appendix A.5, along with other robustness checks for the validity analyses conducted in this section.

[21] I have tested other, more complex, methods of item correction, which are presented in Online Appendix A.5. However, none of these additional scales of political knowledge exceeded the predictive validity of the simpler scales #1 and #2.

Table 2
Convergent validity analysis of the initial knowledge scale (OLS regression)

	Main effect			Interacted with cheating		
	Coef.	Std. Err.	Std. Coef.	Coef.	Std. Err.	Std. Coef.
Intercept	0.366***	0.019	-	-	-	-
Age (in decades)	0.007**	0.002	0.043	−0.019***	0.005	−0.152
Sex (woman)	−0.058***	0.008	−0.104	−0.036*	0.016	−0.046
Education	0.011***	0.002	0.080	0.004	0.004	0.032
Income: middle level[a]	0.038***	0.010	0.067	0.026	0.018	0.030
Income: high level[a]	0.076***	0.012	0.118	0.004	0.022	0.004
Region: French-speaking[b]	−0.096***	0.010	−0.142	−0.025	0.019	−0.021
Region: Italian-speaking[b]	−0.109***	0.020	−0.083	0.009	0.035	0.004
Dwelling place: small town[c]	−0.014	0.010	−0.025	0.048*	0.019	0.058
Dwelling place: big city[c]	−0.048***	0.011	−0.075	0.068**	0.022	0.062
Political interest	0.106***	0.006	0.296	−0.022*	0.011	−0.072
Cheating	0.106**	0.038	0.167	-	-	-
Adj. R^2	0.208					
F-test	67.2***					
N	5286					

[a] reference category=low level [b] ref. category=German-speaking [c] ref. category= countryside
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

rable across models. However, because models are identical except for a slightly different measure of political knowledge, one should not be surprised to find very close results from one model to the next. Starting with models explaining direct democratic participation, regression coefficients of political knowledge are only slightly higher for the two corrected scales, compared to the initial scale. In terms of information criteria (AIC, BIC), however, the results suggest a better goodness of fit and higher predictive accuracy for the models containing the corrected scales. In particular, differences between BIC values for Model 1 and for alternative models exceed 2 (ΔBIC M2 vs. M1=3.3; ΔBIC M3 vs. M1 = 2.6), suggesting positive evidence in favor of the corrected scales (Raftery, 1995); in turn, BIC values are indistinguishable for the two types of corrected knowledge measures. In contrast, evidence is more conclusive with respect to electoral participation. For one thing, regression coefficients are clearly higher for alternative Models 5 and 6, compared to Model 4 featuring the initial knowledge scale. Information criteria tell a similar story. Differences between BIC values (ΔBIC M5 vs. M4 = 5.9; ΔBIC M6 vs. M8 = 16.8) offer positive evidence in favor of Scale #1 and very strong evidence in favor of Scale #2. In addition, comparing BIC values for the two alternative models (ΔBIC M6 vs. M5 = 10.8) yields very strong evidence in favor of Scale #2.

Overall, it can be argued that Scale #2, which entails a full penalty for all cases of suspected cheating, is the best available means of improving the validity of the political knowledge measure in the 2015 Selects survey. It should be stressed that this is an entirely ad hoc recommendation, tailored to the specifics of this particular survey. Obviously, more research is needed to establish whether the methods proposed in this contribution can apply to other contexts and other election surveys. This issue is discussed further in the conclusion.

7 Conclusion

The main conclusion of my study is that a small but non-negligible share of respondents (typically between 10 and 20 percent, depending on items) can be suspected of "cheating" in their responses to some political knowledge questions, most probably by looking up the correct answers on the internet. Based on an analysis of response times (also dubbed as "latencies" or "reaction times"), I have proposed a method for detecting cheating behavior and for correcting knowledge scores which are inflated by cheating. Importantly, the method is not a ready-to-use algorithm, and its implementation requires judgment and adaptability. As a matter of fact, it was developed for the Swiss post-electoral survey 2015, with contains only four knowledge questions of medium difficulty. Therefore, this empirical background raises concerns of external validity—one may ask to what extent the method proposed here applies to other contexts and other surveys.

I argue that the response-time method is not tied to any particular type of question or questionnaire. Because baselines and residuals computed from RTs are both individual and item-specific, they can accommodate all types of items—

Table 3
Explaining political participation. Left panel: direct democratic participation (OLS regression coefficients); right panel: electoral participation (log odds from logistic regression)

	Direct democratic participation (0–10)						Electoral participation (0: no; 1: yes)					
	Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.	Coeff.	S.E.
Intercept	1.592***	0.180	1.589***	0.180	1.619***	0.179	−3.677***	0.195	−3.723***	0.196	−3.723***	0.195
Age	0.024***	0.002	0.024***	0.002	0.024***	0.002	0.025***	0.002	0.024***	0.002	0.024***	0.002
Sex (woman)	0.394***	0.071	0.400***	0.071	0.401***	0.071	0.151*	0.075	0.169*	0.075	0.178*	0.075
Education	0.098***	0.018	0.098***	0.018	0.100***	0.018	0.108***	0.019	0.109***	0.019	0.111***	0.019
Income: middle level[a]	0.192*	0.084	0.192*	0.083	0.195*	0.083	0.071	0.085	0.069	0.085	0.073	0.085
Income: high level[a]	0.238*	0.100	0.234*	0.100	0.236*	0.100	0.014	0.105	0.005	0.105	0.006	0.105
Region: French-speaking[b]	0.323***	0.086	0.327***	0.086	0.322***	0.086	−0.152	0.087	−0.139	0.088	−0.138	0.088
Region: Italian-speaking[b]	0.502**	0.168	0.517**	0.168	0.522**	0.168	0.485**	0.18	0.519**	0.18	0.539**	0.181
Dwelling place: small town[c]	−0.114	0.084	−0.111	0.084	−0.109	0.084	−0.288***	0.089	−0.285**	0.089	−0.280**	0.089
Dwelling place: big city[c]	−0.211*	0.098	−0.207*	0.098	−0.207*	0.098	−0.443***	0.103	−0.437***	0.103	−0.433***	0.103
Political interest	1.401***	0.057	1.396***	0.057	1.397***	0.057	0.938***	0.061	0.931***	0.061	0.931***	0.061
Overall exposure to information	0.534***	0.053	0.537***	0.053	0.544***	0.053	0.588***	0.055	0.590***	0.055	0.599***	0.055
Political knowledge (orig. scale)	0.836***	0.141	-	-	-	-	1.118***	0.143	-	-	-	-
Political knowledge (scale #1)	-	-	0.891***	0.143	-	-	-	-	1.263***	0.147	-	-
Political knowledge (scale #2)	-	-	-	-	0.854***	0.139	-	-	-	-	1.284***	0.143
N	4983		4983		4983		5230		5230		5230	
Adj. R^2/Pseudo R^2 (Nagelkerke)	0.301		0.302		0.302		0.344		0.347		0.348	
F-test / Chi-square	180.0***		180.4***		180.3***		1419.0***		1432.3***		1439.0***	
AIC	8862.8		8859.5		8860.2		4737		4731.1		4720.2	
BIC	8947.5		8944.2		8944.9		4822.3		4816.4		4805.6	

[a] reference category=low level [b] ref. category=German-speaking [c] ref. category=countryside.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

from most simple to most difficult, from shortest to longest, and so on. Arguably, the method is not even linked to a specific survey mode. Although cheating is probably prevalent among CAWI respondents, it may also occur in telephone interviews, as we have seen above. Finally, beyond cheating, RTs may also be useful to discover other problematic aspects of questions—an issue to which I return below.

In the following, I will zoom out from the particular case of the Selects survey and try to provide simple steps for applying the RT method to all kinds of surveys. The procedure presupposes that answers to knowledge questions have been collected with their respective RTs and that an initial additive scale of knowledge has been (or can be) constructed on the basis of separate items.

1. Trim the RT data to remove the undue influence of extreme outliers. The M3SD method presented in Section 4.3 is a rule of thumb that is probably suitable for most RT distributions. However, another threshold may be more chosen if, for example, RTs for a particular item comprise an unusually large number of extreme outliers.

2. Estimate a cross-classified model of trimmed RTs, using a fair number of individual-level and item-level predictors. The selection of predictors should be both theory-driven and data-driven, so as to maximize explained variance.

3. For each item, analyze the association between residuals from the cross-classified model (expressing the "true" accessibility of knowledge) and "hit rates" (shares of correct answers). Paying attention to hit rates of respondents who answered much more slowly than expected, determine whether there is a significant increase in hit rates among "laggards" and whether this increase is at odds with the general pattern for a given item.

4. For all items which satisfy the conditions required in the third step, compute a specific "window of cheating" corresponding to an interval of RT residuals where cheating is most likely. Next, compute a dummy variable indicating whether an individual belongs to one of the windows.[22] Then, proceed to an analysis of convergent validity where the dummy variable is interacted with plausible predictors of the initial knowledge scale. If cheating is indeed present, the interaction terms should show that the influence of usual predictors of political knowledge (e.g., political interest, age) is reduced for respondents belonging to the cheating group.

5. To sustain conclusions from the convergent validity analysis, run a predictive validity analysis where the initial scale is pitted against two corrected scales for explaining a widely recognized consequence of knowledge (e.g., political participation). In the first corrected scale, respondents in the window of cheating for a particular item receive a half point; in the second scale, a full penalty (= 0) is assigned. Compare the predictive accuracy of the three scales and chose the best one—it may be the initial scale after all…

All of these steps require sound judgment from researchers. As stressed above, there is no irrefutable evidence that someone has cheated in answering a survey question. However, we are certainly better off if we take the issue of cheating seriously and try to ascertain cheating behavior using the RT method. This could be done in complement to other methods to prevent cheating, such as self-reports of cheating, control for internet browsing, or time limits (see Section 2.2 above). In fact, because longer questionnaires probably decrease the impact or motivation to cheat, especially when combined with time constraints (Strabac & Aalberg, 2010, p. 180; B. Smith et al., 2020, p. 152), the RT method may be most useful when the number of items is low and opportunities for cheating are readily available, regardless of item difficulty.

As I see it, cheating in knowledge questions is a "crime of opportunity", and many respondents may not even be aware that they are breaking an implicit rule. Yet, very different patterns of results obtain when respondents are explicitly instructed that "it is alright (…) if you use the internet to double check your answer or look for the correct response if you do not already know it" (B. Smith et al., 2020, p. 143). For example, the gender gap in political knowledge (which I found to be widened by cheating) may be reduced when outside search is encouraged, most notably because women tend to put more effort in answering survey questions (B. Smith et al., 2020, p. 151). Accordingly, response times (and hit rates) may be inflated not only by cheating, but also by the desire of some respondents to optimize their answers by checking that their answers are correct using outside sources. Testing this hypothesis is beyond the scope of this article, but it should be given serious consideration in future research.[23]

A further observation is that the exploration of knowledge items through RT analysis is not only useful for detecting cheating behavior, but also for uncovering other types of "abnormal" responding. Thus, for example, the out-of-range pattern of answers to the question of the number of parties in the governing coalition (K1) feeds my suspicion that this question was a tricky one. Many respondents were misled because of their long-standing knowledge that four parties (and not five) were represented in the government, which

[22] For the (rather unlikely) case where a large number of items are deemed suspicious, it might be more convenient to compute an ordinal scale reflecting the total number of cheating windows to which a respondent belongs.

[23] Fortunately, recent computer-assisted systems allow to distinguish between "first-click", "last-click" and "page-submit" RTs as part of their paradata collection tools. Arguably, if there is a distinct "editing timepiece" (see Section 3.2) devoted to fact-checking rather than to first-hand cheating, then the first and last clicks should be synchronous (i.e., they are one and the same event) or closely sequential, and they should occur significantly earlier than the page submit click. Any other sequence (e.g., dissimilar first and last clicks with a late page submit) should be more indicative of cheating.

was actually the case for nearly 50 years in the context of the historical coalition agreement known as the "magic formula" (1959–2008), which only came to an end with the inclusion of the BDP (Bourgeois Democratic Party).[24] In sum, RT analysis has undeniable heuristic merits beyond and above the simple detection of cheating. In the internet era, even less than before, response times are not a simple kind of "paradata". They are a valuable (and perhaps indispensable) resource for evaluating the quality of survey responses, especially when "shifting referents" (Page & Shapiro, 1992, pp. 58–59) result in the same questions being interpreted differently (i.e., being related to different attitudes and beliefs) at different points of time.

In concluding, I may repeat that the analysis presented here is of a "technical" nature and leaves untouched a couple of important questions—the *why* questions. To begin with, it has nothing to say about why cheating occurs in the first place. In a nutshell, social desirability has often been offered as an explanation (e.g. Clifford & Jerit, 2016; Munzert & Selb, 2017; Shulman & Boster, 2014). In particular, social expectations within groups may elicit a desire "to cultivate a desirable reputation" as a politically sophisticated citizen, which may encourage cheating even in the absence of effective control by peers (Marshall, 2019). Relatedly, since "knowledge is power", as the aphorism goes, it is probably not by chance that cheating in the completion of various tasks is enhanced in competitive settings (Rigdon & D'Esterre, 2015; Schwieren & Weichselbaumer, 2010). At least for those who have internalized the pressure for achievement, being unknowledgeable about politics is to appear powerless and outperformed by others. This interpretation of cheating as the result of self-presentation biases and competitive pressures is buttressed by evidence that political knowledge is not a distinct construct, but instead "resides on the same dimension as knowledge of other subjects" (Burnett & McCubbins, 2020, p. 194).

A second "why" question has to do with the normative and political consequences of cheating. Why should we be concerned with cheating in political knowledge tests? In this regard, I may reiterate that nearly all political knowledge tests available today purport to measure declarative (explicit) knowledge rather than procedural (implicit) knowledge. In some respects, however, cheating can be considered a form of procedural knowledge that has intrinsic value—knowing how and where to find the relevant information for answering questions (P. R. Johnson, 2009; Prior & Lupia, 2008; B. Smith et al., 2020). Hence, "(t)raditional surveys, while having many virtues, prevent or inhibit exactly the kinds of search activities that are in fact strongly encouraged by people who want others to make informed decisions" (Prior & Lupia, 2008, p. 180). This viewpoint suggests a need to refocus our attention away from encyclopedic and abstract knowledge toward more practical and useful forms of knowledge. In other words, we should place more emphasis on "necessary knowledge" for making politically competent choices (Lupia, 2006) and on "operative knowledge" gathered from practical experience which enables individuals to make sense of politics and to take political action (Cramer & Toff, 2017; P. R. Johnson, 2009).

From there it is but a short step to claiming that "cheating behavior" (as defined by usual declarative/academic criteria) is actually reflective of the use of a "politically relevant skill". Put another way, what is the point in naming "cheating" a behavior performed in the context of a political knowledge test, while a similar practice is widespread (and valued) among the most politically sophisticated individuals in other settings—for example wiki-wandering to prepare a political science lecture? On this point, I agree with other scholars who argue that both declarative and procedural aspects of political knowledge are important, but that they should not be conflated because they are driven by different motivations and mechanisms (e.g. Delli Carpini, 2009, pp. 39–40; B. Smith et al., 2020, p. 151). Insofar as most existing knowledge tests do not allow to distinguish declarative and procedural knowledge, cheating entails a blurring of these two knowledge types. In this regard, RT analysis is a valuable tool to measure political knowledge in its pure declarative dimension.

8 Acknowledgement

I would like to thank Simon Lanz and four anonymous reviewers for their valuable comments on previous versions of this article

References

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

Barabas, J., Jerit, J., Pollock, W., & Rainey, C. (2014). The question(s) of political knowledge. *American Political Science Review*, *108*(4), 840–855.

Bassili, J. N. (1995). On the psychological reality of party identification: Evidence from the accessibility of voting intentions and of partisan feelings. *Political Behavior*, *17*(4), 339–358.

Bassili, J. N. (2000). Editor's introduction: Reflections on response latency measurement in telephone surveys. *Political Psychology*, *21*(1), 1–6.

[24]The fact that Federal Councillors Eveline Widmer-Schlumpf and Samuel Schmid were previous members of the SVP (until their eviction from that party and the creation of the BDP in 2008) probably added to the confusion. Accordingly, quick respondents (in the first two or three deciles of RT residuals) were particularly prone to provide incorrect answers. In that process, respondents with high political interest were no more knowledgeable than their less interested counterparts.

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly*, *60*, 390–399.

Bizer, G. Y., & Krosnick, J. A. (2001). Exploring the structure of strength-related attitude features: The relation between attitude importance and attitude accessibility. *Journal of Personality and Social Psychology*, *81*(4), 566–586.

Bullock, J. G., Gerber, A. S., Hill, S. J., & Huber, G. A. (2015). Partisan bias in factual beliefs about politics. *Quarterly Journal of Political Science*, *10*, 519–578.

Burnett, C. M. (2016). Exploring the difference in participants' factual knowledge between online and in-person survey modes. *Research and Politics*, *3*(2), 1–7.

Burnett, C. M., & McCubbins, M. D. (2020). Is political knowledge unique? *Political Science Research and Methods*, *8*, 188–195.

Cizek, G., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests.* New York: Routledge.

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*(2), 120–131.

Clifford, S., & Jerit, J. (2016). Cheating on political knowledge questions in online surveys: An assessment of the problem and solutions. *Public Opinion Quarterly*, *80*(4), 858–887.

Converse, P. E. (1964). The nature of belief systems in mass publics. In D. Apter (Ed.), *Ideology and discontent* (pp. 206–261). New York: Free Press.

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society*, *176*(1), 271–286.

Cramer, K. J., & Toff, B. (2017). The fact of experience: Rethinking political knowledge and civic competence. *Perspectives on Politics*, *15*(3), 754–770.

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. J. (2016). Toward a formalized account of attitudes: The causal attitude network (CAN) model. *Psychological Review*, *123*(1), 2–22.

Davis, S. F., Drinan, P. F., & Bertram Gallant, T. (2009). *Cheating in school: What we know and what we can do.* Chichester: Wiley-Blackwell.

Delli Carpini, M. X. (2009). The psychology of civic learning. In E. Borgida, C. M. Federico, & J. L. Sullivan (Eds.), *The political psychology of democratic citizenship* (pp. 23–51). New York: Oxford University Press.

Delli Carpini, M. X., & Keeter, S. (1993). Measuring political knowledge: Putting first things first. *American Journal of Political Science*, *37*(4), 1179–1206.

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

DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.

Diedenhofen, B., & Musch, J. (2017). Pagefocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods*, *49*, 1444–1459.

Erber, M. W., Hodges, S. D., & Wilson, T. D. (1995). Attitude strength, attitude stability, and the effects of analyzing reasons. In R. Petty & J. Krosnick (Eds.), *Attitude strength: Antecedents and consequences.* (pp. 433–454). Mahwah (NJ): Lawrence Erlbaum.

Evans, J. R., & Mathur, A. (2018). The value of online surveys: A look back and a look ahead. *Internet Research*, *28*(4), 854–887.

Fabrigar, L. R., MacDonald, T. K., & Wegner, D. T. (2005). The structure of attitudes. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 79–124). Mahwah (NJ): Lawrence Erlbaum.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. Clark (Eds.), *Research methods in personality and social psychology* (pp. 74–97). Newbury Park: Sage.

Fazio, R. H. (2001). On the automatic activation of associated evaluations: An overview. *Cognition and Emotion*, *15*(2), 115–141.

Fazio, R. H., & Williams, C. J. (1986). Attitude accessibility as a moderator of the attitude-perception and attitude-behavior relations: An investigation of the 1984 presidential election. *Journal of Personality and Social Psychology*, *51*(3), 505–514.

Fox, J.-P., Klein Entink, R., & van der Linden, W. (2007). Modeling of responses and response times with the package cirt. *Journal of Statistical Software*, *20*(7), 1–14.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of web and telephone surveys. *Public Opinion Quarterly*, *69*(3), 370–392.

Glasman, L. R., & Albarracín, D. (2006). Forming attitudes that predict future behavior: A meta-analysis of the attitude—behavior relation. *Psychological Bulletin*, *132*(5), 778–822.

Gooch, A. (2015). Measurements of cognitive skill by survey mode: Marginal differences and scaling similarities. *Research and Politics*, *2*(3), 1–11.

Harms, C., Jackel, L., & Montag, C. (2017). Reliability and completion speed in online questionnaires under consideration of personality. *Personality and Individual Differences*, *111*, 281–290.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review*, *21*(3), 360–373.

Holbrook, A. L., & Krosnick, J. A. (2010). Operative and meta-attitudinal manifestations of attitude accessibility: Two different constructs, not two measures of the same construct. In J. P. Forgas, J. Cooper, & W. D. Crano (Eds.), *The psychology of attitudes and attitude change* (pp. 109–124). Florence (KY): Psychology Press.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*(2), 161–172.

Huckfeldt, R., Levine, J., Morgan, W., & Sprague, J. (1999). Accessibility and the political utility of partisan and ideological orientations. *American Journal of Political Science*, *43*(3), 888–911.

Jacobs, L. R., Lomax Cook, F., & Delli Carpini, M. X. (2009). *Talking together. public deliberation and political participation in America.* Chicago: University of Chicago Press.

Jensen, C., & Thomsen, J. P. F. (2014). Self-reported cheating in web surveys on political knowledge. *Quality & Quantity*, *48*(6), 3343–3354.

Jerit, J., Barabas, J., & Bolsen, T. (2006). Citizens, knowledge, and the information environment. *American Journal of Political Science*, *50*(2), 266–282.

Johnson, M. (2004). Timepieces: Components of survey question response latencies. *Political Psychology*, *25*(5), 679–702.

Johnson, P. R. (2009). What knowledge is of most worth? In E. Borgida, C. M. Federico, & J. L. Sullivan (Eds.), *The political psychology of democratic citizenship* (pp. 52–70). New York: Oxford University Press.

Johnson, T. P., Basic, M., & Joscelyn, S. (2016). Diffusion of web survey methodology: An update. *Survey Research (Newsletter from the Survey Research Laboratory)*, *47*(3), 1–2.

Kinsey, S., Iannacchione, V., Shook-Sa, B., Peytcheva, E., & Triplett, S. (2013). Examination of data collection methods for the national crime victimization survey: Final report. RTI International, Project n. 0211889.001. Retrieved from https://www.bjs.gov/content/pub/pdf/Examination_Data_Collection.pdf.

Krosnick, J. A. (1989). Attitude importance and attitude accessibility. *Personality and Social Psychology Bulletin*, *15*(3), 297–308.

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

Lord, C. G., & Lepper, M. R. (1999). Attitude representation theory. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 31, pp. 265–343). San Diego (CA): Academic Press.

Lowenstein, L. M., Richards, V. F., Leal, V. B., & et al. (2016). A brief measure of smokers' knowledge of lung cancer screening with low-dose computed tomography. *Preventive Medicine Reports*, *4*, 351–356.

Lupia, A. (2006). How elitism undermines the study of voter competence. *Critical Review*, *18*(1–3), 217–232.

Luskin, R. C. (1987). Measuring political sophistication. *American Journal of Political Science*, *31*(4), 856–899.

Lutz, G. (2016). *Elections fédérales 2015. Participation et choix électoral.* Lausanne: Selects/FORS.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426–451.

Marquis, L. (2014). The psychology of quick and slow answers: Issue importance in the 2011 Swiss parliamentary elections. *Swiss Political Science Review*, *20*(4), 697–726.

Marshall, J. (2019). Signaling sophistication: How social expectations can increase political information acquisition. *Journal of Politics*, *81*(1), 167–186.

Mayerl, J., & Sellke, P. (2005). *Analyzing cognitive processes in CATI-surveys with response latencies: An empirical evaluation of the consequences using different baseline speed measures.* Working paper, University of Stuttgart.

Mayerl, J., & Urban, D. (2008). *Antwortreaktionszeiten in Survey-Analysen. Messung, Auswertung und Anwendungen.* Wiesbaden: VS Verlag.

Meyer, M., & Schoen, H. (2014). Response latencies and attitude-behavior consistency in a direct democratic setting: Evidence from a subnational referendum in Germany. *Political Psychology*, *35*(3), 431–440.

Montgomery, J. M., & Cutler, J. (2013). Computerized adaptive testing for public opinion surveys. *Political Analysis*, *21*, 172–192.

Motta, M. P., Callaghan, T. H., & Smith, B. (2017). Looking for answers: Identifying search behavior and improving knowledge-based data quality in online surveys. *International Journal of Public Opinion Research*, *29*(4), 575–603.

Mulligan, K., Grant, J. T., Mockabee, S. T., & Monson, J. Q. (2003). Response latency methodology for survey research: Measurement and modeling strategies. *Political Analysis*, *11*(3), 289–301.

Munzert, S., & Selb, P. (2017). Measuring political knowledge in web-based surveys: An experimental valida-

tion of visual versus verbal instruments. *Social Science Computer Review*, *35*(2), 167–183.

Page, B. I., & Shapiro, R. Y. (1992). *The rational public. Fifty years of trends in Americans' policy preferences.* Chicago.

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections.* New York: Cambridge University Press.

Prior, M. (2014). Visual political knowledge: A different road to competence? *Journal of Politics*, *76*(1), 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*(1), 169–183.

Prior, M., Sood, G., & Khanna, K. (2015). You cannot be serious: The impact of accuracy incentives on partisan bias in reports of economic perceptions. *Quarterly Journal of Political Science*, *10*, 489–518.

Prislin, R. (1996). Attitude stability and attitude strength: One is enough to make it stable. *European Journal of Social Psychology*, *26*, 447–477.

Radosevich, D. M., Partin, M. R., Nugent, S., & et al. (2004). Measuring patient knowledge of the risks and benefits of prostate cancer screening. *Patient Education and Counseling*, *54*, 143–152.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.

Rapeli, L. (2014). *The conception of citizen knowledge in democratic theory.* Basingstoke: Palgrave Macmillan.

Reichert, F. (2010). Political competences and political participation: On the role of "objective" political knowledge, political reasoning, and subjective political competence in early adulthood. *Journal of Social Science Education*, *9*(4), 63–81.

Rigdon, M. L., & D'Esterre, A. P. (2015). The effects of competition on the nature of cheating behavior. *Southern Economic Journal*, *81*(4), 1012–1024.

Sass, S., Wittwer, J., Senkbeil, M., & Köller, O. (2012). Pictures in test items: Effects on response time and response correctness. *Applied Cognitive Psychology*, *26*, 70–81.

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, *25*(5), 638–656.

Schwieren, C., & Weichselbaumer, D. (2010). Does competition enhance performance or cheating? A laboratory experiment. *Journal of Economic Psychology*, *31*, 241–253.

Selects. (2015). Swiss electoral studies (selects) 2015. Distributed by FORS, Lausanne, 2016. doi:10.23662/FORS-DS-726-5

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.

Smith, B., Clifford, S., & Jerit, J. (2020). How internet search undermines the validity of political knowledge measures. *Political Research Quarterly*, *73*(1), 141–155.

Smith, E. R., & Conrey, F. R. (2007). Mental representations are states, not things. implications for implicit and explicit measurement. In B. Wittenbrink & N. Schwarz (Eds.), *Implicit measures of attitudes* (pp. 247–264). New York: Guilford Press.

Stern, M. J. (2008). The use of client-side paradata in analyzing the effects of visual layout on changing responses in web surveys. *Field Methods*, *20*(4), 377–398.

Strabac, Z., & Aalberg, T. (2010). Measuring political knowledge in telephone and web surveys: A cross-national comparison. *Social Science Computer Review*, *29*(2), 175–192.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge: Cambridge University Press.

Turner, G., Sturgis, P., & Martin, D. (2015). Can response latencies be used to detect survey satisficing on cognitively demanding questions? *Journal of Survey Statistics and Methodology*, *3*, 89–108.

van der Linden, W. J. (2011). Modeling response times with latent variables: Principles and applications. *Psychological Test and Assessment Modeling*, *53*(3), 334–358.

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384.

van der Linden, W. J., & Pashley, P. J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3–30). New York: Springer.

Van Harreveld, F., & van der Pligt, J. (2004). Attitudes as stable and transparent constructions. *Journal of Experimental Social Psychology*, *40*, 666–674.

Van Harreveld, F., van der Pligt, J., de Vries, N. K., Wenneker, C., & Verhue, D. (2004). Ambivalence and information integration in attitudinal judgment. *British Journal of Social Psychology*, *43*, 431–447.

Vannette, D. L., & Krosnick, J. A. (2014). Answering questions: A comparison of survey satisficing and mindlessness. In A. Ie, C. T. Ngnoumen, & E. J. Langer (Eds.), *The Wiley Blackwell handbook of mindfulness* (pp. 312–327). West Sussex, UK: John Wiley & Sons.

Wang, T., & Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, *29*(5), 323–339.

Watson, V., Porteous, T., Bolt, T., & Ryan, M. (2019). Mode and frame matter: Assessing the impact of survey

mode and sample frame in choice experiments. *Medical Decision Making*, *39*(7), 827–841.

Wegener, D. T., Downing, J., Krosnick, J. A., & Petty, R. E. (1995). Measures and manipulations of strength-related properties of attitudes: Current practice and future directions. In R. E. Petty & J. A. Krosnick (Eds.), *Attitude strength: Antecedents and consequences* (pp. 455–487). Hillsdale (NJ): Lawrence Erlbaum.

Wilson, T. D., & Hodges, S. D. (1992). Attitudes as temporary constructions. In L. Martin & A. Tesser (Eds.), *The construction of social judgments* (pp. 37–65). Hillsdale (NJ): Lawrence Erlbaum.

Wingfield, A. (1998). Comprehending spoken questions: Effects of cognitive and sensory change in adult aging. In N. Schwarz, D. C. Park, B. Knäuper, & S. Sudman (Eds.), *Cognition, aging, and self-reports* (pp. 201–228). Philadelphia: Psychology Press.

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38.

Wyer, J., R. S. (2008). The role of knowledge accessibility in cognition and behavior: Implications for consumer information processing. In C. P. Haugtvedt, P. M. Herr, & F. R. Kardes (Eds.), *Handbook of consumer psychology* (pp. 31–76). New York: Lawrence Erlbaum.

Yan, T., & Olson, K. (2013). Analyzing paradata to investigate measurement error. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 73–95). Hoboken (NJ): John Wiley & Sons.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, *22*, 51–68.

Yu, H., Glanzer, P. L., & Johnson, B. (2017). Why students cheat: A conceptual framework of personal, contextual and situational factors. In D. M. Velliaris (Ed.), *Handbook of research on academic misconduct in higher education* (pp. 35–59). Hershey (PA): IGI Global.

Zaller, J. R., & Feldman, S. (1992). A simple theory of the survey response: Answering questions versus revealing preferences. *American Journal of Political Science*, *36*(3), 579–616.

Zukin, C., Keeter, S., Andolina, M., Jenkins, K., & Delli Carpini, M. X. (2006). *A new engagement? Political participation, civic life, and the changing american citizen.* New York: Oxford University Press.