Lionel Marquis, University of Lausanne

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

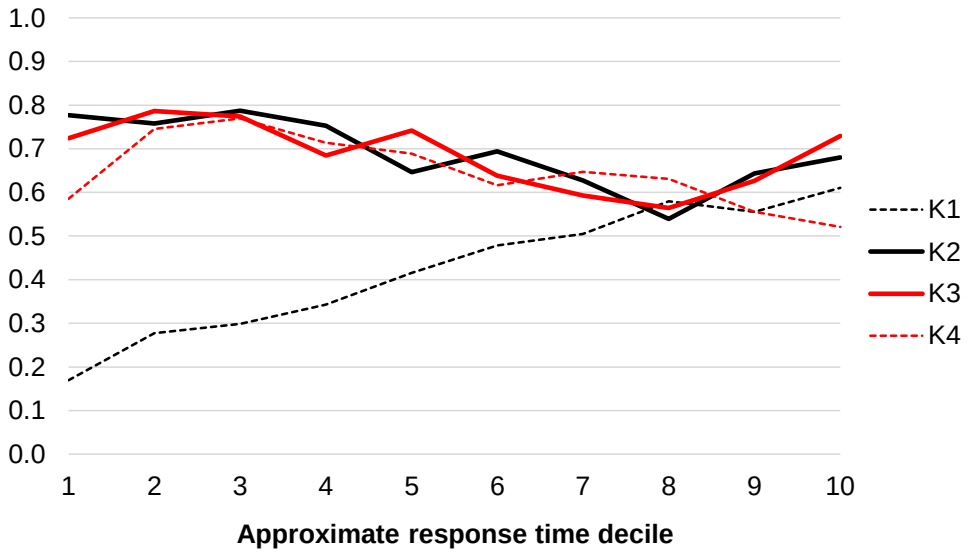
Survey Research Methods (2021) Vol. 15, No. 4

Online Appendix

Appendix A.1: First-hand analysis of cheating behavior

Figure A.1 presents the results of a simple crosstabulation between the share of correct answers and corrected RTs, broken down in deciles. Relationships are mostly monotonic (though not linear), as the share of correct answers either tends to decrease (K2 to K4) or to increase (K1) as a function of RTs. In the case of K2 and K3, however, responses in the last two or three RTs deciles break with the overall form of the relationship.

Figure A.1: Relationship between the share of correct answers on the four knowledge items and RTs expressed in approximate decile intervals



A simple crosstabulation such as the one performed in Figure A.1 should be the first step in the exploration of the validity of a knowledge scale. Whenever the share of correct answers tends to increase at longer RTs *and* contradicts the overall trend, then there is cause for concern. The method described in this article will establish whether this concern is founded and whether available solutions to correct for the problem of cheating in knowledge scales are fit for purpose.

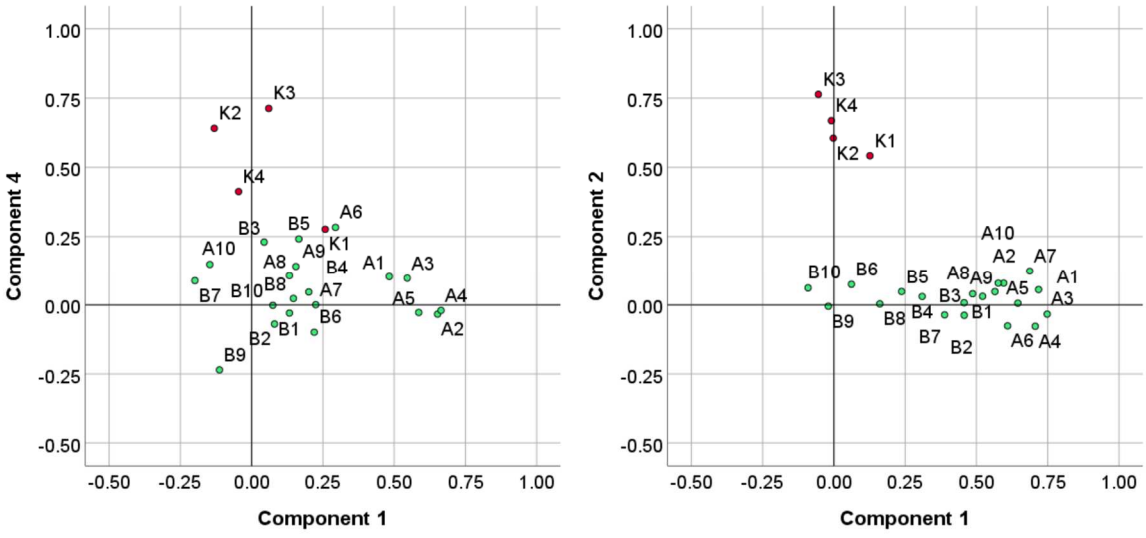
Appendix A.2: Component loadings for RTs

The results of a principal component analysis of RTs to all 24 items suggest that the time taken to answer any one question has much to do with the time taken to answer the others (model KMO=.97). Components were rotated (oblimin method) to ease the interpretation of extracted dimensions. The first dimension alone explains 36% of the total variance in RTs. Most filler items (i.e., 17 out the 20 attitudinal and behavioral questions) are strongly correlated to this dimension ($.50 \leq r \leq .87$). The remaining three filler items (B4, B8, and B10) are related to a second dimension (6% explained variance; $.60 \leq r \leq .68$); interestingly, these items share the feature of being retrospective questions, which might tap a specific "biographic memory" dimension. Conversely, knowledge RTs are unrelated to the first two components (all *r*s < .35) and contribute almost entirely to the extraction of a third component, accounting for an additional 5% in total variance ($.57 \leq r \leq .69$) while RTs to other questions are essentially unrelated to it (all *r*s < .20).

One may conclude that RTs are generally interrelated, but also that answers to knowledge questions tend to proceed at their own pace, independently from answers to other types of questions. Upon closer inspection, however, it appears that this description is reflective of the pattern of RTs in the online (CAWI) sample — among CATI respondents, in contrast, RTs for knowledge items are less different from RTs for other items. Figure A.2 displays the associations between RTs and the first component of the factorial solution as well as the most important component *for knowledge items in particular*. As it turns out, this particular component corresponds to the fourth component in the CATI sample (5% explained variance), and to the second component in the CAWI sample (6% explained variance). As for the first component, it explains 19% variance in the CATI sample (model KMO=.84) and 32% variance in the CAWI sample (model KMO=.96). The figure is based on standardized regression coefficients from a model where each item's RT is regressed on both components ("pattern matrix"); it thus displays components' *unique* contribution to the prediction of RTs.

In sum, there is a clear distinction between knowledge items and other types of items. However, this distinction is enhanced by the online interviewing mode. All detailed analyses can be obtained from the author upon request.

Figure A.2: Component loadings for the rotated principal component analysis of RTs among CATI respondents (left panel) and CAWI respondents (right panel)



Appendix A.3: Measurement of variables

Individual-level and contextual variables

Table A.1 describes all individual and contextual variables used in the paper and in the following appendices. Item-level variables and raw RTs are discussed below.

Table A.1: Description of individual-level and contextual variables

Variable	*Range*	*Descriptive statistics*	*% valid*
Age	18 – 96	M=49.8; SD=17.8	100.0
Sex	0 – 1	Man: 49.2%; Woman: 50.8%	99.2
Education	1 (compulsory) – 7 (university)	Median=4	100.0
Linguistic region	1 – 3	German-speaking: 73.3%; French-speaking: 22.0%; Italian-speaking: 4.7%	100.0
Political interest	1 (not interested at all) – 4 (very interested)	Median=3	99.4
Intensity of left-right self-placement	0 – 5 (folded 0–10 left-right scale; DK and NA coded 0)	M=1.91; Median=2; SD=1.57	100.0
Overall exposure to information	0 (no exposure) – 3 (very high exposure)	Additive scale (α=.66; M=1.60; SD=.78) from 3 items: (1) attention to news; (2) # of information sources; (3) interpersonal discussion	99.7
Nationality at birth	0 – 1	Swiss: 85.2%; foreign: 14.8%	99.2
Professional role/function	1 – 6	Executive: 7.1%; Supervisory: 14.9%; Operative: 32.9%; Self-employed: 7.8%; Non-active: 31.9%; Missing information: 5.4%	100.0
Income (gross monthly household income, recoded)	1 (low) – 3 (high)	Low (up to 5000 CHF/missing): 31.8%; Middle (5000–10,000 CHF): 43.0%; High (more than 10,000 CHF): 25.2%	100.0
Dwelling place	1 (densely populated area) – 3 (thinly populated area)	Densely populated (city): 25.1%; Intermediate density (towns/suburbs): 48.6%; Thinly populated (rural areas): 26.2%	100.0
Interviewing mode	0 (CATI) – 1 (CAWI)	CATI: 18.1%; CAWI: 81.9%	100.0
Days elapsed since election	1 – 42	M=11.79; Median=10; SD=9.78	100.0
Internet use	0 (occasional/no use) – 2 (daily use)	Occasional/no use: 12.6%; Regular use: 11.6%; Daily use: 75.9%	99.3

Item-level variables

Item type varies between attitudinal (10 items), behavioral (10 items) and knowledge (4 items); see Table 1 for description and classification of items. *Item uncertainty* is based on the total percentage of "don't know" responses and refusals to answer; it ranges between 0.3 (items B6, B7, B9 and B10) and 17.1 (item K4).[1] *Item position in the questionnaire* indicates the item's rank in the overall order of questions in the questionnaire; it ranges between 4 (item A8) and 112 (item B4). *Item categories: Scale vs. non-scale* specifies whether the question required respondents to provide their answer on a predetermined (ordinal) scale or not; given this definition, one half of the questions qualify as "scales". *Item categories: Number*

[1] For some items, these values differ slightly from the percentage of missing values provided in Table 1 (column 4). This is because Table 1 reports *all* non-valid answers, whereas the item uncertainty measure does *not* include "system missing" cases (e.g., a question was not posed to some respondents).

indicates the number of answer options provided to respondents; it ranges between 2 and 20[2]. Finally, *item length* corresponds to the number of words comprised in a question; it can vary between the CATI and CAWI versions of a question, and it ranges between 7 (B6, B10) and 44 (A1, CATI).

Descriptive analysis of raw RTs

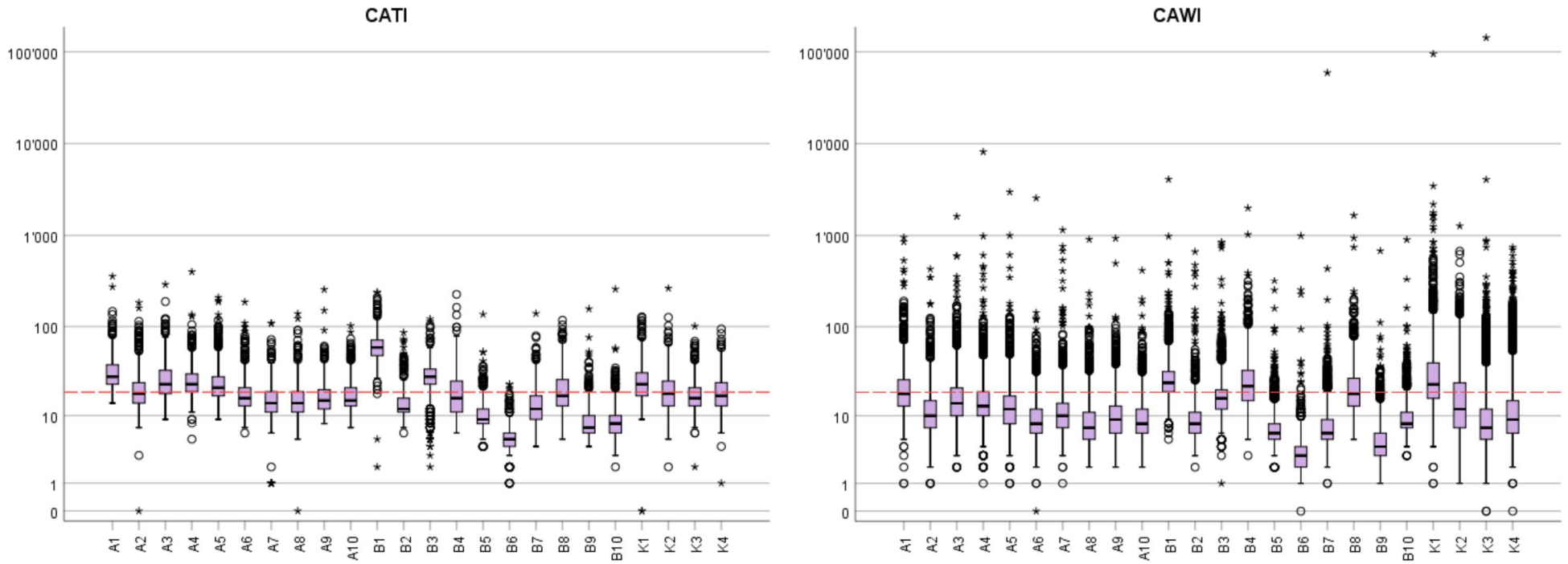
Figure A.3 below confirms the presence of strong outliers in the distribution of RTs and shows further that outliers tend to be concentrated among CAWI participants. In fact, CAWI respondents typically take less time to answer questions than CATI respondents (as indicated by lower median values), but they also comprise a disproportionate share of laggards. Among CAWI respondents, 4 percent took at least one 5-minute or longer pause, and 20 of them took more than 5 hours to complete the questionnaire.[3] Likewise, as one can see from Figure A3, even the use of *logged response times* (or "logtimes") does not prevent the occurrence of extremely deviant observations.

As Figure A.3 also suggests, there is actually a second type of "outliers" in raw RTs, which might be called "false starters" (Marquis 2014). False starters are respondents who provide an amazingly quick answer — most probably before the question has been completely read or heard (see Faas & Mayerl 2010; Meyer & Schoen 2014). A lot of answers of this type are presumably produced by "accident" or negligence, as when a CATI interviewer knows a question by heart and presses the 'Next' button before the entire question has been read, or when a CAWI respondent inadvertently selects an answer option before reading the question and, for independent reasons, sticks to his first choice (Stern 2008: 5). For example, based on the 125,686 answers given by all respondents to all 24 items, 1.3% of questions were answered in zero to two seconds after the keystroke, i.e., after the interviewer began to read the question (CATI) or after the question appeared on the screen (CAWI). However, the model presented in Appendix A.5 does a fair job at predicting very short latencies (most notably through its item-level variables such as item length), and thus we do not need a trimming procedure for false starters.

[2] In the case of an open-ended question with a great number of pre-established categories (B8: Year since respondent lives in current canton), the number of categories is fixed at 15. For B1, which is comprised of five items, the total number of categories is considered (i.e., 20).

[3] This descriptive account is based only on the 20 filler items and 4 knowledge items. The number of outliers would be higher if the analysis were based on the whole questionnaire.

Figure A.3: Logged RTs for filler items (A1–B10) and knowledge items (K1–K4), according to the interviewing mode (CATI vs. CAWI)



Notes: See Table 1 for a legend of the item symbols. The dotted red line is the grand mean of all items (regardless of interviewing mode). Circles and asterisks denote outliers, in this case values that are higher than the 75th percentile or lower than the 25th percentile by more than 1.5 times the interquartile range (i.e., the height of the box) for circles, by more than 3 times the interquartile range for asterisks. Unlike Table 1, only RTs which are valid for *all* items are represented; this explains some differences between table and graph (e.g., extreme outliers for A1 are not displayed).

Appendix A.4: Cross-classified model for the prediction of RTs

Data preparation and model specification

A *cross-classified model* is a variant of a multilevel model where observations of a dependent variable (at level 1) can be *simultaneously* classified in several higher-level units relevant to independent variables. Unlike most multilevel models, however, these higher-level units are not related to one another in a hierarchical way. In the present case, we indeed have a data structure where RTs are nested *both* within individuals and within items (see Yan & Tourangeau 2008). Hence, the model is able to provide an estimated RT *for each individual on each item*, taking into account all data provided at the individual and item levels.[4] In turn, residuals from this model can be considered to assess the *accessibility* of the underlying attitudes or knowledge units, since these residuals are RTs "purged" from the individual, item-specific, and contextual *tendencies* to provide short or long answers.

First, the data is restructured (or "stacked") so as to contain as many "observations" as there are questions (and thus RTs) which have been answered by all interviewees. This produces a multilevel structure where Level 1 is the **RTs level**, which consists of RTs to all relevant survey questions answered by all respondents.[5] Importantly, the level-1 (RT) observations are nested in *two* types of higher-level variables. On the one hand, the **item-level characteristics** are made up of question features which are susceptible to increase or decrease RTs. Following, in part, suggestions by Yan and Tourangeau (2008), I selected the following characteristics: question length (number of words); question position in the questionnaire; item uncertainty (percentage of DK/NA answers); number of answer categories; type of answer categories (scale vs. non-scale); and type of subject (attitudinal vs. factual/behavioral vs. knowledge). On the other hand, RTs are also nested in individuals, because each respondent answers several questions. Hence, **individual-level characteristics** represent another level-2 unit related to a different set of variables. These consist of personal attributes of the respondents which are supposed to make a difference in the speed with which responses are provided. In line with similar analyses by other authors and suggestions by reviewers, I selected the following variables: age, gender, education, nationality at birth (Swiss vs. foreigners), professional activity, linguistic region, political interest, intensity of left-right self-placement, internet use, and time elapsed since the election.[6] Finally, the model will also include a cross-level interaction between internet use and item type.[7] Details about the construction of all item-specific and individual-level variables are provided in Appendix A.3.

---

[4] For the sake of simplicity, the contextual variables in my model (linguistic region and interviewing mode, i.e., CATI vs. CAWI) are considered as individual-level variables.

[5] In fact, to avoid an accumulation of missing observations across items with different sets of non-respondents, RTs for non-respondents (i.e., those who gave a "don't know"/"refuse to answer" response) were recoded to the median value, provided that the number of missing values did not exceed 10 out of 24.

[6] In one version of the models (not shown here), I also included a squared term for age, to test for nonlinear effects of this variable. As one reviewer noted, it may be that not only the elderly, but also the younger respondents are slower, as the latter have less experience with the political system. However, adding age squared did not bring any improvement in model fit and the coefficient was systematically immaterial and nonsignificant.

[7] I assume that internet use is related to political information and thus may have a stronger influence on RTs to knowledge questions (compared to attitudinal and behavioral/factual questions). As a matter of fact, internet use is positively related to respondents' attention to political news (more detail available upon request to the author).

It is important to realize that the two level-2 units (items and individuals) are not related to one another in a hierarchical way, as in most multilevel data structures and models. As a matter of fact, items are not nested within individuals, because the same item is answered by many different individuals; likewise, individuals are not nested within items. Rather, the combination of items and individuals defines *cells* (with items considered, for example, as the row data and individuals as the column data) in which all RTs are classified. In other words, items and individuals are interwoven in a "complex data structure in which the lower-level units are cross-classified by two or more higher-level units" (Raudenbush & Bryk 2002: 373). Accordingly, the level-1 model can be thought of as a "within-cell model", and the level-2 model as a "between-cell model".

However, cross-classified models can be estimated much in the same way as other types of multilevel models. This is the second stage of my procedure. I estimate a model in which RTs are the dependent variable, with fixed effects for all item-level and individual-level characteristics, and random intercepts for individual-specific and item-specific effects. This conditional model takes the following form:

[1] $Y_{ijk} = \theta_0 + \gamma_{01}\text{education} + \gamma_{02}\text{age} + \beta_{01}\text{item type} + \beta_{02}\text{item length} + b_{00j} + c_{00k} + e_{ijk}$,

where the subscript *ijk* refers to the *i*th RT measured for respondent *j* to question *k*. [1] is the fixed-effects model containing, *for illustration purposes*, only two predictors for respondents and two predictors for items. Of course, the full model specified above — with ten respondent-level predictors and six item-level predictors — will be tested.

To assess the contribution of predictors in explaining RTs variance, I compare the fixed-effects model in [1] to an intercept-only model which serves as the null (or unconditional) model[8]:

[2] $Y_{ijk} = \theta_0 + b_{00j} + c_{00k} + e_{ijk}$

To get as much information as possible, I will test three conditional models. Model 1 will explore the role of the interviewing mode, to determine whether the data can be analyzed as a whole or should be analyzed separately for CATI and CAWI respondents. Accordingly, Model 1 includes a dummy for the CAWI/CATI mode as well as two *interactions*: (1) between CAWI and internet use (assuming that familiarity with the internet will decrease RTs first and foremost among CAWI interviewees); and (2) between CAWI and item type. Models 2 and 3 will include the same variables as Model 1 but will test their effects separately for CATI and CAWI respondents — hence, they will dispense with the CAWI variable and its interactions with other variables.

Model estimation

Table A.2 presents the results of the three predictive models. By design, variance in RTs is only explained at the respondent and item levels (i.e., in the "between-cell" part of the model).[9] Starting with Model 1, it should be noted that the predictors account for 42% of the

[8] This unconditional model will not be presented in the following section. It can be obtained from the author upon request.

[9] This is understandable enough, given the nature of level-1 units and the fact that no predictor was measured at this level (for a similar conclusion, see Yan & Tourangeau 2008). In principle, there *could* be level-1 predictors, such as the time elapsed between a given answer and the beginning of the interview. For practical purposes, however, the item position in the questionnaire (measured at the item level) is a good proxy for this level-1 measure.

variance at the level of respondents, for 67% of the variance at the item level, and for 30% of the overall variance in RTs.[10] Fixed effects demonstrate that younger, more educated, and Swiss-born respondents are especially likely to provide fast responses. Likewise, quick respondents tend to be concentrated among respondents living in the German-speaking area, among people who have more autonomy in their professional activities (professionals, executive employees, self-employed), among people with intense ideological positions, as well as people who use the internet every day. At the item level, item length is the most important predictor of RTs — the more words in a question, the longer it takes to answer that question. Likewise, a high number of answer categories undermines reaction times. Interestingly, all other things being equal, item uncertainty (as indicated by a higher percentage of DK and non-responses) tends to *reduce* RTs, which may indicate a general propensity to "satisfice" rather than to provide the "best" answer as possible. Importantly for my present purpose, knowledge questions take more time to answer than attitudinal and behavioral questions, even after controlling for other item-specific factors. In addition, internet use has less influence on knowledge questions than on other questions — RTs to knowledge questions are the least facilitated by heavy internet use.

Finally, some of the above tendencies are reinforced or attenuated in the case of on-line respondents, as indicated by the interactions in the last rows of the fixed effects panel of Table 2. Thus, heavy internet use speeds up responses mainly among CAWI respondents. Likewise, on-line responding widens the time gap between knowledge and behavioral items, but it reduces the gap between knowledge and attitudinal items. For the sake of simplicity, I did not probe further interactions with interviewing mode, but the evidence just presented justifies a test of two separate models for CATI and CAWI respondents.

Model 2 was estimated for the telephone-administered part of the 2015 Selects survey (total N=969). As CATI respondents represented only 18% of all respondents in 2015, it is possible that substantial differences between the CATI and CAWI modes went unnoticed in the total sample; therefore, I will focus here on the major differences between Model 2 and the two other models. To begin with, education, profession, linguistic region and ideological polarization play a lesser role or no role at all. Importantly, internet use is *not* related to RTs in the CATI mode, at least for knowledge and behavioral questions.[11] In contrast, the number of days elapsed since election day is negatively related to RTs.[12] Next, the time gap between knowledge items and attitudinal or behavioral items is lesser than for CAWI respondents (cf. Model 3). This is consistent with the idea that cheating is a time-consuming behavior which tends to be concentrated among online respondents. Further, contrary to CAWI, item uncertainty does not play a role for CATI respondents, while the placement of questions in the questionnaire does — suggesting that the telephone mode elicits less incentives for satisficing overall, but that satisficing might become more frequent with interviewing fatigue and

[10] These estimates are based on a comparison of variance components between the unconditional model (not shown here) and the conditional model displayed in Table 2 (see Snijders & Bosker 1999: chap. 7; Bickel 2007: 131-134). The same holds for Models 2 and 3.

[11] This may suggest that familiarity with the internet *per se* does not matter in a human interaction setting (and despite the fact that 43% of CATI respondents *are* daily internet users). However, it seems that daily internet users are quicker than non-users in providing answers to attitudinal questions.

[12] This difference between CAWI and CATI respondents may largely occur for artificial reasons. Online participation may be mainly driven by political interest and subjective competence; hence, the first to complete the online questionnaire would be citizens with the most ready-to-tell attitudes and beliefs. In contrast, the timing of telephone interviews may be more dependent on the sheer availability of respondents. If the most interested and competent respondents are also the most "busy" and least likely to be found home, it may take more time to reach them for an interview.

saturation. Finally, it is worth mentioning that Model 2 explains 25% of variance at the respondent level, 90% at the item level, and 53% overall.

Model 3 was estimated for the online part of the survey (total N=4368). The fixed effects displayed in Model 3 are essentially similar to Model 1, because the empirical bases of the two models are largely redundant, and the main results need not be repeated here. However, Model 3 has the worst fit of all models, with 23% of explained variance at the respondent level, 75% at the item level, and 27% overall.[13] This is an important aspect for the next stage of my analysis, because it means that there is more unexplained variance (residuals) for CAWI respondents, and thus more deviation from the expected baseline of RTs.

[13] Arguably, Model 1 shows a better fit than Model 3 only because it estimates some part of the differences between the predictors of Models 2 and 3 through the CATI vs. CAWI dummy. Overall, though, the results in Table 2 are entirely consistent with those obtained, with similar methods, by Yan and Tourangeau (2008).

Table A.2: Cross-classified models predicting RTs (restricted maximum likelihood estimates)

	M1: All respondents		**M2: CATI resp.**		**M3: CAWI resp.**	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Fixed effects						
Intercept	15.49*	7.18	17.62**	5.84	18.49**	6.15
Age	0.07***	0.00	0.05***	0.01	0.07***	0.00
Gender (male)	-0.18	0.11	-0.38	0.23	-0.15	0.13
Education	-0.21***	0.03	-0.11	0.06	-0.22***	0.03
Swiss at birth (ref: foreign-born)	-0.93***	0.14	-1.69***	0.30	-0.78***	0.16
Profession: executive (ref: missing)	-1.05***	0.30	-0.99	0.89	-1.10***	0.32
Profession: supervisory (ref: missing)	-0.51	0.27	-0.42	0.77	-0.50	0.30
Profession: operative (ref: missing)	-0.04	0.25	0.36	0.66	-0.07	0.27
Profession: self-employed (ref: missing)	-0.92**	0.31	-0.12	0.75	-1.05**	0.34
Profession: inactive (ref: missing)	0.23	0.25	0.70	0.65	0.18	0.27
German-speaking region (ref: Italian-sp.)	-1.25***	0.16	0.50	0.32	-1.64***	0.18
French-speaking region (ref: Italian-sp.)	-0.29	0.17	0.30	0.35	-0.45*	0.20
Political interest	0.00	0.08	-0.02	0.15	0.00	0.09
Intensity of left-right self-placement	-0.09*	0.03	-0.04	0.07	-0.11**	0.04
Internet use	-0.60***	0.15	0.12	0.20	-0.47**	0.18
Days elapsed since election	0.01	0.01	-0.05***	0.01	0.02**	0.01
Item type: Attitudinal (ref: Knowledge)	-22.39***	5.39	-14.93**	5.08	-18.57**	4.74
Item type: Behavioral (ref: Knowledge)	-24.11***	5.89	-12.44*	5.30	-20.92***	5.12
Internet use × Attitudinal item	-0.66***	0.12	-0.67***	0.18	-0.65***	0.16
Internet use × Behavioral item	-0.38**	0.12	-0.14	0.18	-0.54***	0.16
Item uncertainty (% DK/NA)	-1.28*	0.49	-0.21	0.45	-1.07*	0.43
Item position in the questionnaire	-0.01	0.03	-0.10**	0.03	0.00	0.03
Item categories: Scale (ref: non-scale)	-3.33	2.78	-3.34	2.27	-3.10	2.35
Item categories: Number	1.52***	0.26	1.37***	0.21	1.37***	0.22
Item length (# words)	0.87***	0.01	0.67***	0.09	0.55***	0.10
CAWI (ref: CATI)	0.68*	0.34				
CAWI × Internet use	-1.00***	0.18				
CAWI × Attitudinal item	1.31***	0.24				
CAWI × Behavioral item	-0.95***	0.22				
Random effects (variance components)						
Respondent variance: var(b_{00j})	11.835***	0.304	8.316***	0.531	12.523***	0.351
Item variance: var(c_{00k})	18.131**	6.246	11.700**	4.164	12.859**	4.554
Residuals: var(e_{ijk})	88.730***	0.361	73.689***	0.710	89.137***	0.400
Information criteria						
–2 Restricted Log Likelihood	930665.0		162011.3		765103.8	
AIC (Akaike's Information Criterion)	930671.0		162017.3		765109.8	
BIC (Schwarz's Bayesian Criterion)	930700.2		162041.4		765138.4	
Percent variance explained						
Respondent level	41.6		24.7		22.6	
Item level	67.0		89.9		75.4	
Overall	30.2		53.3		27.2	
$\mathbf{N}_i$	126023		22508		103515	
$\mathbf{N}_j$	5252		938		4314	
$\mathbf{N}_k$	24		24		24	

Note: Model estimated on unweighted cases. The indicated N_j is the maximum value; depending on items, the real N_j can be slightly lower (but anyway ≥5247 for M1, ≥934 for M2, and ≥4312 for Model 3). ***: p<.001; **: p<.01; *: p<.05.

Appendix A.5: Robustness checks for the validity analysis

Convergent validity analysis

As explained in the paper, my analysis of convergent validity focused on one particular way of defining potential "cheaters", namely all respondents who fell in the highest two deciles of the residuals' distribution for K2 **or** in the highest decile for K3. According to this rather inclusive measure (hereafter Definition #1A), about 26% of respondents can be considered as potential cheaters. However, I considered other measures varying along a dimension of *inclusivity* of the "cheating group".

First, I constructed two less inclusive measures of cheating comprising (a) only CAWI respondents in the last two K2 deciles **or** all respondents in the last K3 decile, and (b) only CAWI respondents in the last two K2 deciles **or** CAWI respondents in the last K3 decile. Secondly, I also determined a more restrictive variant for all three measures, identifying the respondents who qualify as cheaters for **both** K2 and K3 items (i.e., **AND** operator), rather than potential cheaters on **either** K2 or K3 (i.e., **OR** operator). Combining these two perspectives on cheating yields five additional definitions of potential cheaters (i.e., Definitions 1B to 2C), which are displayed in Table A.3.

The rationale for including or excluding CATI respondents depends on a rather subjective assessment of how likely it is that CATI respondents actually did look up answers on the internet, just like their CAWI counterparts. In fact, there is a bump in correct answers from CATI respondents in the last deciles of RT residuals for K3, but not for K2. However, the **threshold T** beyond which cheating appears most likely is the same for online and telephone respondents; it corresponds to the eighth decile of the residuals' distribution for K2 and to the ninth decile for K3.

Table A.3: Definition of potential cheaters

Definition	Operator	CAWI K2 > T	CAWI K3 > T	CATI K2 > T	CATI K3 > T
#1A (26.3%)	OR {	✓	✓	✓	✓
#1B (23.0%)	OR {	✓	✓		✓
#1C (21.2%)	OR {	✓	✓		
#2A (3.2%)	AND	✓	✓	✓	✓
#2B (3.0%)	AND	✓	✓		✓
#2C (3.0%)	AND	✓	✓		

As indicated in the first column of Table A.3, the proportion of potential cheaters varies slightly (between 26% and 21%) as a result of including or excluding CATI respondents; however, it decreases very sharply when cheaters are defined on the basis of highly positive RT residuals on *both* items (about 3%).

Table A.4 now displays the results of a regression model testing convergent validity for the various measures of potential "cheating". The models suggest two main conclusions. First, cheating appears to have the broadest impact on knowledge acquisition when measured according to Definition 1A (i.e., the definition adopted in the article). Second, given that no more than 3% of all respondents can be considered as cheaters on both items, the results for Definitions 2A to 2C are less conclusive.

Table A.4: Convergent validity analysis of the initial knowledge scale according to different definitions of the "cheating group" (OLS coefficients)

	Def. #1A	**Def. #1B**	**Def. #1C**	**Def. #2A**	**Def. #2B**	**Def. #2C**
Intercept	.366***	.374***	.381***	.386***	.386***	.386***
Age	.001**	.001*	.001*	.000	.000	.000
Sex (woman)	-.058***	-.065***	-.069***	-.064***	-.064***	-.064***
Education	.011***	.011***	.011***	.012***	.012***	.012***
Income: middle level[a]	.038***	.045***	.044***	.046***	.046***	.046***
Income: high level[a]	.076***	.077***	.078***	.082***	.081***	.081***
Region: French-speaking[b]	-.096***	-.099***	-.096***	-.101***	-.101***	-.101***
Region: Italian-speaking[b]	-.109***	-.106***	-.109***	-.107***	-.106***	-.106***
Dwelling place: small town[c]	-.014	-.015	-.018†	-.003	-.003	-.003
Dwelling place: big city[c]	-.048***	-.047***	-.046***	-.035***	-.034***	-.034***
Political interest	.106***	.104***	.102***	.102***	.102***	.102***
Cheating	.106**	.069†	.026	.156	.145	.145
Cheating × age	-.002***	-.001*	-.001†	.001	.001	.001
Cheating × sex	-.036*	-.012	.005	-.066	-.064	-.064
Cheating × education	.004	.004	.005	-.002	-.002	-.002
Cheating × middle income	.026	-.003	.002	-.054	-.078	-.078
Cheating × high income	.004	.004	-.002	-.131*	-.124*	-.124*
Cheating × French-speaking	-.025	-.017	-.029	-.048	-.047	-.047
Cheating × Italian-speaking	.009	-.006	.002	.001	-.015	-.015
Cheating × small town	.048*	.059**	.075***	.051	.057	.057
Cheating × big city	.068**	.067**	.068**	.123*	.130*	.130*
Cheating × political interest	-.022*	-.015	-.003	-.033	-.016	-.016
Adj. R^2	.208	.208	.210	.204	.205	.205
F-test	67.2***	67.2***	67.9***	65.6***	65.7***	65.7***
N	5286	5286	5286	5286	5286	5286

Notes: a: reference category=low level; b: ref. category=German-speaking; c: ref. category= countryside. ***: p<.001; **: p<.01; *: p<.05; †: p<.10.

Finally, additional analyses based on *separate* measures of potential cheating on the two knowledge items indicate that the effects of age and sex on political knowledge tend to be moderated by cheating on K2 (name of the president), while the effects of political interest and education tend to be moderated by cheating on K3 (required number of signatures for a federal initiative) — the effect of being an urbanite tends to be moderated by cheating on both K2 and K3 items. Detailed analyses can be obtained from the author upon request.

Predictive validity analysis

As shown above, alternative measures of cheating yielded similar results in terms of convergent validity, though Definition #1A seems most in line with expectations regarding the impact of cheating. To test for predictive validity, I constructed four main types of scales. Only the first two scales (Scales 1 and 2) were presented in the article. In Scale 1, correct responses of potential cheaters are cut down to a half point; in Scale 2, a full penalty (0 point) is given to suspected cheaters. In Scale 3, my procedure is based on the alternative idea that response times are *in themselves* a key component of knowledge. Accordingly, item

knowledge is no longer measured as a simple binary score but as a quantitative scale. For all knowledge items, a 0.05-point increment is attributed to each step down on the residuals' distribution broken down in twenty 5-percentile intervals — provided, of course, that the response is correct. In Scale 4, I follow the same logic as in the measurement of Scale 3, but this time using a logarithmic function: $\ln(1 \ldots 20) \cdot \ln(20)^{-1}$, where (1 … 20) is the simplified residuals' distribution described above. This function is used to increase the penalty for longer RTs. Finally, all scales are computed as the mean of the four items, no matter how the latter were measured.

For each of the four scales, I also distinguish between three variants (A, B, and C). For Scales 1 and 2, these three variants correspond to the varying inclusivity of the "cheating group" outlined by Definitions 1A to 1C in Table A.3 above. For Scales 3 and 4, the three variants correspond to the range of data to which the indicated transformations of knowledge scores are applied: (A) to all respondents on all items; (B) only to CAWI respondents on all items; or (C) only to CAWI respondents on items K2 and K3.

Table A.5 provides a short descriptive analysis of all scales, which shows that the various scales differ in terms of central tendency, dispersion, and skewness. For comparison purposes, the table also displays information about the initial knowledge scale (i.e, Scale #0). Quite logically, the recoded knowledge scales tend to exhibit lower scores in comparison with the initial scale; the difference is more pronounced with respect to Scales 3 and 4.

Table A.5: Descriptive analysis of knowledge scales (N=5317; all scales ranging 0–1)

Scale	Mean	Median	Standard Deviation	Skewness
0	0.62	0.75	0.28	-0.37
1A	0.59	0.63	0.27	-0.32
1B	0.59	0.63	0.27	-0.33
1C	0.59	0.63	0.27	-0.33
2A	0.56	0.50	0.28	-0.24
2B	0.57	0.50	0.28	-0.25
2C	0.57	0.50	0.28	-0.27
3A	0.33	0.33	0.19	0.21
3B	0.37	0.35	0.22	0.48
3C	0.49	0.50	0.25	-0.10
4A	0.44	0.45	0.22	-0.12
4B	0.46	0.48	0.23	-0.06
4C	0.53	0.50	0.26	-0.23

To assess the predictive validity of the various scales, I tested their impact on direct democratic participation (Table A.6) and on electoral participation (Table A.7). Compared to Scales 1A and 2A, which are fully discussed in the article, other scales appear less discriminant in terms of predictive validity — in fact, judging from the Bayesian information criterion, most of them (especially Scales 3 and 4) are even inferior to the initial scale. In other words, alternative measures do not allow to improve the predictive accuracy of knowledge scores in relation to political participation. Accordingly, it makes sense to recommend the use of Scale #2A, which has the best overall record as a predictor of both direct democratic and electoral participation.

Table A.6: Explaining direct democratic participation (0–10); OLS regression coefficients

	Scale 0	**Scale 1A**	**Scale 1B**	**Scale 1C**	**Scale 2A**	**Scale 2B**	**Scale 2C**	**Scale 3A**	**Scale 3B**	**Scale 3C**	**Scale 4A**	**Scale 4B**	**Scale 4C**
Intercept	1.59***	1.59***	1.60***	1.61***	1.62***	1.64***	1.66***	1.76***	1.89***	1.73***	1.68***	1.77***	1.67***
Age	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***
Sex (woman)	.39***	.40***	.40***	.39***	.40***	.40***	.39***	.39***	.34***	.38***	.40***	.37***	.39***
Education	.10***	.10***	.10***	.10***	.10***	.10***	.10***	.11***	.11***	.11***	.10***	.11***	.10***
Income: middle level[a]	.19*	.19*	.19*	.20*	.19*	.20*	.20*	.20*	.22**	.21*	.20*	.22**	.21*
Income: high level[a]	.24*	.23*	.24*	.24*	.24*	.24*	.24*	.27**	.29**	.27**	.25*	.27**	.25*
Region: French-speaking[b]	.32***	.33***	.33***	.32***	.32***	.32***	.31***	.26**	.24**	.30***	.29***	.27**	.31***
Region: Italian-speaking[b]	.50**	.52**	.51**	.51**	.52**	.52**	.51**	.47**	.42*	.49**	.49**	.46**	.50**
Dwelling place: small town[c]	-.11	-.11	-.11	-.11	-.11	-.11	-.11	-.11	-.11	-.10	-.11	-.11	-.10
Dwelling place: big city[c]	-.21*	-.21*	-.21*	-.21*	-.21*	-.21*	-.21*	-.22*	-.23*	-.21*	-.22*	-.22*	-.20*
Political interest	1.40***	1.40***	1.40***	1.40***	1.40***	1.40***	1.41***	1.42***	1.46***	1.42***	1.41***	1.43***	1.41***
Overall exposure to information	.53***	.54***	.54**	.54***	.54***	.55***	.55***	.56***	.57***	.55***	.56***	.57***	.55***
Political knowledge	.84***	.89***	.86***	.82***	.85***	.80***	.73***	.82***	.06	.65***	.85***	.45**	.75***
N	4983	4983	4983	4983	4983	4983	4983	4983	4983	4983	4983	4983	4983
Adj. R^2	.301	.302	.301	.301	.302	.301	.300	.299	.296	.299	.300	.297	.300
F-test	180.0***	180.4***	180.1***	179.7***	180.3***	179.7***	179.1***	177.7***	175.8***	177.9***	178.6***	176.7***	178.8***
AIC	8862.8	8859.5	8862.3	8865.3	8860.2	8865.4	8870.6	8882.4	8898.0	8880.1	8874.6	8890.4	8872.7
BIC	8947.5	8944.2	8947.0	8950.0	8944.9	8950.0	8955.3	8967.0	8982.7	8964.8	8959.2	8975.1	8957.4

Notes: a: reference category=low level; b: ref. category=German-speaking; c: ref. category=countryside. ***: p<.001; **: p<.01; *: p<.05.

Table A.7: Explaining electoral participation (0–1); log odds from logistic regression models

	Scale 0	**Scale 1A**	**Scale 1B**	**Scale 1C**	**Scale 2A**	**Scale 2B**	**Scale 2C**	**Scale 3A**	**Scale 3B**	**Scale 3C**	**Scale 4A**	**Scale 4B**	**Scale 4C**
Intercept	-3.68***	-3.72***	-3.71***	-3.70***	-3.72***	-3.70***	-3.69***	-3.51***	-3.38***	-3.56***	-3.63***	-3.54***	-3.64***
Age	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***	.02***
Sex (woman)	.15*	.17*	.17*	.16*	.18*	.17*	.17*	.15*	.11	.15*	.17*	.14	.16*
Education	.11***	.11***	.11***	.11***	.11***	.11***	.11***	.12***	.12***	.12***	.12***	.12***	.11***
Income: middle level[a]	.07	.07	.07	.07	.07	.07	.08	.08	.11	.09	.08	.10	.08
Income: high level[a]	.01	.00	.01	.01	.01	.01	.01	.05	.08	.04	.03	.05	.02
Region: French-speaking[b]	-.15	-.14	-.14	-.14	-.14	-.14	-.14	-.22**	-.24**	-.17	-.18*	-.20*	-.15
Region: Italian-speaking[b]	.48**	.52**	.52**	.51**	.54**	.53**	.53**	.46*	.40*	.49**	.50**	.47**	.52**
Dwelling place: small town[c]	-.29**	-.28**	-.28**	-.28**	-.28**	-.28**	-.28**	-.29**	-.28**	-.27**	-.28**	-.27**	-.27**
Dwelling place: big city[c]	-.44***	-.44***	-.44***	-.44***	-.43***	-.43***	-.43***	-.46***	-.45***	-.43***	-.45***	-.44***	-.43***
Political interest	.94***	.93***	.93***	.93***	.93***	.93***	.93***	.96***	.99***	.95***	.95***	.96***	.94***
Overall exposure to information	.59***	.59***	.59***	.59***	.60***	.60***	.60***	.62***	.63***	.60***	.61***	.62***	.60***
Political knowledge	1.12***	1.26***	1.23***	1.22***	1.28***	1.24***	1.21***	1.36***	.57**	1.12***	1.34***	.98***	1.21***
N	5230	5230	5230	5230	5230	5230	5230	5230	5230	5230	5230	5230	5230
Pseudo R^2 (Nagelkerke)	.344	.347	.346	.346	.348	.347	.346	.339	.333	.342	.343	.338	.344
Chi-square (all *p*s<.001)	1419.0	1432.3	1428.8	1427.3	1439.0	1432.4	1429.7	1397.6	1367.3	1407.3	1413.7	1391.4	1419.7
AIC	4737.0	4731.1	4734.5	4736.0	4720.2	4726.8	4731.8	4767.4	4797.7	4756.1	4751.2	4773.5	4742.0
BIC	4822.3	4816.4	4819.9	4821.3	4805.6	4812.1	4817.1	4852.7	4883.0	4841.4	4836.5	4858.8	4827.3

Notes: a: reference category=low level; b: ref. category=German-speaking; c: ref. category=countryside. ***: p<.001; **: p<.01; *: p<.05.

In judging which scale is the best predictor of participation, one should not put too much emphasis on regression coefficients, whose size can easily vary as a result of distributional differences between the various scales (see Table A.5). Instead, my analysis borrows from Raftery's (1995) suggestion to compare BIC values across models (be they nested or non-nested). The results of this analysis are reported in Table A.8. They show that Scales #1A and #2A are more "informative" than all other scales, including the original scale (#0). In turn, Scales #1A and #2A are on par with respect to direct democratic participation; but Scale #2A is clearly a better predictor of electoral participation. Overall, then, Scale #2A may be considered the most successful in terms of predictive validity.

Table A.8: Differences between BIC values (as reported in Tables A.6 and A.7); a positive value (in bold) indicates that the model including the scale in column 1 yields a better information criterion (i.e., a lower BIC value) than the scale in row 1

	Difference with Scale #0		**Difference with Scale #1A**		**Difference with Scale #2A**	
Scale	**DDP[a]**	**EP[b]**	**DDP**	**EP**	**DDP**	**EP**
#0	–	–	-3.3•	-5.9•	-2.6•	-16.8•••
#1A	**3.3•**	**5.9•**	–	–	0.7	-10.8•••
#1B	0.5	**2.5•**	-2.8•	-3.5•	-2.1•	-14.3•••
#1C	-2.5•	1.0	-5.8•	-5.0•	-5.1•	-15.8•••
#2A	**2.6•**	**16.8•••**	-0.7	**10.8•••**	–	–
#2B	-2.6•	**10.3•••**	-5.9•	**4.3•**	-5.1•	-6.5••
#2C	-7.9••	**5.2•**	-11.1•••	-0.7	-10.4•••	-11.6•••
#3A	-19.6•••	-30.3•••	-22.9•••	-36.3•••	-22.1•••	-47.1•••
#3B	-35.2•••	-60.6•••	-38.5•••	-66.6•••	-37.8•••	-77.4•••
#3C	-17.4•••	-19.0•••	-20.6•••	-25.0•••	-19.9•••	-35.8•••
#4A	-11.8•••	-14.1•••	-15.1•••	-20.1•••	-14.3•••	-30.9•••
#4B	-27.6•••	-36.4•••	-30.9•••	-42.4•••	-30.2•••	-53.2•••
#4C	-9.9••	-4.9•	-13.2•••	-10.9•••	-12.5•••	-21.7•••
#5	-3.6•	-7.1••	-6.9••	-13.0•••	-6.1••	-23.9•••
#6	-8.8••	-9.1••	-12.1•••	-15.1•••	-11.4•••	-25.9•••

Notes: **a**: Direct democratic participation; **b**: Electoral participation; •: positive evidence (△BIC=[2–6]); ••: strong evidence (△BIC=[6–10]); •••: very strong evidence (△BIC=[10+]); according to Raftery (1995, p. 139, Table 6); evidence based on absolute differences.

One may notice that Table A.8 comprises BIC comparisons for two additional scales (i.e., Scales #5 and #6), the purpose of which is explained in the next paragraph.

Is it really cheating or simply RTs?

In a final specification, I also tested the hypothesis that RTs *in themselves*, and not cheating, are involved in determining the validity of knowledge scales. If such were the case, it should be established that the validity of the scales is impaired by long RTs which are not confounded with cheating. Therefore, for this demonstration, I use the two items which are least likely to lead to cheating behavior (K1 and K4).

For the test of convergent validity, I define the "cheating group" as the sum of all respondents who gave a late response (i.e., a response falling in the last decile of RT residuals) to either item K1 or item K4. According to this definition, 18 percent of respondents can be considered as potential cheaters. However, when the cheating group is defined only in terms of RTs, and *not* on any substantiated suspicion of cheating, there is very little evidence that "cheating" modifies the impact of the usual predictors of political knowledge. The only exception is a significant interaction between cheating and having a high income (B= –0.05, p<.04) — a result which is not consistent with my previous analysis of convergent validity based on truly suspicious cases (see Definition 1, Table A.3). Likewise, a test of predictive validity does not confirm the alternative hypothesis that RTs in themselves account for the results obtained in this article. Thus, when "cheating" (as defined above) is given a half-point penalty (Scale #5) or a full penalty (Scale #6), the resulting knowledge scales perform badly as predictors of participation, in comparison with the original scale and with the "better" scales discussed above (see Table A.8 for evidence based on BIC values).

In sum, I find strong evidence that slow responding does not moderate the impact of the usual predictors of knowledge (convergent validity) and that both the initial scale and the revised scales described above do a better job at predicting political participation than the scales penalizing slow responding on K1 and K4 (predictive validity). The full results of this analysis can be obtained from the author.

References for all appendices

Bickel, Robert (2007). *Multilevel Analysis for Applied Research. It's Just Regression!* New York and London: The Guilford Press.

Faas, Thorsten and Jochen Mayerl (2010). "Michigan reloaded: Antwortlatenzzeiten als Moderatorvariablen in Modellen des Wahlverhaltens", pp. 259-276 in *Information – Wahrnehmung – Emotion. Politische Psychologie in der Wahl- und Einstellungsforschung*, T. Faas, K. Arzheimer und S. Rossteutscher (Hrsg.). Wiesbaden: VS Verlag.

Marquis, Lionel (2014). "The Psychology of Quick and Slow Answers: Issue Importance in the 2011 Swiss Parliamentary Elections", *Swiss Political Science Review* 20(4): 697-726.

Meyer, Marco and Harald Schoen (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.

Raudenbush, Stephen W. and Anthony S. Bryk (2002). *Hierarchical Linear Models. Applications and Data Analysis Methods (2nd ed.)*. Thousand Oaks (CA), etc.: Sage.

Snijders, Tom and Roel Bosker (1999). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*. London: Sage.

Stern, Michael 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.

Yan, Ting and Roger Tourangeau (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.