Online Supplemental Material

Indicators of Response Burden

Table S1. Overview of burden indicators.

Indicators of burden	Coding
Subjective burden indicators:	
The questionnaire was interesting.	1 = Completely agree 2 = Somewhat agree 3 = Neither agree nor disagree 4 = Somewhat disagree 5 = Completely disagree
The length of the questionnaire was adequate.	1 = Completely agree . . . 5 = Completely disagree
The questions were comprehensible.	1 = Completely agree . . . 5 = Completely disagree
Filling in the questionnaire presented no difficulty.	1 = Completely agree . . . 5 = Completely disagree
Mean score on all subjective burden items	1 = Low subjective burden . . . 5 = High subjective burden
Mean of interesting and length items	1 = Low subjective burden . . . 5 = High subjective burden
Objective burden indicators:	
No. of questions answered	App: 91 questions applicable to all respondents Browser: 96 questions applicable to all respondents with a smartphone; 83 questions applicable to all respondents without a smartphone
Completion duration in minutes	App: Sum of module completion times Browser: Sum of screen-by-screen times

Note. All indictors were measured in wave 1. Analytical sample contained 621 cases. For more information see also Tables 2 and 3.

Output of Principal Component Analysis of subjective burden indicators

Descriptive Statistics

	Mean	Std. deviation	Analysis N
Interesting	2.2480	.89786	621
Lengthadeq	2.1771	1.02274	621
Understandable	1.3865	.59466	621
Easy	1.4573	.74932	621

Correlation Matrix

		Interesting	Lengthadeq	Understandable	Easy
Correlation	interesting	1.000	.556	.316	.301
	lengthadeq	.556	1.000	.370	.302
	understandable	.316	.370	1.000	.562
	easy	.301	.302	.562	1.000
Sig. (1-tailed)	interesting		< .001	< .001	< .001
	lengthadeq	.000		.000	.000
	understandable	.000	.000		.000
	easy	.000	.000	.000	

Total Variance Explained

	Initial eigenvalues			Extraction sums of squared loadings		
Component	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	2.204	55.108	55.108	2.204	55.108	55.108
2	.917	22.926	78.035	.917	22.926	78.035
3	.467	11.669	89.704			
4	.412	10.296	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix [a]

	Component	
	1	2
interesting	.729	.504
lengthadeq	.753	.453
understandable	.761	-.437
easy	.726	-.517

Extraction Method: Principal Component Analysis.

[a.] 2 components extracted.

Computation of Weights

Because our comparisons of interest are confounded by selection effects on the samples responding using different devices/software, we used a propensity score weighting approach to try to balance the samples using auxiliary data available for all sample members from the register-based sampling frame. This follows general recommendations for addressing questions of causal inference in social research (Harder et al. 2010; Rosenbaum and Rubin 1983). We first computed an inverse probability weight (p_1) to address selectivity due to nonresponse at the first wave, using logistic regression to estimate the probability (π) of responding at wave 1 as follows:

$$\log\left(\frac{\pi_{iw1}}{1-\pi_{iw1}}\right) = \beta_0 + \sum\nolimits_{j=1}^{J} \beta_j^* \beta_{ij} \tag{1}$$

where π_i denotes the response probability of sample member i approached at wave 1 (w*1*); β_0 is the regression constant at wave 1, β_j is the effect coefficient for covariates j (frame variables, hence time constant); and β_{ij} is the value of covariate j for respondent i. Based on the logistic regression model (Equation 1), the estimated response propensities ($Y_{iw1} = 1|X_i$) at wave 1 translate into

$$Pr\left(Y_{iw1} = 1|X_i\right) = \frac{e^{\left[\beta_0 + \sum\nolimits_{j=1}^{J} \beta_j^* \beta_{ij}\right]}}{1 + e^{\left[\beta_0 + \sum\nolimits_{j=1}^{J} \beta_j^* \beta_{ij}\right]}} \tag{2}$$

where X_i is a vector of the characteristics from the frame variables. The weight (p_1) was then calculated as the inverse of the predicted probability of responding

$$p_1 = Pr\left(Y_{iw1} = 1|X_i\right)^{-1} = \frac{1}{\Pr(Y_i = 1|X_i)} \tag{3}$$

For the dependent variable, we used an indicator of whether the respondent had completed the survey up to and including the questionnaire evaluation items used as indicators of subjective burden (n = 621). As these measures were included in one of the last modules in the questionnaire, they were not always completed by those who 'broke off', but who nevertheless completed all substantive modules of the questionnaire. App respondents were significantly more likely to skip this module (labelled 'Your evaluation of this study'), presumably because it was of less direct relevance to the survey topic. Thus, the weight is intended to adjust for this difference also. Covariates were sociodemographic variables, including sex, age, marital status, household size, and whether resident in an urban or rural area (details of coding are available in Table S3). We additionally included an indicator of whether the respondent was randomly assigned to the app group or not.

For each of the pairwise comparisons across devices, we computed separate weights (p_2, p_3, and p_4) to control for the differential probability of responding using one software/device type compared to the other. To compute the propensity scores, we again estimated the parameters of logistic regression equations predicting the probability of responding on a given device for different pairs of device groups. Here, the estimated response propensities $Pr(Y_{iw1} = 1|X_i)$ refer to the probability that a respondent participated on

a) mobile browser ($Y_{iw1} = 1$) compared to a PC ($Y_{iw1} = 0$) (p_2);
b) mobile browser ($Y_{iw1} = 1$) compared to the app ($Y_{iw1} = 0$) (p_3); and
c) the app ($Y_{iw1} = 1$) compared to a PC browser ($Y_{iw1} = 0$) (p_4).

The covariates were the same as for (p_1), with the exception of the control for the experimental treatment. For respondents in the predicted category, the weight is equal to p, while for respondents in the reference category, the weight is equal to

$$p_{ref} = 1 - \Pr\left(Y_i = 1|X_i\right)^{-1} = 1 - \frac{1}{\Pr(Y_i = 1|X_i)} \tag{4}$$

The individual weights were normalised to preserve the totals responding in each device group (none of the propensity scores were trimmed). For the bivariate analyses comparing response burden between device/software groups, we multiplied the propensity score from the general nonresponse weight (p_1) by the propensity scores for each of the pairwise comparisons (e.g., $p_1 * p_2$), and then normalised the resultant weights to preserve the device group totals. For the remaining analyses, in which we used regression-based methods to test our mediation hypothesis, we use the general nonresponse weight (p_1) on its own. Parameter coefficients from all models are presented in Table S2.

Table S2. Coefficients of logistic regression models used to compute propensity score weights.

	(1) Responded at wave 1 (p1)		(2) Mobile versus PC (p2)		(3) Mobile versus App (p3)		(4) App versus PC (p4)	
	B	SE	*B*	SE	*B*	SE	*B*	SE
Respondent sex: Female	-0.03	0.1	0.42†	0.22	0.26	0.23	0.13	0.19
Respondent age (Ref. aged 18–30):								
Aged 31–55	-0.01	0.15	-0.33	0.31	-0.17	0.35	-0.14	0.31
Aged 56 and over	-0.19	0.19	-1.01***	0.41	-0.33	0.46	-0.80*	0.39
Respondent marital status (Ref. single):								
Married	0.34*	0.15	-0.07	0.31	-0.14	0.34	0.08	0.3
Divorced, widowed	0.02	0.18	0.09	0.39	0.2	0.44	-0.17	0.38
Respondent household size (Ref. 1hh):								
2 members	0.1	0.16	-0.38	0.35	-0.86*	0.39	0.45	0.34
3 or more members	0.03	0.16	-0.18	0.34	-0.82*	0.38	0.62	0.34
Lives in a rural area	0.15	0.11	-0.12	0.25	0.19	0.25	-0.3	0.21
Assigned to the app group	-0.23*	0.1	-	-	-	-	-	-
Constant	-1.04***	0.19	-0.13*	0.42	0.21	0.46	-0.29	0.4
N	2175		411		346		485	
Nagelkerke R^2	0.14		0.08		0.05		0.07	

Notes: †p-value $<$ 0.1, *p-value $<$ 0.05, ** p-value $<$ 0.01, *** p-value $<$ 0.001. B = Unstandardized beta coefficient; SE = standard errors, ref. = reference category.

Controlling for eonfounding of Assumptions in the Mediation Analysis

Table S3. Control variables and coding.

Potential confounding	Control variables	Coding
Confounding between the exposure (wave 1 device used) and the outcome (wave 2 drop-out)	*Sociodemographic variables from the sampling frame:*	
	Respondent sex	Female (1) Male (0)
	Age group	18–30 (0) 31–55 (1) 56 years and older (1)
	Marital status	Single, never married (0) Married (1) Divorced, widowed or separated (1)
	Household size	Single person household (0) 2 members (1) 3 or more members (1)
	Urbanicity of residential area	Rural (1) Urban (0)
	Self-report measures of respondent characteristics:	
	Occupational activity	In full-time paid work (0) In part-time work (1) Student/ apprentice/ in training (1) Not in paid work (retired, unemployed, home-maker) (1)
	Treatment group	Group 2 (invited to use the app at wave 1) (1) Group 1 (0)
Confounding between the mediator (experienced burden) and the outcome (wave 2 drop-out)	Level of education	Completed secondary-level or equivalent qualification (0)[1] Tertiary-level education (1)
	Topic interest	Very or somewhat interested in politics (1) Not at all or rather not interested (0)[2].

Table S3. Continued

Potential confounding	Control variables	Coding
	Intrinsic motivation (motivated to participate by the possibility to contribute to science)	Very or extremely important motivation for participating (1) Somewhat or unimportant motivation for participating (0)
	Self-report measures of digital literacy and device familiarity:	
	Frequency of internet use	Uses internet less than once a day (1) Uses internet more than once a day (0)
	Smartphone use	Does not use a smartphone to access internet (1) Uses a smartphone to access internet (0)
	Tablet use	Uses a tablet to access internet (1) Does not use a tablet (1)
Confounding between the exposure (wave 1 device used) and the mediator (experienced burden).	*Paradata:*	
	Survey completion time	Completion time in minutes, normalised based on +-2 standard deviations (SD) from the mean[3]

Notes: [1]15 cases with 'other' educational qualifications besides those listed were coded 0. [2]Two cases with missing values for interest in politics were coded to 'rather not interested'. [3]Completion times for a total of 16 cases (9 for the app, 7 for the browser) were two SD above the mean and were substituted with the average time taken by the remaining respondents.

In the following, we consider each of the confounding assumptions discussed by VanderWeele (2016, 19–21) and describe how we control for potential confounds (details of all control variables included in the mediation analysis are provided in Table S3). The first assumption to be addressed is that there is no confounding between the exposure (i.e., the device used at wave 1), and the outcome (i.e., the decision [not] to participate at wave 2). Key limitations of our research design are a) that there was no random assignment to browser type (PC versus mobile), and b) that there is selective non-response across the treatments, even where there was random assignment (browser versus app). As a result, the responding samples in each of the comparison groups are not directly comparable. If respondents with certain characteristics are, at the same time, more likely to respond on one type of device over another, *and* more likely to drop out of panel studies, then assumption 1 may be violated by this feature of our design. As mentioned, statistically significant differences were observed between the respondents using different devices/

software, which is why we used a propensity score weighting procedure to adjust for nonresponse at wave 1 and to balance the composition of the samples for the bivariate analyses. For the same reason, we include the same covariates (taken from the register-based sampling frame), with the same coding used to estimate the propensity score weights in each of the models for the mediation analysis.

In addition, we included in the mediation analysis the following other self-report measures of respondent characteristics from the wave 1 questionnaire that may be related to attrition: binary indicators representing main occupational activity (coded as: in full-time paid work [reference category [ref.]), in part-time work, student/ apprentice/ in training, or not employed (retired, unemployed, home-maker; two cases with missing values for this variable were coded to 'other', which was combined with the 'not in paid work' category). Because the groups are compared on the basis of their chosen response device/ software, and not on the basis of their original assignment, we also include an indicator of whether or not the respondent was invited to use the app at wave 1 (coded 1 if the respondent was in group 2, 0 if not). This also provides some control for the fact that the protocol for the group 1 respondents included a 'mode switch' between waves 1 and 2 (though they could continue using their original response device) and this could potentially have impacted their decision to take part in wave 2 (Sakshaug and Kreuter 2011; Sakshaug et al. 2010).

The second assumption to be addressed is that there is no confounding between the mediator (experienced burden) and the outcome (non-participation at wave 2). This assumption may be violated if the people who are more likely to experience burden are also more likely to drop out of panel studies (but not because of experienced burden per se). This would imply a common cause other than experienced burden, which could potentially be controlled in the analysis (Groves 2006). One such cause could be a respondent's *level of education*. Previous research finds, for example, that people with lower levels of education are less likely to participate in surveys generally and are also more likely to drop out of panel studies (Groves and Couper 1998). Respondents with lower education levels have also been found to have, on average, longer completion times, and have lower levels of interest in the survey topic (i.e., they experience greater burden [Yan and Tourangeau 2008]). *Interest in the survey topic* (politics) on its own could also account for differences in experienced burden and willingness to participate in subsequent panel waves (Groves et al. 2004). Thus, to control for this potential confounding between the mediator variable and the outcome, we also included the covariates for highest level of education (a binary indicator for having completed primary- secondary-level or other qualifications [ref.] versus tertiary-level education [coded 1]); and interest in politics (coded 1 if the respondent reporting being very or somewhat interested in politics, 0 if not at all, or rather not interested [two cases with missing values for interest in politics were coded to 'rather not interested']). Other *intrinsic motivations* to participate in the survey besides topic interest may also independently explain experiences of burden and willingness to continue participating. For this reason, we also included a binary indicator of whether the respondent was motivated to participated by the possibility to contribute to science (coded 1 if this was a very or extremely important motivation for participating, 0 if only somewhat or unimportant).

Other variables relating to *digital/IT literacy* and *device familiarity* have been shown elsewhere (e.g., Herzing and Blom 2019; Revilla et al. 2016; Wenz et al. 2019) to explain willingness to complete surveys on mobile devices (whether via an app or browser). On the assumption that these factors similarly influence experienced burden when completing online surveys, we additionally include the following covariates here: frequency of internet use (coded 1 if the respondent uses the internet less than once a day and 0 if more often) and binary indicators for whether the respondent uses a smartphone (coded 1 if not) or a tablet (coded 1 if used) to access the Internet. While each of these control covariates may have its own indirect effect on the outcome, as well as moderate the effects of others, their inclusion here is strictly to control for the confounding assumptions to enable us to draw conclusions about the possible mediating role of subjective burden.

The third assumption to be addressed is that there is no confounding between the exposure (i.e., device used at wave 1) and the mediator (i.e., experienced burden). As mentioned earlier, there were differences in the number of questions included in the browser and app questionnaires, which would have affected response duration (objective burden). Furthermore, the number of questions asked of respondents varied as a function of their characteristics (e.g., people in paid work were asked follow-up questions about their occupation). To the extent that these characteristics co-vary with the likelihood to complete the survey on a given response device, there may be systematic differences in the mean number of questions answered by device, resulting in varied experiences of objective burden. To control for this, we planned to include the indicator of the *number of questions answered,* however, as mentioned, due to collinearity with the device indicators and the response duration measure, we had to exclude this. Nevertheless, having decided to focus the mediation analysis on the subjective burden measure, we retained response duration in the model as a control for objective burden.

Finally, the fourth assumption to be addressed according to VanderWeele (2016, 20), is that none of the mediator-outcome confounders relating to assumption 2 (in this case, education, political interest, frequency of internet use, and being motivated to participate by the possibility to contribute to science) ‘are themselves affected by the exposure’ (i.e., the response device used at wave 1). Given that these characteristics are antecedent to the exposure and cannot be directly affected by it, we conclude that this assumption is not violated (though it is plausible that completing a questionnaire online improves familiarity with the device and that participating in a survey about politics increases interest in politics, thereby reducing burden, we assume such effects to be negligible).

References

Groves, R.M. 2006. “Nonresponse rates and nonresponse bias in household surveys.” *Public Opinion Quarterly* 70(5): 646–675. DOI: https://doi.org/10.1093/poq/nfl033.

Groves, R.M., and M.P. Couper. 1998. *Nonresponse in household interview surveys*. New York: John Wiley.

Groves, R.M., F. Fowler, M.P. Couper, J. Lepkowski, E. Singer, and R. Tourangeau. 2004. *Survey Methodology*. New York: John Wiley.

Harder, V.S., E.A. Stuart, and J.C. Anthony. 2010. “Propensity Score Techniques and the Assessment of Measured Covariate Balance to Test Causal Associations in

Psychological Research." *Psychological Methods* 15: 234–249. DOI: https://doi.org/10.1037/a0019623.

Herzing, J.M.E., and A.G. Blom. 2019. "The Influence of a Person's Digital Affinity on Unit Nonresponse and Attrition in an Online Panel. *Social Science Computer Review*, 37(3): 404–424. DOI: https://doi.org/10.1177/0894439318774758.

Rosenbaum, P.R., and D.B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55. DOI: https://doi.org/10.1093/biomet/70.1.41.

Revilla, M., Toninelli, D., and Ochoa, C. 2016. "PCs versus Smartphones in answering web surveys: does the device make a difference?. *Survey Practice*, 9(4): 1–6. DOI: https://doi.org/10.29115/sp-2016-0021Wenz et al., 2019.

Sakshaug, J.W., and Kreuter, F. 2011. "Using paradata and other auxiliary data to examine mode switch nonresponse in a "recruit-and-switch" telephone survey. *Journal of Official Statistics*, 27(2): 339–357. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/using-paradata-and-other-auxiliary-data-to-examine-mode-switch-nonresponse-in-a-34recruit-and-switch34-telephone-survey.pdf.

Sakshaug, J.W., T. Yan, and R. Tourangeau. 2010. "Measurement error, and mode of data collection: Tradeoffs in a multi-mode survey of sensitive and non-sensitive Items. *Public Opinion Quarterly*, 74(5): 907–933. DOI: https://doi.org/10.1093/poq/nfq057.

VanderWeele, T.J. 2016. "Mediation analysis: A practitioner's guide.? *Annual Review of Public Health* 37: 17–32. DOI: https://doi.org/0.1146/annurev-publhealth-032315-021402.

Wenz, A., A. Jäckle, and M.P. Couper. 2019. "Willingness to use mobile technologies for data collection in a probability household panel. *Survey Research Methods*, 13(1): 1– 22. DOI: https://doi.org/10.18148/srm/2019.v13i1.7298.

Yan, T., and R. Tourangeau. 2008. "Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology*, 22(1): 51–68. DOI: http://dx.doi.org/10.1002/acp. 1331.