This is a pre-copyedited, author-produced version of an article accepted for publication in the "Journal of Survey Statistics and Methodology" following peer review. The version of record is available online at: https://academic.oup.com/jssam/article/12/5/1254/7510744
DOI: 10.1093/jssam/smad044

Investigating Respondent Attention to Experimental Text Lengths

Tobias Rettig

Research Data Center, Mannheim University Library, University of Mannheim, Germany

Tobias.Rettig@uni-mannheim.de

Annelies G. Blom

SOCIUM Research Center on Inequality and Social Policy, University of Bremen, Germany

Digital Social Science Core Facility (DIGSSCORE), University of Bergen, Norway

Blom@uni-bremen.de

Abstract

Whether respondents pay adequate attention to a questionnaire has long been of concern to survey researchers. In this study, we measure respondents' attention with an instruction manipulation check. We investigate which respondents read question texts of experimentally varied lengths and which become inattentive in a probability-based online panel of the German population. We find that respondent attention is closely linked to text length. Individual response speed is strongly correlated with respondent attention, but a fixed cutoff time is unsuitable as a standalone attention indicator. Differing levels of attention are also associated with respondents' age, gender, education, panel experience, and the device used to complete the survey. Removal of inattentive respondents is thus likely to result in a biased remaining sample. Instead, questions should be curtailed to encourage respondents of different backgrounds and abilities to read them attentively and provide optimized answers.

Statement of Significance

To draw accurate conclusions from survey data, researchers rely on the accuracy of respondents' answers to all survey questions. To ensure that all responses were given with the necessary care and attention, implementing attention checks in surveys and excluding the low-quality data from inattentive respondents from analyses has become common advice. In this study, we investigate whether respondents from a probability-based online panel read a treatment text depending on its length. We further compare results from our attention measurement to respondents' response time, as fast response times are commonly used as an indicator for inattention. We find that this practice is likely to result in systematically misclassifying fast readers as inattentive and thus not suitable as a standalone indicator. We also investigate links between attention levels and the respondents' (socio-demographic) characteristics. Previous research on socio-demographic correlates of inattention has largely been conducted using nonprobability samples and often yielded mixed results, which is another gap in the existing literature we aim to fill with this study.

Keywords

Respondent attention, instruction manipulation check, online survey, experimental treatment, text length

Word count: 7,096

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – [Project-ID 139943784 – SFB 884].

Acknowledgments

The authors gratefully acknowledge the opportunity to include the experiment presented in this paper in wave 38 of the German Internet Panel (GIP, DOI: 10.4232/1.1339, Blom et al., 2019). GIP data are available to the scientific community from the GIP data archive as Scientific Use Files at https://dbk.gesis.org/dbksearch/GDesc2.asp?no=0109&tab=&ll=10&notabs=1&db=E. In addition, our analyses used wave 38 paradata (device and response time). These variables are available at the GIP On-Site Data Access (ODA) facilities (for further information, see https://www.uni-mannheim.de/en/gip/for-data-users/on-site-data-access/). A study description of the GIP can be found in Blom et al. (2015). The documentation of the GIP wave 38 source questionnaire is freely available at https://paneldata.org/gip/instruments/W38.

The authors also gratefully acknowledge the fruitful discussions with Dr. Jan Karem Höhne during the development stage of the project.

Corresponding author

Tobias Rettig, University of Mannheim

University Library, Research Data Center

Schloss Schneckenhof Nord, 68131 Mannheim, Germany.

Email: tobias.rettig@uni-mannheim.de

The Stata code for all analyses can be obtained from the authors upon request.

1. Introduction

Whether survey respondents answer a question with as much effort, care, and attention as the task at hand requires, has long been a concern among survey researchers (see, for example, Krosnick, 1991). Ideally, at each question respondents undergo a four-step cognitive response process (Strack & Martin, 1987; Tourangeau et al., 2000): Comprehending the question and understanding which information is requested, retrieving relevant information from memory, forming a judgment based on the retrieved information, and selecting the appropriate response. As Krosnick (1991) argues, not all respondents have the motivation to fully perform this cognitive process for each item and may instead use shortcuts to reach an acceptable response with lower cognitive effort (i.e., they "satisfice"). Krosnick distinguishes weak and strong satisficing. In weak satisficing, respondents undergo all four steps of the response process less thoroughly than would be required to reach an optimal response. In strong satisficing, respondents may skip directly from a superficial comprehension of the question to selecting a response that will appear reasonable while not reflecting any actual view that the respondent holds (Krosnick, 1991).

Self-administered surveys such as online surveys allow for a more extreme form of satisficing, because respondents can omit the first step of the cognitive process by selecting a response option without having read the question, instructions, text stimuli, or the response options (Anduiza & Galais, 2017; Shamon & Berning, 2020). Consequently, this disruption of the first step makes it impossible for respondents to undergo the other three steps. Without knowing which information is needed, respondents cannot retrieve any relevant information, form a judgment based on this information, or select a response that corresponds to this judgment. Clearly, such skipping behavior cannot result in a meaningful response. Consequently, researchers have devised both direct and indirect measurements to detect survey respondents who pay insufficient attention (overviews can be found in Curran, 2016 and Huang et al., 2012).

1.1 Indicators of respondent attention

A popular ad hoc indicator of attention is to look at the time it took respondents to fill out a questionnaire and exclude those with unreasonably fast response times, i.e. speeders (Aust et al., 2013; Curran, 2016; Meade & Craig, 2012). This practice follows the idea that a serious effort to read and carefully answer all items in a survey should take respondents a certain minimum amount of time and that saving time (and effort) is a primary motive for satisficing (Curran, 2016). However, while very fast response times are a good indication of inattention, a cutoff value to distinguish attentive from inattentive respondents has proven difficult to determine in practice (Aust et al., 2013; Curran, 2016; Meade & Craig, 2012; Niessen et al., 2016; R. K. Thomas, 2014).

Direct measurements of assessing respondent attention by using items specifically created for this purpose include instructed response items, so-called "bogus" items, and instruction manipulation checks (IMCs). Instructed response items directly ask respondents to pick a specific response option (such as "If you read this, check 'completely disagree'"), making the instructed response the "correct" option for attentive respondents and flagging the selection of any other response option (Curran, 2016; Huang et al., 2012; Meade & Craig, 2012). Bogus items work similarly, but instead of asking respondents to select a specific option, they have only one obvious correct response. For example, respondents should disagree with the statement "I was born on the 30^{th} of February" (Curran, 2016; Meade & Craig, 2012; Paolacci et al., 2010).

Whereas in theory both measurements lead to right or wrong answers that allow distinguishing attentive from inattentive respondents, they can also produce both false negatives (inattentive respondents picking the correct option by chance) and false positives, especially when respondents overthink the meaning of the item or take an impossible statement as hyperbole rather than literally (Curran & Hauser, 2019). Furthermore, respondents are often explicitly assured that researchers are interested in their opinions and that there are no right or wrong answers. The introduction of items with right and wrong answers runs counter to this assertion. Finally, as Goodman et al. (2013) noted, unusual questions and response options bear the risk of drawing attention and thus alerting respondents who were just skimming over the questionnaire.

In contrast, IMCs are inserted in the instructions preceding the actual survey question. They instruct respondents to perform an action that would not normally be expected of them, such as writing "I read the instructions" into the answer field (Hauser et al., 2016; Hauser & Schwarz, 2016; Oppenheimer et al., 2009). Asking for an unusual action in this way eliminates the possibility of respondents randomly passing the attention check without having read the instructions. One might retort that skipping such instruction texts is common among respondents who feel that they do not need additional help to answer the question. Thus, it would not necessarily mean that the response itself was not valid (Maniaci & Rogge 2014). However, questionnaire designers usually consider instructions a vital part of the question and, thus, a full response necessitates reading, comprehending and considering all provided information (see, e.g., Tourangeau et al., 2000). This is especially true when the additional text before a question is not merely instructional but includes substantive information which researchers want respondents to consider when answering and which may be experimentally varied across respondents (i.e., experimental treatments).

Receipt of such treatments is sometimes assessed with comprehension checks which require respondents to answer a factual question based on the received information (Kane et al., 2023; Kane & Barabas, 2019; Meisters et al., 2020; K. A. Thomas & Clifford, 2017). Beyond measuring attention, a comprehension check investigates whether respondents understood the provided information. Answering correctly requires respondents to hold the information in memory and be able to correctly repeat or draw inference from it. Depending on the specific research interest, this may be preferable to measuring mere attention (K. A. Thomas & Clifford, 2017). However, researchers may also risk excluding attentive respondents for not understanding the provided information and thus overestimate the effectiveness of complex treatments, as comprehension has been noted to correlate with education (Meisters et al., 2020).

The routine use of one or more measurements of respondents' attention has become a common suggestion in the literature (Abbey & Meloy, 2017; Curran, 2016; Meade & Craig, 2012; Shamon & Berning, 2020; K. A. Thomas & Clifford, 2017). However, due to the reactive nature of most direct attention measurements (i.e., at least respondents who pass the attention check are aware that their attention has been checked), some authors have also voiced concerns that they may anger respondents, reduce motivation, or alter response behavior (Gummer et al., 2021; Hauser et al., 2016; Peer et al., 2014). Evidence on whether these concerns are warranted is mixed. Hauser and Schwarz (2015) suggest that manipulation checks alter attention rather than just measure it and prompt more systematic thinking. Hauser et al. (2016), however, find that this is true for complex reasoning tasks but not regular survey questions. While Oppenheimer et al. (2009) note that they did not receive any complaints, Silber et al. (2022) find that respondents held mostly favorable, but also some unfavorable attitudes towards attention checks and that a small yet non-negligible portion of respondents intentionally did not comply with the instructions (see also Liu & Wronski, 2018). Compliance was, however, improved by providing respondents with a better explanation for the need for attention checks (Silber et al., 2022).

As a consequence of these potential drawbacks, some authors suggest more measured approaches to checking attention, considering factors such as the type, placement, and frequency of attention checks (Chandler et al., 2020), emphasizing the need for unobtrusive, low salience measurements and improving the survey materials themselves to be more comprehensible and better hold respondents' attention (K. A. Thomas & Clifford, 2017; R. K. Thomas, 2014).

1.2 Treating inattentive respondents

Respondents who answer a questionnaire with insufficient attention have been shown to provide lower quality data (Aust et al., 2013; Gummer et al., 2021; Huang et al., 2012; Peer et al., 2014), introduce error into the data (Chandler et al., 2020; Meade & Craig, 2012), reduce statistical power (Aust et al., 2013), and, as a consequence, mask effects present in attentive respondents (Goodman et al., 2013; Maniaci & Rogge, 2014). Consequently, it is now common practice to remove inattentive respondents from analyses (Aronow et al., 2019; Berinsky et al., 2014; Meade & Craig, 2012; Paolacci et al., 2010), especially when post-hoc identification methods are used after the end of fieldwork. Such removal can improve data quality (Aust et al., 2013; Shamon & Berning, 2020), improve scale measurement (Huang et al., 2012), and increase statistical power (Maniaci & Rogge, 2014). However, removing inattentive respondents may also introduce or worsen existing biases in the sample composition (Anduiza & Galais, 2017; Aronow et al., 2019; Berinsky et al., 2014; Oppenheimer et al., 2009).

Attention checks also give researchers the option to intervene during the interview. Respondents who failed an attention check may either be allowed to continue without intervention or be made aware of their failure to pass an attention test. Most researchers opt for the passive approach of letting inattentive respondents continue (and later remove them). A few studies have experimented with making those who failed an attention check retry to pass (Hauser et al., 2016; Hauser & Schwarz, 2015; Oppenheimer et al., 2009). Researchers' reservations with the latter approach may stem from the aforementioned concerns of angering respondents, especially those who failed the attention check, by making them feel caught.

Some authors suggest adding a warning about the importance of good data quality and the presence of attention tests throughout the survey. This procedure seems to increase attention but may also increase social desirability bias, decrease attitudes towards the survey, and the effects of such a warning may also depend on its wording (Clifford & Jerit, 2015; Meade & Craig, 2012). Furthermore, it is unclear whether the observed increase in attention is caused by respondents paying more attention to the survey questions or whether such a warning just puts them on alert to look out for attention checks.

1.3 Prior findings on respondent attention

The practice of removing respondents who were identified as not paying sufficient attention has sparked two main concerns about its effects: on the sample size and on the sample composition when certain groups of respondents are removed more frequently than others.

The size of the problem varies widely across studies, samples, types of attention tests, and cognitive effort required to pass these tests (Anduiza & Galais, 2017; Hauser & Schwarz, 2016;

Mancosu et al., 2019; Meade & Craig, 2012). In some settings, a large majority of respondents passed the attention checks and only a small percentage was flagged for inattention (Gummer et al., 2021; Hauser et al., 2016; Hauser & Schwarz, 2015; Johnson, 2005; Paolacci et al., 2010). In contrast, other studies have reported higher rates of attention check failures up to the point where a majority of respondents were flagged as inattentive (Hauser & Schwarz, 2016; Liu & Wronski, 2018; Shamon & Berning, 2020). Ex ante predicting how many respondents will be identified as inattentive in a given survey is thus difficult.

The effect on the sample composition of removing inattentives equally varies across studies. Whereas some report that male respondents tend to be less attentive than female respondents (Berinsky et al., 2014; Maniaci & Rogge, 2014; R. K. Thomas, 2014), others find no difference in attention between genders (Gummer et al., 2021; Mancosu et al., 2019; Oppenheimer et al., 2009). There seems to be some evidence that younger respondents are less attentive than older respondents (Anduiza & Galais, 2017; Berinsky et al., 2014; Gummer et al., 2021; Maniaci & Rogge, 2014; R. K. Thomas, 2014), but again, not all studies find this effect (Mancosu et al., 2019; Oppenheimer et al., 2009). Furthermore, lower attention may also correlate with lower education (Anduiza & Galais, 2017; Gummer et al., 2021; Mancosu et al., 2019; Maniaci & Rogge, 2014), although Thomas (2014) finds mixed results for different attention indicators and Berinsky et al. (2014) barely find any difference. Gummer et al. (2021) report slightly higher attention from respondents with prior survey experience, which the authors attribute to higher participation ability. This does, however, seem to contradict other research indicating that more experienced panel respondents tend to respond less carefully and engage in more satisficing behavior (Schonlau & Toepoel, 2015; Toepoel et al., 2008). Finally, Gummer et al. (2021) report no difference in attention across the devices respondents used to complete the survey.

It should be noted that these results are not only mixed, they also almost exclusively stem from studies that were conducted in nonprobability samples or samples of specific population subgroups, such as respondents from commercial access panels or student samples. This is notable for two reasons: First, both data quality and sample accuracy tend to be lower in nonprobability samples and can vary widely and unpredictably across different nonprobability samples (Brüggen et al., 2016; Cornesse et al., 2020; Cornesse & Blom, 2020; Pasek & Krosnick, 2020). Second, probability-based recruitments of respondents are generally associated with considerably higher costs and efforts (Baker et al., 2010; Brüggen et al., 2016; Pasek & Krosnick, 2020; Sakshaug et al., 2019; Wiśniowski et al., 2020). For researchers aiming to infer to the general population and working with a probability-based sample, removing inattentive respondents may be prohibitively expensive. Both the loss of any number of respondents and the resulting introduction of bias into the sample are undesirable outcomes. In addition, concerns about losing respondents who were angered by attention checks may contribute to the scarcity of attention research in probability-based samples.

2. Research questions

We have revealed several gaps in the literature that our study aims to fill. So far, direct attention measurements have generally been applied to individual questions or their instruction text, while comprehension checks for treatment texts measure more than mere attention. We aim to bridge between these approaches by measuring attention to a treatment text using an IMC.

Q1: Which proportion of the population skips reading a treatment text in an online survey?

In addition, an experimental variation of the text length allows us to investigate whether and to what extent respondent attention to the text treatment is dependent on its length.

Q2: How much does respondent attention vary with text length?

We investigate which sociodemographic groups (gender, age, education, and occupation) are at risk of being disproportionately excluded from analyses if they were removed following a failed attention check.

Q3: Which sociodemographic characteristics correlate with respondent inattention?

In addition, our experimental design in combination with a wealth of background characteristics and paradata also enables us to investigate the relationship between respondent attention and participation characteristics, in particular their level of panel experience, the device used to complete the survey, and how early in the field and at which time of the day respondents participated.

Q4: Are other participation indicators correlated with skipping a treatment text?

Finally, very short completion times are commonly used as an indicator for inattention. We, therefore, investigate how this indirect detection method compares to our direct attention measurement. Specifically, we investigate the relationship between the time respondents spend on the survey page that contains the treatment text and the result of the attention check. To be a useful measure of respondent attention, response time should both correlate strongly with the result of the attention check and allow researchers to set a sensible cutoff that distinguishes between attentive and inattentive respondents.

Q5: Is response time a suitable predictor of respondent attention?

3. Data and method

To answer our research questions, we implemented a survey experiment in the November 2018 wave of the German Internet Panel (GIP; Blom et al., 2019). The GIP is a probability-based online panel of the general population of Germany (see Blom et al., 2015, 2017). GIP panelists are surveyed online bimonthly with a questionnaire of 20 to 25 minutes. Recruitments into the panel were conducted in 2012, 2014, and 2018 among the general population aged 16 to 75 years living in private households in Germany at the time of recruitment. The November 2018 wave constitutes wave 38 of the GIP but was the first regular wave for the newly recruited 2018 sample. Placing our experiment in this particular wave thus allows us to make comparisons between freshly recruited panelists and experienced panelists from the 2012 and 2014 recruitments. The questionnaire was open for completion from November 1 to November 30, 2018.

3.1 Experimental design

Our attention check experiment was inserted in the middle of the questionnaire within a question on vote choice in a hypothetical referendum on Germany leaving the European Union. As similar topics are frequently surveyed in the GIP, this question was unlikely to draw any unusual attention. Within the text leading up to the question, we added an instruction asking respondents not to answer the question, but instead to click on the GIP logo in the top left corner of the page. This logo is shown on every page of each GIP questionnaire but usually has no function. Thus, we can be certain that no respondent clicked on the logo without having read our prompt to do so.

To maximize compliance and avoid angering respondents, the instruction was carefully worded to not give respondents the feeling that they are being monitored. The instruction thus read:

> "It is not always easy to stay concentrated throughout. To improve our study, we would like to know whether our participants carefully read the texts. Please do not answer the question below on Germany's membership in the EU. Instead, to continue please click on the logo of 'Society in Change' in the top left corner. This way, you show us that you have read this text." (survey fielded in German, own translation).

To investigate respondent attention to texts of different lengths, we designed four versions of the substantive question. The first version only displayed the instruction itself right above the survey question (see Figure 1, panel 1). In the other three versions, the instruction was placed within a text with one, two, or four paragraphs (see Figure 1, panels 2–4). To minimize the chance of respondents becoming alerted to the instruction while skimming over the text or after reading it partially, the instruction was placed before the final sentence of the last paragraph of each text. Respondents were randomly allocated to one of the attention check treatments. The length of the texts in the four versions ranged from 58 words to 347 words (not counting the question and response options).

Both clicking on the GIP logo (and thus passing the attention check) and clicking on the continue button (and thus failing the attention check) took respondents to the next survey page. Respondents continued with the survey independently of whether they had passed or failed the attention check. Those who failed were not made aware of this. There was no "back" button on the subsequent page.

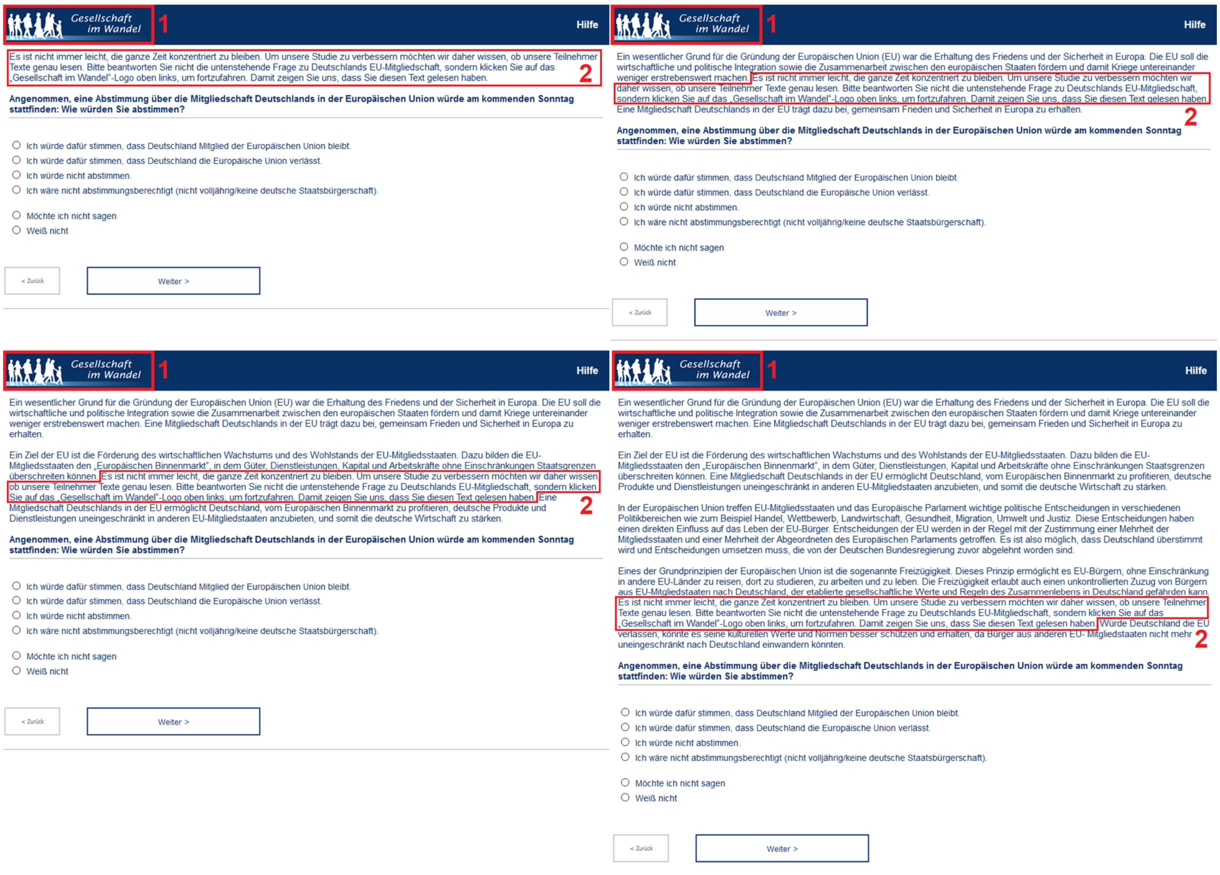

Figure 1. The four versions of the attention check with the logo (1) and instruction (2).

Note. Red boxes and numbers were not displayed on the original questionnaire.
[1] Respondents were instructed to click on this logo to continue.
[2] The position of the instruction text on the question page.

3.2 Sample description

In total, 5,827 panelists of the GIP were invited to participate in the November 2018 wave. Of these, 4,294 started the survey and 4,236 completed it, corresponding to a response rate of 72.7% and a cumulative response rate (from panel recruitment) of 12.0% (see DiSogra & Callegaro, 2009). Of the 4,294 respondents who started the survey, 38 (0.9%) broke off the survey before answering the attention check, 134 (3.1%) were not included in our experiment because the necessary JavaScript was not enabled on their device, and 83 (1.9%) were missing other information used in our analyses. These respondents were removed from our analytical sample (listwise case deletion). We also excluded 13 respondents (0.3%) who remained on the experiment's questionnaire page for an excessively long time (>20 minutes). Therefore, our analytical sample consists of 4,026 respondents who were included in our experiment and provided information on all variables used in our analyses.

Our sample had a median age of 51 years and 48.2% were female. 15.5% had low education, 28.9% medium-low, 22.3% medium-high, and 33.2% high education (see Appendix Table A1 for classification). 46.5% worked full-time, 17.3% part-time, 8.9% were unemployed or not in the labor force, 19.6% were retired, and 7.7% were in education or voluntary (civil or military) service. The newly recruited 2018 respondents accounted for 43.9% of the sample. 21.7% of respondents used smartphones to complete the survey.

During the field time, 34.7% of the sample participated by the end of day 3, 55.9% by the end of the first week, and 76.5% by the end of the second week. Finally, 25.6% participated in the morning (6am–noon), 38.7% in the afternoon (noon–6pm), 28.4% in the evening (6–10pm), and 7.3% at night (10pm–6am).

A random 25.3% of respondents (1,018) received the instruction-only version of the attention check, 25.1% (1,009) received one paragraph of text, 24.8% (999) two paragraphs, and 24.8% (1,000) four paragraphs. The random assignment to treatment delivered a uniform distribution of respondents with different backgrounds across treatment groups (no significant differences in gender, education, occupational status, panel experience, or smartphone use). There were also no differences in response date or time of day. Due to chance, younger respondents are slightly overrepresented in the one-paragraph condition and slightly underrepresented in the instruction-only condition (see Appendix Table A2 for distributions of respondents across experimental groups and χ^2-tests).

3.3 Analytical strategy

To answer our research questions formulated above, we conduct a number of analyses. First, we descriptively look at the proportions of respondents who passed and failed our IMC for each text length (*Q1* and *Q2*). To further investigate the relationship of text length and other respondent and participation characteristics with inattention, we compute three binary-logistic regression models of passing the IMC: (1) with text length as the only predictor (*Q2*), (2) with age (in 10-year categories, see Appendix Table A1), gender, education, and occupation as additional predictors (*Q3*), and (3) with panel experience, device, and the point during the field time and time of day respondents participated as additional predictors (*Q4*). We furthermore investigate the relationship between passing an IMC and response time (*Q5*). First, we descriptively look at the distributions of response times (collected server-side in full seconds) of respondents who passed and failed the IMC for each text length. We further analyze differences in the response times of respondents who passed and failed the IMC by conducting t-tests and compute Pearson's r as a measure of the strength of association between response time and attention. Due to the skewed nature of the distribution of response times, we conduct these analyses using the natural logarithm of response times. We then compare the results of our IMC to a response time-based classification of respondents as attentive or inattentive. For this we chose a cutoff time of 300ms per word (rounded down to the nearest full second), corresponding to cutoff times for the four experimental attention check length conditions of instruction only, one, two, and four paragraphs of 17, 35, 54 and 104 seconds, respectively. Some authors have suggested such a cutoff as an indicator for speeding based on respondents' reading of survey materials (Keusch & Yan, 2018; Zhang & Conrad, 2014). We compare the level of agreement between both attention indicators. We further compare the composition of the remaining sample following an exclusion of inattentive respondents based on the IMC and response time, and of cases in which the two attention indicators disagree (*Q3*, *Q4*, and *Q5*). We conduct all analyses without applying weights to the data.

4. Results

4.1 Passing the attention check

Overall, 60.8% of respondents complied with the instruction to click on the logo and thus passed the IMC. This passing rate varied greatly across experimental groups, ranging from 79.4% for the instruction-only condition and falling with each increase in text length to 41.1% for the four-paragraphs condition (Figure 2). A logistic regression model of the association between text length and the likelihood of passing the attention check (Model 1 in Table 1) confirms this significant correlation between longer text conditions and a lower likelihood of respondents passing the attention check. Computing separate models for each of the text lengths as the respective reference category (not shown) revealed significant differences across comparisons of all pairs of text lengths. This indicates that each increase in text length significantly decreases the proportion of respondents passing the IMC.

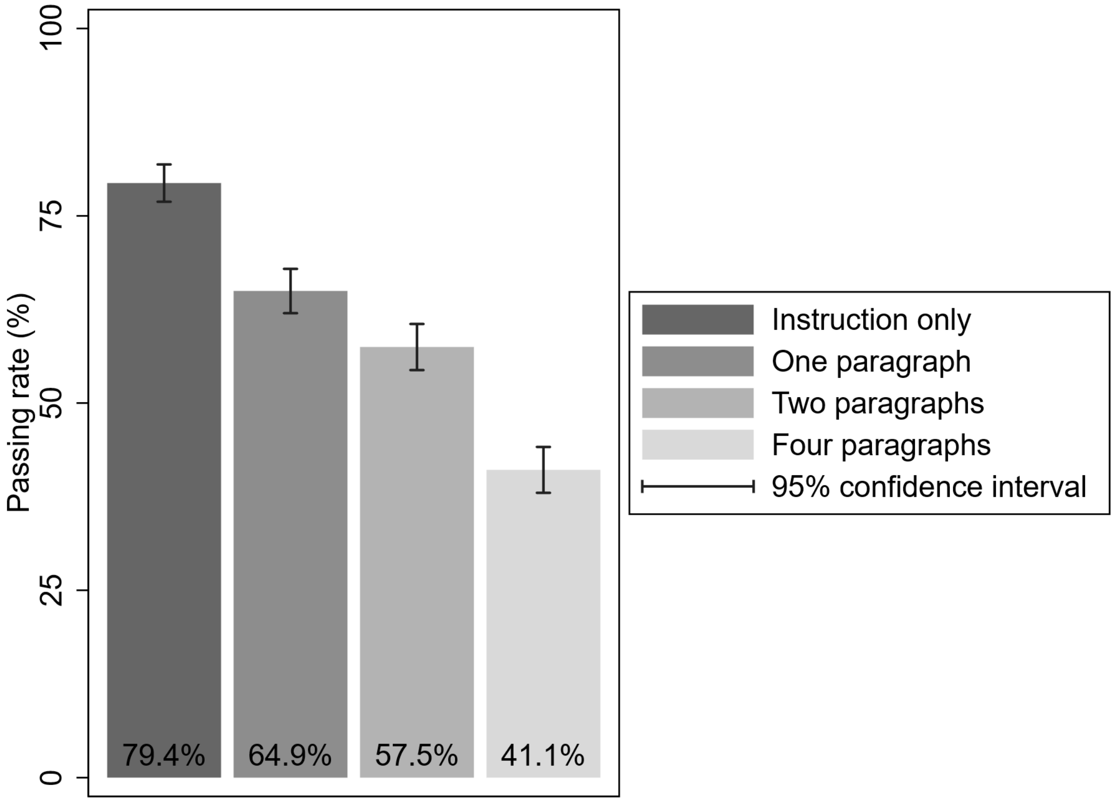

Figure 2. Passing rates for the attention check across the four text length conditions.

Note. Lower and upper boundaries of the displayed 95% confidence intervals:
instruction only: [76.9%, 81.9%]; one paragraph: [62.0%, 67.9%]; two paragraphs: [54.4%, 60.5%]; four paragraphs: [38.0%, 44.2%].

To further investigate which groups of respondents were more likely to pass the attention check (or which respondents were, in turn, more likely to skip reading the text), we computed two additional logistic regression models (Models 2 and 3 in Table 1). First, we have a look at the sociodemographics of respondents who passed and who failed the attention check (Model 2). Older respondents are more likely to pass the attention check. Changing the reference category (not shown) reveals that the age groups of 49 years and up do not significantly differ from each other but are consistently more likely to pass than the youngest respondents. Female respondents show a significantly higher chance of passing the attention check than male respondents.

Attention also differs across levels of education: Respondents with medium-low, medium-high, and high education are significantly more likely to pass the attention check than those with a low level of education. Changing the reference category (not shown) reveals that respondents with medium-high and high education are also significantly more likely to pass than those with a medium-low level of education. However, respondents with medium-high and high education do not significantly differ. Across occupations, we mostly find no differences. Only respondents in education or voluntary service seem to be more attentive compared to those who work full-time.

The effects observed in Model 2 remain stable when adding participation indicators in Model 3. We find that new GIP recruits are significantly more likely to pass the attention check than experienced respondents who have been panel members for four to six years. Curiously, respondents who completed the survey on a smartphone were significantly more likely to pass the attention check than those who used computers or tablets. It should be noted that in the displayed model we find this effect while adding (and thereby controlling for) respondents' age in the same model. That is, smartphone respondents are more attentive than non-smartphone respondents of the same age. However, smartphone use was not evenly distributed across age groups. A sensitivity analysis in a reduced model (question length and smartphone use as the only correlates) confirms the positive effect of smartphone completion on the passing rates (Appendix Table A3).

Table 1. Logistic regression models of passing the attention check.

	Model 1			**Model 2**			**Model 3**		
	OR	SE	p	OR	SE	p	OR	SE	p
Text length (ref.: Instruction only)									
One paragraph	.481	.049	.000	.474	.049	.000	.469	.049	.000
Two paragraphs	.351	.035	.000	.339	.035	.000	.336	.035	.000
Four paragraphs	.181	.018	.000	.171	.018	.000	.165	.017	.000
Age (ref.: <29 years)									
29–38 years				1.648	.243	.001	1.739	.260	.000
39–48 years				1.657	.247	.001	1.873	.285	.000
49–58 years				2.137	.304	.000	2.625	.390	.000
59–68 years				1.916	.302	.000	2.400	.396	.000
>68 years				2.069	.417	.000	2.743	.576	.000
Female				1.442	.108	.000	1.402	.106	.000
Education (ref.: low)									
Medium-low				1.288	.138	.018	1.286	.139	.020
Medium-high				1.821	.211	.000	1.812	.212	.000
High				2.009	.214	.000	2.099	.227	.000
Occupation (ref.: full-time work)									
Part-time work				1.188	.124	.099	1.192	.126	.097
Unemployed				1.169	.157	.245	1.153	.156	.293
Retired				1.001	.137	.991	.984	.136	.905
Education/service				1.559	.260	.008	1.508	.255	.015
New recruitment sample							1.625	.120	.000
Smartphone respondent							1.538	.145	.000
Field time (ref.: day 1)									
Days 2–3							1.150	.179	.369
Days 4–7							1.396	.222	.036
Days 8–14							1.387	.221	.040
After day 14							1.115	.173	.484
Daytime (ref.: morning)									
Afternoon							1.257	.111	.010
Evening							1.264	.126	.019
Night							1.396	.208	.025
Constant	3.848	.298	.000	1.152	.195	.403	.513	.122	.005
Observations	4,026			4,026			4,026		
Pseudo-$R^2_{\text{McKelvey \& Zavoina}}$	.103			.140			.168		
AIC	5,069			4,979			4,907		
BIC	5,094			5,087			5,071		

Note. OR = odds ratios, SE = standard errors.

The point in time during the 30-day field period that respondents participated in the survey did not consistently significantly correlate with attention. However, both early and late responders seem to be less likely to pass the attention check than those in between. Finally, we find that respondents who completed the questionnaire in the afternoon (from noon to before 6 pm), in the evening (from 6 pm to before 10 pm), or at night (from 10 pm to before 6 am) were significantly more likely to pass the attention check than those who participated in the morning (from 6 am to before noon). Selecting a different reference category did not reveal any further significant differences.

Adding interaction terms for age, gender, education, panel experience and smartphone use with text length did not reveal any consistent interaction effects or substantively different results (Appendix Table A4). Computing individual models with only one of these interaction terms each (not shown) did not result in any substantially different findings. In addition, the AIC and BIC values of these models did not indicate that the inclusion of any interaction terms resulted in an improved model fit over Model 3. We thus find no evidence that longer texts may disproportionately affect the attentiveness of certain respondent groups.

4.2 Response time as an indicator of attention

When web surveys do not include attention tests, researchers often resort to identifying inattentive respondents by setting a fixed response time cutoff. We evaluate this method by comparing the response times between respondents who passed and those who failed the attention check across the four text lengths. Figure 3 displays density curves for the response times of respondents who passed and those who failed the attention check separately for each of the four text lengths.

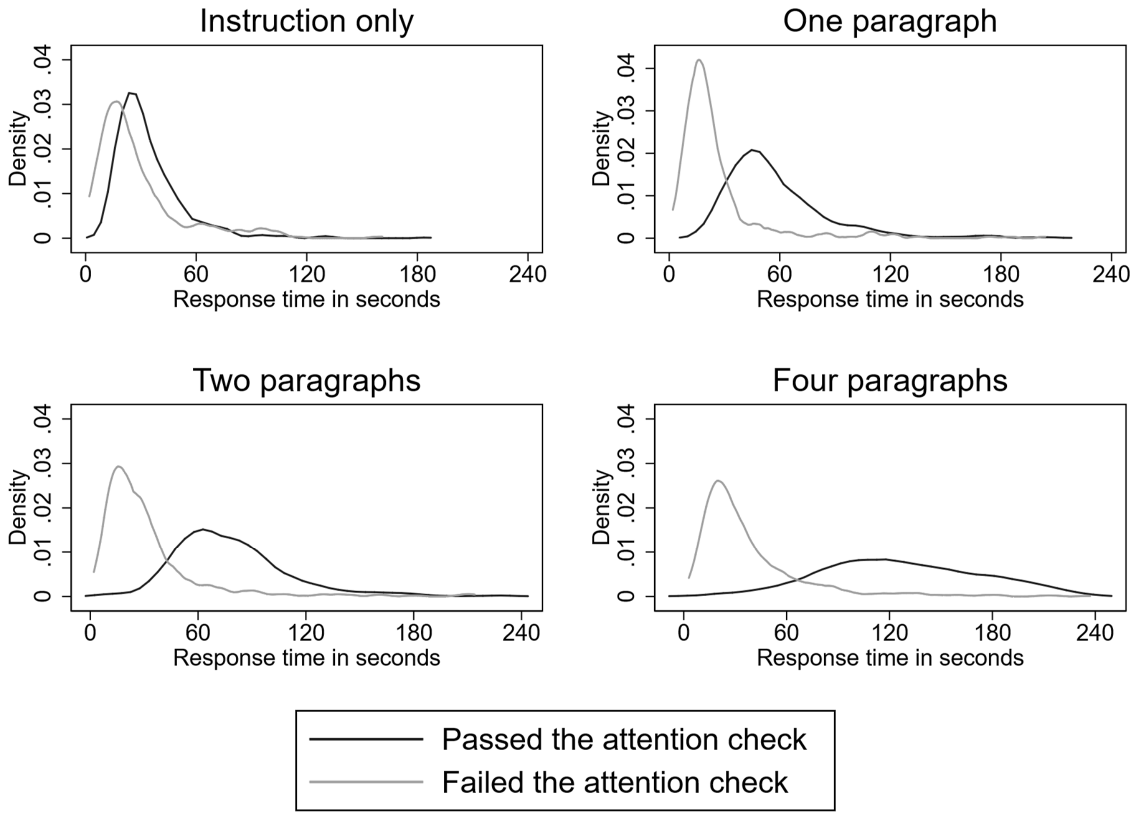

Figure 3. Density curves of response times in the attention check.

Note. Response times limited to 240 seconds (4 minutes) for display purposes.

With increasing text length, the mean response time of respondents who passed the attention check increases, as does the variance. As would be expected if failing the attention check equates to skipping the text, the mean and variance of the response times of respondents who failed the attention check are hardly affected by text length. We also see that there is considerable overlap between the two curves, especially in the shorter text conditions. Using a fixed response time cutoff to identify inattentive respondents would thus bear the risk of either not identifying a portion of the slower inattentive respondents, incorrectly identifying fast readers as inattentive, or both, especially in shorter texts.

This observation is confirmed by comparing the correlation of response time with the results from the attention check for each text length condition (Table 2). While the (logarithmized) mean response time significantly differs between respondents who passed and failed the attention check for each text length, the association of response time with attention is stronger for longer text conditions. Response time and attention are only weakly correlated for the instruction-only condition. However, for the longer texts, response time and attention are moderately to strongly correlated. For the longest text, whether respondents read the text accounts for about half of the variance in (logarithmized) response times.

Table 2. Key indicators of response time and two-tailed t-tests and Pearson's correlations of logarithmized response time with passing the attention check by text length.

	Response time			T-test			Pearson's corr.		
	mean	SD	flagged	t	df	p	r	p	R^2
Overall	58.6	67.2	42.6%	- 33.571	4,024	.000	.468	.000	.219
Instruction only	33.5	25.1	14.6%	- 8.172	1,016	.000	.248	.000	.062
One paragraph	49.5	49.9	40.3%	- 25.963	1,007	.000	.633	.000	.401
Two paragraphs	65.7	72.9	48.0%	- 28.579	997	.000	.671	.000	.450
Four paragraphs	86.0	90.6	67.8%	- 31.731	998	.000	.709	.000	.502

Note. Mean and standard deviation (SD) of response time (in seconds). "flagged" = percentage of respondents flagged as inattentive based on response time (<300ms per word). T-tests and Pearson's correlations based on logarithmized response time.

Comparing the results from the IMC to a time-based cutoff value of 300ms per word, the two attention measurements agree in their classification of 82.9% of respondents. About 10.2% were identified as too fast despite passing the IMC, while 6.9% of respondents passed the time-based classification despite failing the IMC. A kappa value of .646 (SE = .016, p = .000) for interrater agreement indicates moderate to good agreement between these measurements. Compared to the IMC, a response time cutoff would indicate an even stronger association between attention and text length, as passing rates range from 85.4% for the instruction-only condition to 32.2% for the longest text (compared to 79.4% and 41.1% for the IMC).

We further investigate potential biases in the remaining sample after removing inattentive respondents and to check whether a misidentification of respondents by using a fixed time cutoff may disproportionately affect certain groups of respondents. Table 3 shows the distribution of respondents' characteristics in the full sample compared to those in the remaining samples following an exclusion of respondents identified as inattentive by our IMC, a response time cutoff, and of cases in which the two indicators disagree.

Excluding respondents based on either the IMC or response time results in a remaining sample that is significantly different from the excluded respondents (and thus biased) in their age, gender, occupation, panel experience, smartphone use, and for the IMC additionally in their education. However, especially respondents with high education are, likely due to their higher reading speed, overrepresented among the respondents flagged by the response time cutoff despite passing the IMC. Respondents with high education are also underrepresented among respondents not flagged by the response time cutoff despite failing the IMC. The opposite pattern can be observed for low education. The result may be a remaining sample that seems unbiased with respect to education at the expense of a systematic misclassification of respondents with different levels of education. While we found older respondents to be more attentive, the older age groups (59 years and up) are also less represented among respondents who passed the IMC but were flagged for their fast response time, which may add to the resulting age bias in the remaining sample. We further find differences between respondents who passed and failed the IMC among those flagged for their response time (gender and occupation), as well as between those who passed and failed the IMC among those not flagged for their response time (gender, occupation, and panel experience).

Table 3. Distribution of respondents' characteristics in full and remaining samples following exclusion of intattentives and results of χ^2-tests for differences.

	Full sample	IMC	Response time		
		Remaining sample	Remaining sample	Flagged, IMC passed	Not flagged, IMC failed
Age		[a]	[a]	[b]	
<29 years	13.3%	11.8%	11.2%	15.2%	11.6%
29–38 years	15.2%	15.4%	14.6%	16.5%	11.2%
39–48 years	16.2%	16.0%	15.4%	19.4%	15.9%
49–58 years	24.4%	25.6%	25.2%	25.7%	22.0%
59–68 years	20.4%	20.7%	21.9%	16.7%	24.5%
>68 years	10.5%	10.5%	11.7%	6.5%	14.8%
Female	48.2%	51.6% [a]	49.8% [a]	55.7% [b]	42.6% [c]
Education		[a]		[b]	[c]
Low	15.5%	13.2%	15.2%	10.6%	25.3%
Medium-low	28.9%	27.7%	29.6%	21.1%	34.3%
Medium-high	22.3%	23.3%	22.1%	24.5%	14.8%
High	33.2%	35.8%	33.1%	43.8%	25.6%
Occupation		[a]	[a]	[b]	[c]
Full-time work	46.5%	44.6%	43.2%	49.2%	39.4%
Part-time work	17.3%	19.0%	18.1%	23.0%	16.6%
Unemployed	8.9%	9.4%	9.8%	7.0%	9.7%
Retired	19.6%	19.3%	21.7%	10.9%	27.1%
Education/service	7.7%	7.7%	7.2%	9.9%	7.2%
New recruitment sample	43.9%	48.0% [a]	48.7% [a]	38.7%	40.8% [c]
Smartphone respondent	21.7%	23.4% [a]	23.2% [a]	22.5%	20.2%

Note. See Appendix Table A5 for χ^2-test statistics of differences in sample composition.
[a]significant differences between excluded respondents and remaining sample.
[b]significant differences between flagged respondents who passed and failed the IMC.
[c]significant differences between not flagged respondents who passed and failed the IMC.

Overall, these results suggest that the exclusion of inattentive respondents leads to a biased remaining sample and that a response time cutoff as a proxy for inattention systematically misclassifies respondents. Comparing the sample composition of respondents who passed the two attention measurements for each text length shows few differences between treatments, again confirming the observation that longer texts do not seem to consistently increase inattention in specific groups of respondents disproportionately (see Appendix Tables A6 and A7).

5. Summary

This paper addresses respondent attention to survey questions in a probability sample of the general population. Five research questions guide our analyses. We discuss the results in turn:

Q1: Which proportion of the population skips reading a treatment text in an online survey?

Q2: How much does respondent attention vary with text length?

To answer these questions, we implemented an attention experiment in the middle of a 20–25 minutes long survey wave of a probability-based online panel. In the text leading up to a survey question, we instructed respondents to click the survey's logo instead of answering the question and experimentally varied the length of this text. We recorded whether respondents complied with this instruction to measure their attention to the text. With increasing text length, the proportion of respondents who passed the IMC decreased. Respondents who received the instruction only passed the check most frequently with 79.4%, whereas 64.9%, 57.5%, and

41.1% of respondents who received one, two, and four paragraphs, respectively, passed the check.

These findings have several implications for applied survey research. The association of respondent attention and text length reveals potential for optimizing respondent attention by keeping texts as short as possible. When designing questionnaires, the use of excessively long question texts should thus be avoided, or else a majority of respondents will likely not read the treatment carefully. This may be of particular importance in experiments that test different stimuli against each other. The effect of the stimuli may be confounded with the proportion of respondents who read the respective text. As a consequence, a detected treatment effect may only be an artifact of the text length.

Q3: Which sociodemographic characteristics correlate with respondent inattention?

We find that higher rates of failing the attention check in our experiment were associated with younger, male, and lower educated respondents, which is in line with previous results. The common practice of excluding inattentive respondents from analyses will therefore likely result in a biased remaining sample. Measures to optimize for respondent attention should thus be preferred over a post-hoc identification and exclusion of respondents. In addition, researchers should be cautious about making comparisons of treatment effects across sociodemographic groups, as the effectiveness of treatments across sociodemographic groups may be confounded with different rates of receiving the treatment.

While we found differences in the general attentiveness across certain groups of respondents, we did not find any consistent interaction effects with text length. Longer texts will generally be read by considerably fewer respondents and some groups of respondents are more likely to skip reading a text than others, but we did not find evidence that these group differences are more pronounced for longer texts.

Q4: Are other participation indicators correlated with skipping a treatment text?

In line with previous findings, we find that more experienced respondents are less attentive than newly recruited, inexperienced respondents. In addition, we find that respondents who completed the survey using smartphones were less likely to skip reading a treatment text than those using computers or tablets. This is notable, as concerns that smartphone respondents may provide lower quality data are commonly voiced in the literature (see, e.g., Lugtig & Toepoel, 2016; Struminskaya et al., 2015). Furthermore, while we may think of early responders as more motivated to participate in the survey, we find that both those who participated early during the field period (day 1 or days 2–3) and in the morning were less attentive. This indicates that some early responders may be more motivated to have the survey "off their desk" instead of taking the time to participate when it is most convenient for them. Late responders who participated after more than two weeks are, however, also associated with lower attention. Survey practitioners should thus consider up to which point leaving the survey open for participants and sending out additional reminders are worthwhile, as fewer and less attentive respondents seem to participate late in the field period.

Q5: Is response time a suitable predictor of respondent attention?

Our findings are twofold: The time respondents spent on the survey page containing the treatment text was moderately to strongly correlated with whether they read it, particularly for the longer text conditions. The interrater agreement between our IMC and a response time-based identification method was moderate to good. In this sense, response time is a good predictor of respondent attention.

However, we also find that using a hard response time cutoff value as a proxy for respondent attention bears considerable risk of misidentifying slower inattentive respondents as attentive (false negatives) and fast readers as inattentive (false positives). In addition, we found evidence that such misidentifications may disproportionately affect specific respondent groups. For instance, we found that older respondents who passed the IMC were less likely to be (falsely) flagged as inattentive based on their response time. In addition, older respondents were also less likely to fail the IMC. Removing respondents based on their response time may thus result in an age bias in the remaining sample that is driven by two factors: Genuinely lower attentiveness among younger respondents and the disproportionate misidentification of younger respondents as inattentive. In addition, we found evidence that other biases may be masked. Respondents with high education were more likely to pass the attention check but also tended to have faster response times and were thus more likely misidentified as inattentive, and less likely misidentified as attentive. The result may be a sample that looks unbiased regarding respondents' education levels at the cost of a disproportionate exclusion of attentive respondents with high education. While an exceedingly fast response time may be an indication of respondents who answer a survey without the required care and attention, we advise against the use of a fixed cutoff time as the sole indicator of respondent attention.

6. Conclusion

Whenever inferences are made from survey data, researchers need to be certain that respondents paid adequate attention to the questions that they have been asked. By means of an IMC experiment, our study investigates the prevalence of inattention, the relationship between inattention and text length, sociodemographic and other participation indicator correlates of inattention, as well as the suitability of response times as a proxy for attention.

We find that respondent attention to text stimuli decreases with increasing text length. Furthermore, individual response speed is moderately to strongly correlated with respondent attention. However, the use of a fixed cutoff time to separate attentive from inattentive respondents bears a considerable risk of creating both false positives and false negatives. In addition, we find that such misidentification is linked to respondents' age, gender, education level, occupation, and level of panel experience. A fixed response time cutoff is therefore not suitable as a standalone indicator of respondent attention. Differing levels of attention are also correlated with respondents' age, gender, education, panel experience, and the device used to complete the survey. An exclusion of inattentive respondents based on either an attention check or response time is likely to result in a biased remaining sample. Instead, researchers should optimize questions and treatment texts to encourage respondents of different backgrounds and abilities to read them carefully and provide optimized responses.

In addition to new insights on the topic of respondent attention and practical suggestions for both applied survey research and survey management, our study also points towards areas that require further research. We conducted one specific experiment in a specific way. Variations on this experiment would give insights into the generalizability of our findings across different settings. We placed the experiment in the middle of a 20–25-minute online questionnaire within a question on vote choice in a hypothetical referendum on Germany leaving the European Union. Our data thus do not allow us to investigate whether attention levels would have been different earlier or later in the same survey, in a different sample or survey mode, for a substantially longer or shorter questionnaire, or for a different topic. Respondents may, for example, already have strong feelings on a specific issue such as EU membership and feel that they do not need to read any additional materials to give their answer. While this would still constitute inattention in the sense that the survey materials were not read, it would not necessarily indicate satisficing or providing poor-quality responses. Respondents who do not read treatment texts because they have already made up their mind on the topic may in fact be a legitimate factor in evaluating the effectiveness of treatments. However, present techniques to measure attention to treatment texts, such as IMCs, comprehension checks, or response time cutoffs, are not equipped to distinguish such cases from actual satisficing.

Moreover, a systematic evaluation of respondents' reactions to the attention check, potential intentional noncompliance, and interventions on inattentive respondents were beyond the scope of our study. Concerns about angering respondents and potentially altering their response behavior seem to be the key reason why such interventions are hardly explored and used in practice. Evidence on whether these concerns are justified is sparse, yet growing in recent years and offers great value in guiding researchers' decision on whether to implement an attention check or not. Further investigation into unintended consequences of (reactive) attention measurement and noncompliance is therefore needed.

Furthermore, the literature on respondent attention is currently divided on whether inattention at one point in time is a good predictor of whether respondents will be inattentive at a later point (Anduiza & Galais, 2017; Berinsky et al., 2014; Gummer et al., 2021). Our study cannot contribute to this important debate, as our experiment has only been implemented once so far. Much value may be generated from insights on how to treat respondents in longitudinal studies because the decision to remove a long-term panelist from the survey is particularly costly here. However, it may also be undesirable to keep respondents in longitudinal studies if they consistently participate without the required care and attention.

We thus encourage researchers and survey practitioners to implement different attention measurements in their surveys to gather more evidence on how to best prevent, detect, and handle inattentive respondents, as well as respondents' reactions and attitudes towards the (repeated) exposure to attention checks.

References

Abbey, J. D., & Meloy, M. G. (2017). Attention by Design: Using Attention Checks to Detect Inattentive Respondents and Improve Data Quality. *Journal of Operations Management*, *53–56*(1), 63–70. https://doi.org/10.1016/j.jom.2017.06.001

Anduiza, E., & Galais, C. (2017). Answering Without Reading: IMCs and Strong Satisficing in Online Surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519. https://doi.org/10.1093/ijpor/edw007

Aronow, P. M., Baron, J., & Pinson, L. (2019). A Note on Dropping Experimental Subjects who Fail a Manipulation Check. *Political Analysis*, *27*(4), 572–589. https://doi.org/10.1017/pan.2019.5

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness Checks Are Useful to Improve Data Validity in Online Research. *Behavior Research Methods*, *45*(2), 527–535. https://doi.org/10.3758/s13428-012-0265-2

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D., Frankel, M. R., Garland, P., Groves, R. M., Kennedy, C., Krosnick, J., & Lavrakas, P. J. (2010). AAPOR Report on Online Panels. *Public Opinion Quarterly*, *74*(4), 711–781. https://doi.org/10.1093/poq/nfq048

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-administered Surveys. *American Journal of Political Science*, *58*(3), 739–753. https://doi.org/10.1111/ajps.12081 [dataset] Blom, A. G., Fikel, M., Friedel, S., Höhne, J. K., Krieger, U., Rettig, T., Wenz, A., & SFB 884 ´Political Economy of Reforms´ Universität Mannheim. (2019). *German Internet Panel, Wave 38 (November 2018)*. GESIS Data Archive, Cologne. ZA6958 Data file Version 1.0.0. https://doi.org/10.4232/1.13391

Blom, A. G., Gathmann, C., & Krieger, U. (2015). Setting Up an Online Panel Representative of the General Population: The German Internet Panel. *Field Methods*, *27*(4), 391–408. https://doi.org/10.1177/1525822X15574494

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J. W., Krieger, U., & Bossert, D. (2017). Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence From the German Internet Panel. *Social Science Computer Review*, *35*(4), 498–520. https://doi.org/10.1177/0894439316651584

Brüggen, E., van den Brakel, J., & Krosnick, J. A. (2016). Establishing the Accuracy of Online Panels for Survey Research. *Statistics Netherlands*, *11*(04), 43.

Chandler, J., Sisso, I., & Shapiro, D. (2020). Participant Carelessness and Fraud: Consequences for Clinical Research and Potential Solutions. *Journal of Abnormal Psychology*, *129*(1), 49–55. https://doi.org/10.1037/abn0000479

Clifford, S., & Jerit, J. (2015). Do Attempts to Improve Respondent Attention Increase Social Desirability Bias? *Public Opinion Quarterly*, *79*(3), 790–802. https://doi.org/10.1093/poq/nfv027

Cornesse, C., & Blom, A. G. (2020). Response Quality in Nonprobability and Probability-based Online Panels. *Sociological Methods and Research*. https://doi.org/10.1177/0049124120914940

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. *Journal of Survey Statistics and Methodology*, *8*(1), 4–36. https://doi.org/10.1093/jssam/smz041

Curran, P. G. (2016). Methods for the Detection of Carelessly Invalid Responses in Survey Data. *Journal of Experimental Social Psychology*, *66*, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006

Curran, P. G., & Hauser, K. A. (2019). I'm Paid Biweekly, Just Not by Leprechauns: Evaluating Valid-but-incorrect Response Rates to Attention Check Items. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

DiSogra, C., & Callegaro, M. (2009). Computing Response Rates for Probability-Based Web Panels. *Section on Survey Research Methods – JSM 2009*, 5309–5320.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. https://doi.org/10.1002/bdm.1753

Gummer, T., Roßmann, J., & Silber, H. (2021). Using Instructed Response Items as Attention Checks in Web Surveys: Properties and Implementation. *Sociological Methods and Research*, *50*(1), 238–264. https://doi.org/10.1177/0049124118769083

Hauser, D. J., & Schwarz, N. (2015). It's a Trap! Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open*, *5*(2). https://doi.org/10.1177/2158244015584617

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks than Do Subject Pool Participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hauser, D. J., Sunderrajan, A., Natarajan, M., & Schwarz, N. (2016). Prior Exposure to Instructional Manipulation Checks does not Attenuate Survey Context Effects Driven by Satisficing or Gricean Norms. *Methods, Data, Analyses*, *10*(2), 195–220. https://doi.org/10.12758/mda.2016.008

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and Deterring Insufficient Effort Responding to Surveys. *Journal of Business and Psychology*, *27*(1), 99–114. https://doi.org/10.1007/s10869-011-9231-8

Johnson, J. A. (2005). Ascertaining the Validity of Individual Protocols from Web-based Personality Inventories. *Journal of Research in Personality*, *39*(1 SPEC. ISS.), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009

Kane, J. V., & Barabas, J. (2019). No Harm in Checking: Using Factual Manipulation Checks to Assess Attentiveness in Experiments. *American Journal of Political Science*, *63*(1), 234–249. https://doi.org/10.1111/ajps.12396

Kane, J. V., Velez, Y. R., & Barabas, J. (2023). Analyze the Attentive and Bypass Bias: Mock Vignette Checks in Survey Experiments. *Political Science Research and Methods*, *11*(2), 293–310. https://doi.org/10.1017/psrm.2023.3

Keusch, F., & Yan, T. (2018). Is Satisficing Responsible for Response Order Effects in Rating Scale Questions? *Survey Research Methods*, *12*(3), 259–270. https://doi.org/10.18148/srm/2018.v12i3.7263

Krosnick, J. A. (1991). Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Liu, M., & Wronski, L. (2018). Trap Questions in Online Surveys: Results from Three Web Survey Experiments. *International Journal of Market Research*, *60*(1), 32–49. https://doi.org/10.1177/1470785317744856

Lugtig, P., & Toepoel, V. (2016). The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey: Effects on Survey Measurement Error. *Social Science Computer Review*, *34*(1), 78–94. https://doi.org/10.1177/0894439315574248

Mancosu, M., Ladini, R., & Vezzoni, C. (2019). 'Short is Better'. Evaluating the Attentiveness of Online Respondents Through Screener Questions in a Real Survey Environment. *BMS Bulletin of Sociological Methodology/ Bulletin de Methodologie Sociologique*, *141*(1), 30–45. https://doi.org/10.1177/0759106318812788

Maniaci, M. R., & Rogge, R. D. (2014). Caring about Carelessness: Participant Inattention and Its Effects on Research. *Journal of Research in Personality*, *48*(1), 61–83. https://doi.org/10.1016/j.jrp.2013.09.008

Meade, A. W., & Craig, S. B. (2012). Identifying Careless Responses in Survey Data. *Psychological Methods*, *17*(3), 437–455. https://doi.org/10.1037/a0028085

Meisters, J., Hoffmann, A., & Musch, J. (2020). Can Detailed Instructions and Comprehension Checks Increase the Validity of Crosswise Model Estimates? *PLOS ONE*, *15*(6), e0235403. https://doi.org/10.1371/journal.pone.0235403

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting Careless Respondents in Web-based Questionnaires: Which Method to Use? *Journal of Research in Personality*, *63*, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology*, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Pasek, J., & Krosnick, J. A. (2020). Relations Between Variables and Trends over Time in RDD Telephone and Nonprobability Sample Internet Surveys. *Journal of Survey Statistics and Methodology*, *8*(1), 37–61. https://doi.org/10.1093/jssam/smz059

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P., & Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, *35*(3), 653–681. https://doi.org/10.2478/jos-2019-0027

Schonlau, M., & Toepoel, V. (2015). Straightlining in Web Survey Panels over Time. *Survey Research Methods*, *9*(2), 125–137. https://doi.org/10.18148/srm/2015.v9i2.6128

Shamon, H., & Berning, C. (2020). Attention Check Items and Instructions in Online Surveys with Incentivized and Non-Incentivized Samples: Boon or Bane for Data Quality? *Survey Research Methods*, *14*(1), 55–77. https://doi.org/10.18148/srm/2020.v14i1.7374

Silber, H., Roßmann, J., & Gummer, T. (2022). The Issue of Noncompliance in Attention Check Questions: False Positives in Instructed Response Items. *Field Methods*, *34*(4), 346–360. https://doi.org/10.1177/1525822X221115830

Strack, F., & Martin, L. L. (1987). Thinking, Judging, and Communicating: A Process Account of Context Effects in Attitude Surveys. In H.-J. Hippler, N. Schwarz, & S. Sudman (Eds.), *Social Information Processing and Survey Methodology. Recent Research in Psychology.* (pp. 123–148). Springer.

Struminskaya, B., Weyandt, K., & Bosnjak, M. (2015). The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-based General Population Panel. *Methods, Data, Analyses*, *9*(2), 261–292. https://doi.org/10.4232/1.12245.

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An Assessment of Exclusion Methods and Interactive Experiments. *Computers in Human Behavior*, *77*, 184–197. https://doi.org/10.1016/j.chb.2017.08.038

Thomas, R. K. (2014). Fast and Furious... or Much Ado About Nothing? Sub-Optimal Respondent Behavior and Data Quality. *Journal of Advertising Research*, *54*(1), 17–31. https://doi.org/10.2501/jar-54-1-017-031

Toepoel, V., Das, M., & van Soest, A. (2008). Effects of Design in Web Surveys: Comparing Trained and Fresh Respondents. *Public Opinion Quarterly*, *72*(5), 985–1007. https://doi.org/10.1093/poq/nfn060

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The Psychology of Survey Response*. Cambridge University Press.

Wiśniowski, A., Sakshaug, J. W., Perez Ruiz, D. A., & Blom, A. G. (2020). Integrating Probability and Nonprobability Samples for Survey Inference. *Journal of Survey Statistics and Methodology*, *8*(1), 120–147. https://doi.org/10.1093/jssam/smz051

Zhang, C., & Conrad, F. G. (2014). Speeding in Web Surveys: The Tendency To Answer Very Fast and Its Association With Straightlining. *Survey Research Methods*, *8*(2), 127–135.

Appendix

Table A1. Coding scheme for categorized sociodemographic variables.

Variable	Recoded categories	Original categories
Age		(year of birth categories)
	over 68 years	1935–1939
		1940–1944
		1945–1949
	59 to 68 years	1950–1954
		1955–1959
	49 to 58 years	1960–1964
		1965–1969
	39 to 48 years	1970–1974
		1975–1979
	29 to 38 years	1980–1984
		1985–1989
	under 29 years	1990–1994
		1995–1999
		2000 and later
Education		(highest school & professional degree)
	Low	No degree yet (still student)
		Left school with no degree
		Volks- / Hauptschule (or equivalent)
	Medium-low	Mittlere Reife / Realschule (or equivalent)
	Medium-high	Fachhochschulreife
		Abitur (or equivalent)
	High	Bachelor's degree
		Diploma / Master's (vocational university)
		Diploma / Master's (university)
		Ph.D.
Occupation	Full-time work	Full-time employed
	Part-time work	Part-time employed
		Partial retirement
		Mini job
		One-euro-job
		Irregular employment
	Unemployed / not in labor force	Unemployed
		Parental leave
		Permanently unable to work
		Housewife / househusband
	Retired	Retired / pensioner
	Education / voluntary service	Vocational training
		Occupational retraining
		Federal volunteer service (military or civil)
		Gap year
		Student (school)
		Student (university / vocational university)

Table A2. Distributions of variables of interest across experimental groups and χ^2-tests for differences.

	Overall	Experimental group				χ^2-test		
		Instruction	1 paragraph	2 paragraphs	4 paragraphs	χ^2	df	p
Age						29.702	15	.013
<29 years	535	123 (23.0%)	157 (29.3%)	126 (23.6%)	129 (24.1%)			
29–38 years	613	138 (22.5%)	165 (26.9%)	161 (26.3%)	149 (24.3%)			
39–48 years	651	169 (26.0%)	162 (24.9%)	156 (24.0%)	164 (25.2%)			
49–58 years	981	246 (25.1%)	225 (22.9%)	267 (27.2%)	243 (24.8%)			
59–68 years	822	230 (28.0%)	219 (26.6%)	172 (20.9%)	201 (24.5%)			
>68 years	424	112 (26.4%)	81 (19.1%)	117 (27.6%)	114 (26.9%)			
Gender						4.059	3	.255
Female	1,941	476 (24.5%)	510 (26.3%)	467 (24.1%)	488 (25.1%)			
Male	2,085	542 (26.0%)	499 (23.9%)	532 (25.5%)	512 (24.6%)			
Education						12.061	9	.210
Low	623	156 (25.0%)	155 (24.9%)	168 (27.0%)	144 (23.1%)			
Medium-low	1,165	320 (27.5%)	285 (24.5%)	257 (22.1%)	303 (26.0%)			
Medium-high	900	211 (23.4%)	242 (26.9%)	227 (25.2%)	220 (24.4%)			
High	1,338	331 (24.7%)	327 (24.4%)	347 (25.9%)	333 (24.9%)			
Occupation						9.270	12	.680
Full-time	1,871	455 (24.3%)	477 (25.5%)	467 (25.0%)	472 (25.2%)			
Part-time	699	189 (27.0%)	179 (25.6%)	168 (24.0%)	163 (23.3%)			
Unemployed	358	93 (26.0%)	77 (21.5%)	91 (25.4%)	97 (27.1%)			
Retired	788	205 (26.0%)	186 (23.6%)	203 (25.8%)	194 (24.6%)			
Education/ service	310	76 (24.5%)	90 (29.0%)	70 (22.6%)	74 (23.9%)			
Recruitment						.512	6	.998
2012	676	172 (25.4%)	172 (25.4%)	169 (25.0%)	163 (24.1%)			
2014	1,583	401 (25.3%)	391 (24.7%)	390 (24.6%)	401 (25.3%)			
2018	1,767	445 (25.2%)	446 (25.2%)	440 (24.9%)	436 (24.7%)			
Device						5.483	3	.140
Smartphone	872	223 (25.6%)	238 (27.3%)	193 (22.1%)	218 (25.0%)			
Computer	3,154	795 (25.2%)	771 (24.4%)	806 (25.6%)	782 (24.8%)			
Field time						.737	12	1.000
Day 1	293	73 (24.9%)	76 (25.9%)	71 (24.2%)	73 (24.9%)			
Days 2–3	1,102	273 (24.8%)	278 (25.2%)	276 (25.0%)	275 (25.0%)			
Days 4–7	855	221 (25.9%)	214 (25.0%)	207 (24.2%)	213 (24.9%)			
Days 8–14	828	211 (25.5%)	208 (25.1%)	206 (24.9%)	203 (24.5%)			
After day 14	948	240 (25.3%)	233 (24.6%)	239 (25.2%)	236 (24.9%)			
Daytime						4.435	9	.881
Morning	1,029	277 (26.9%)	256 (24.9%)	245 (23.8%)	251 (24.4%)			
Afternoon	1,559	382 (24.5%)	405 (26.0%)	382 (24.5%)	390 (25.0%)			
Evening	1,143	285 (24.9%)	281 (24.6%)	291 (25.5%)	286 (25.0%)			
Night	295	74 (25.1%)	67 (22.7%)	81 (27.5%)	73 (24.8%)			

Table A3. Model of passing the attention check with only text length and smartphone use as predictors.

	OR	SE	p
Text length (ref.: Instruction only)			
One paragraph	.478	.049	.000
Two paragraphs	.353	.035	.000
Four paragraphs	.181	.018	.000
Smartphone respondent	1.323	.110	.001
Constant	3.633	.288	.000
Observations	4,026		
Pseudo-$R^2_{\text{McKelvey \& Zavoina}}$	.107		

Note. OR = odds ratios, SE = standard errors.

Table A4. Logistic regression model of passing the attention check with interaction terms.

	OR	SE	p	AME	SE	p
Text length (ref.: Instruction only)						
One paragraph	.573	.244	.191	- .145	.019	.000
Two paragraphs	.818	.341	.631	- .216	.020	.000
Four paragraphs	.192	.086	.000	- .386	.019	.000
Age (ref.: <29 years)						
29–38 years	3.177	1.088	.001	.126	.032	.000
39–48 years	2.008	.607	.021	.136	.032	.000
49–58 years	3.776	1.149	.000	.204	.031	.000
59–68 years	2.317	.699	.005	.191	.035	.000
>68 years	3.620	1.350	.001	.210	.043	.000
Female	1.236	.204	.201	.071	.016	.000
Education (ref.: low)						
Medium-low	1.504	.344	.074	.055	.023	.019
Medium-high	2.652	.718	.000	.128	.025	.000
High	2.512	.592	.000	.156	.023	.000
Occupation (ref.: full-time work)						
Part-time work	1.200	.128	.088	.037	.022	.086
Unemployed	1.174	.160	.240	.033	.028	.236
Retired	.985	.136	.912	- .003	.029	.912
Education / voluntary service	1.528	.260	.013	.085	.033	.009
New recruitment sample	1.629	.277	.004	.101	.015	.000
Smartphone respondent	2.032	.474	.002	.089	.019	.000
Field time (ref.: day 1)						
Days 2–3	1.161	.184	.345	.031	.033	.347
Days 4–7	1.409	.227	.034	.071	.034	.035
Days 8–14	1.402	.226	.036	.070	.034	.038
After day 14	1.128	.178	.443	.025	.033	.445
Daytime (ref.: morning)						
Afternoon	1.256	.112	.011	.047	.019	.011
Evening	1.274	.128	.016	.050	.021	.016
Night	1.380	.208	.032	.066	.031	.030

Table A4 (continued).

Age * text length			
29–38 years * one paragraph	.645	.269	.292
29–38 years * two paragraphs	.355	.148	.013
29–38 years * four paragraphs	.525	.223	.129
39–48 years * one paragraph	1.252	.482	.560
39–48 years* two paragraphs	.949	.365	.891
39–48 years * four paragraphs	.642	.253	.261
49–58 years * one paragraph	.731	.280	.413
49–58 years* two paragraphs	.535	.200	.094
49–58 years * four paragraphs	.672	.257	.300
59–68 years * one paragraph	1.186	.446	.651
59–68 years* two paragraphs	.873	.332	.722
59–68 years * four paragraphs	1.183	.457	.664
>68 years * one paragraph	.626	.291	.313
>68 years* two paragraphs	.532	.235	.154
>68 years * four paragraphs	.989	.448	.981
Female * text length			
One paragraph	1.181	.254	.440
Two paragraphs	1.036	.219	.868
Four paragraphs	1.327	.280	.180
Education * text length			
Medium-low * one paragraph	.708	.219	.264
Medium-low * two paragraphs	.715	.221	.277
Medium-low * four paragraphs	1.132	.365	.700
Medium-high * one paragraph	.646	.227	.214
Medium-high * two paragraphs	.467	.162	.028
Medium-high * four paragraphs	.873	.318	.710
High * one paragraph	1.008	.323	.979
High * two paragraphs	.532	.164	.041
High * four paragraphs	1.018	.331	.957
New recruitment sample * text length			
One paragraph	1.335	.298	.196
Two paragraphs	.942	.203	.780
Four paragraphs	.830	.180	.392
Smartphone respondent * text length			
One paragraph	.564	.167	.053
Two paragraphs	.841	.250	.561
Four paragraphs	.785	.229	.408
Constant	.356	.130	.005
Observations	4,026		
Pseudo-$R^2_{\text{McKelvey \& Zavoina}}$	.190		
AIC	4,923		
BIC	5,295		

Note. OR = odds ratios, SE = standard errors, AME = average marginal effects.

Table A5. χ^2-tests for differences in sample composition between attentive and inattentive respondents, and IMC passed vs. failed among those flagged and not flagged for fast response times.

	χ^2	df	p
IMC: Remaining sample vs. excluded			
Age	14.220	5	.014
Female	28.603	1	.000
Education	38.636	3	.000
Occupation	16.327	4	.003
New recruitment sample	41.966	1	.000
Smartphone respondent	11.242	1	.001
Response time: Remaining sample vs. excluded			
Age	34.885	5	.000
Female	5.376	1	.000
Education	1.465	3	.690
Occupation	33.416	4	.000
New recruitment sample	51.999	1	.000
Smartphone respondent	7.435	1	.006
Response time flagged: IMC pass vs. fail			
Age	11.064	5	.050
Female	21.364	1	.000
Education	47.512	3	.000
Occupation	31.006	4	.000
New recruitment sample	.188	1	.665
Smartphone respondent	2.639	1	.104
Response time not flagged: IMC pass vs. fail			
Age	10.232	5	.069
Female	8.053	1	.005
Education	44.234	3	.000
Occupation	10.007	4	.040
New recruitment sample	5.123	1	.024
Smartphone respondent	1.427	1	.232

Table A6. Distribution of respondents' characteristics in remaining samples following exclusion of intattentives by text length.

	IMC				Response time			
	Instruction	1 paragraph	2 paragraphs	4 paragraphs	Instruction	1 paragraph	2 paragraphs	4 paragraphs
Age								[b]
<29 years	11.0%	13.6%	11.7%	10.9%	10.8%	12.3%	11.0%	10.5%
29–38 years	14.9%	16.8%	14.6%	15.1%	13.6%	15.1%	14.6%	16.8%
39–48 years	16.2%	16.9%	16.9%	12.9%	16.4%	16.1%	16.6%	9.6%
49–58 years	25.5%	23.1%	27.9%	26.8%	24.5%	23.9%	27.4%	25.8%
59–68 years	21.3%	22.3%	17.8%	21.2%	23.1%	23.8%	18.7%	20.2%
>68 years	11.1%	7.3%	11.1%	13.1%	11.6%	8.8%	11.7%	17.1%
Female	47.9%	54.0%	50.3%	56.7% [a]	45.9%	51.8%	49.3%	57.1% [b]
Education								
Low	13.1%	13.1%	15.7%	10.2%	15.6%	14.8%	16.8%	11.8%
Medium-low	30.7%	25.2%	25.8%	28.2%	32.1%	27.9%	26.4%	31.4%
Medium-high	22.0%	24.7%	23.2%	23.9%	20.4%	23.6%	22.1%	23.9%
High	34.2%	37.0%	35.3%	37.7%	31.9%	33.7%	34.7%	32.9%
Occupation								
Full-time work	45.2%	47.2%	43.7%	40.9%	45.0%	44.7%	43.0%	36.0%
Part-time work	18.4%	18.6%	20.0%	19.5%	17.6%	18.3%	17.3%	19.9%
Unemployed	9.0%	7.8%	9.9%	11.7%	8.8%	8.4%	11.8%	11.8%
Retired	19.6%	18.2%	19.4%	20.4%	21.2%	21.1%	21.4%	24.8%
Education/service	7.8%	8.2%	7.0%	7.5%	7.4%	7.5%	6.5%	7.5%
New recruitment sample	45.9%	49.8%	48.6%	48.2%	46.4%	48.5%	51.8%	50.6%
Smartphone respondent	23.6%	23.7%	22.1%	24.3%	23.6%	23.1%	22.2%	23.9%

Note. See Appendix Table A7 for χ^2-test statistics of differences in sample composition.

[a]significant differences between text lengths (exclusion based on IMC)

[b]significant differences between text lengths (exclusion based on response time)

Table A7. χ^2-tests for differences in sample composition between attentive respondents for different text lengths.

	IMC			Response time		
	χ^2	df	p	χ^2	df	p
Differences across text lengths (overall)						
Age	22.801	15	.088	29.072	15	.016
Female	10.633	3	.014	13.228	3	.004
Education	12.830	9	.170	11.154	9	.265
Occupation	8.755	12	.724	12.967	12	.371
New recruitment sample	2.317	3	.509	4.399	3	.222
Smartphone respondent	.770	3	.857	.487	3	.922
Instruction vs. other groups						
Age	1.672	5	.892	3.335	5	.649
Female	6.601	1	.010	8.337	1	.004
Education	5.796	3	.122	5.683	3	.128
Occupation	.499	4	.974	2.695	4	.610
New recruitment sample	2.014	1	.156	3.132	1	.077
Smartphone respondent	.036	1	.849	.130	1	.719
1 paragraph vs. other groups						
Age	15.712	5	.008	8.811	5	.117
Female	2.154	1	.142	1.360	1	.244
Education	3.069	3	.381	1.780	3	.619
Occupation	4.655	4	.325	2.152	4	.708
New recruitment sample	1.178	1	.278	.019	1	.891
Smartphone respondent	.033	1	.856	.004	1	.949
2 paragraphs vs. other groups						
Age	5.616	5	.345	5.025	5	.413
Female	.465	1	.495	.056	1	.813
Education	4.378	3	.223	4.012	3	.260
Occupation	1.296	4	.862	3.192	4	.526
New recruitment sample	.126	1	.722	2.549	1	.110
Smartphone respondent	.687	1	.407	.395	1	.530
4 paragraphs vs. other groups						
Age	7.059	5	.216	19.260	5	.002
Female	5.140	1	.023	8.104	1	.004
Education	4.024	3	.259	3.694	3	.296
Occupation	4.830	4	.305	8.386	4	.078
New recruitment sample	.009	1	.923	.527	1	.468
Smartphone respondent	.235	1	.628	.112	1	.738