Bulletin de Méthodologie Sociologique
2019, Vol. 141 30–45
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0759106318812788
journals.sagepub.com/home/bms
SAGE

'Short is Better'. Evaluating the Attentiveness of Online Respondents Through Screener Questions in a Real Survey Environment

Moreno Mancosu
Collegio Carlo Alberto, Turin, Italy

Riccardo Ladini and Cristiano Vezzoni
University of Milan, Milan, Italy

Résumé

'Court, c'est mieux'. Evaluer l'attention des répondants en ligne par des questions-filtre dans un environnement de sondage réel. Dans les enquêtes en ligne, l'évaluation des répondants est la plupart du temps absente : pour cette raison, l'emploi de questions de contrôle ou « screeners » s'est développé pour apprécier le niveau d'attention des répondants. Les *screeners* demandent aux répondants de suivre un certain nombre de consignes décrites dans un texte, lequel contient un nombre variable d'informations erronées. De précédents travaux se sont penchés sur des *designs* expérimentaux *ad hoc*, généralement composés de quelques questions administrées à de petits échantillons. Utilisant une expérimentation intégrée à l'enquête ITANES (*Italian National Election Study*, N=3000), nous montrons que de courts *screeners*, c'est-à-dire des questions comportant un faible nombre de fausses informations, devraient être privilégiés à des tests plus conséquents pour estimer le niveau d'attention des personnes interrogées. Nous montrons en outre que ces *screeners* n'ont aucun effet quant à une éventuelle activation de l'attention de ces dernières.

Abstract

In online surveys, the control of respondents is almost absent: for this reason, the use of screener questions or "screeners" has been suggested to evaluate respondent attention.

Corresponding Author:
Moreno Mancosu, Collegio Carlo Alberto, piazza Arbarello 8, Turin, Italy
Email: moreno.mancosu@carloalberto.org

Screeners ask respondents to follow a certain number of instructions described in a text that contains a varying amount of misleading information. Previous work focused on ad-hoc experimental designs composed of a few questions, generally administered to small samples. Using an experiment inserted into an Italian National Election Study survey (N=3,000), we show that short screeners – namely, questions with a reduced amount of misleading information – should be preferred to longer screeners in evaluating the attentiveness of respondents. We also show there is no effect of screener questions in activating respondent attention.

Mots clés
Screener, effort cognitif, expérimentation, enquêtes en ligne

Keywords
Screener, cognitive strain, survey experiment, online surveys

Introduction

Answering survey questions requires a cognitive engagement that respondents are not always able to maintain (Krosnick, 1991; Lenzner et al., 2010). Previous studies have stressed that respondents sometimes do not pay enough attention in answering survey questions, leading them to answer with an option that does not fully represent their position, or by adopting strategies that reduce the cognitive engagement required to answer correctly (and sincerely). For instance, individuals may answer with the first option that satisfies them or even select an answer randomly (Krosnick, 1991). These sub-optimal strategies contribute to increasing measurement error attributed to the respondent (Groves, 1989; Corbetta, 2003: 62). This type of bias is very likely to increase as long as researcher control of interview conditions decreases, as in the case of surveys done by CAWI (Computer Assisted Web Interviewing). Indeed, CAWI interviews are administered without the control of an interviewer, and their use is dramatically increasing because of their cost-effectiveness compared to other modes of interviewing – such as CATI (computer-assisted telephone interviewing) and CAPI (computer-assisted personal interviewing).

To estimate (and control for) this possible bias in online surveys, practitioners have proposed to use *Instructional Manipulation Checks* (IMC, Oppenheimer et al., 2009), also known as *screener questions*[1] (Berinsky et al., 2014). Screeners are methodological tools that "work by instructing subjects to demonstrate that they are paying attention by following a precise set of instructions when choosing a survey response option" (Berinski et al., 2014: 739). This survey tool, employed in psychology, sociology, and political science,[2] can be seen as a test aimed at distinguishing between "attentive" respondents and those who are not attentive enough to survey questions (Meade and Craig, 2012). Several studies have argued that this tool can even activate respondent attention during an online survey (Oppenheimer et al., 2009 – the rationale being that once the respondent has realized the questionnaire includes trick questions, he/she will be more attentive to the questions that follow).

Despite the growing diffusion of online surveys in social research, methodological work that analyzes the nature and efficacy of screeners in detecting/activating respondents is usually based on *ad hoc* experimental studies, mainly carried out on small convenience samples (Oppenheimer et al., 2009; Hauser and Schwarz, 2015; Liu and Wronski, 2018).

This study aims to increase the knowledge of the empirical impact of screeners under realistic survey conditions by employing data from an online multipurpose survey carried out on a large sample of Italian respondents (N=3,000). First, we descriptively analyze which individual characteristics are likely to increase/decrease the probability of correctly answering or "passing" a screener (namely, completing the task hidden in the question). Second, we question whether or not there is a relation between the difficulty of a screener and its capacity of identifying attentive respondents. Third, we test whether a screener can help increase the attention of a respondent. Both the second and the third questions are tackled by means of an experimental design in which the position and the length of the screener, which determine its complexity, are randomly defined.

Our results show that the educational level and the interest in the topic of the questionnaire increase the likelihood of passing the screener question. In addition, we find that short screeners are passed by a larger number of respondents and that the quality of answers of those individuals is substantially equal to that of those who passed longer and more complex screeners. Finally, we find no evidence whatsoever concerning the activation process that screeners might generate. This article could be of particular interest for researchers who rely on online surveys to collect data and intend to introduce non-conventional questions in CAWI surveys, such as vignettes (Atzmüller and Steiner, 2010; Mutz, 2011: 54-67). To avoid ambiguity in the understanding of this kind of question, it is usually necessary to provide explanations that are as precise and detailed as possible. However, this leads to long and more complex questions, which necessitate careful and accurate reading. Our evidence shows that the CAWI mode might not be suitable for this type of design.

Screeners and Respondents' Attentiveness in Online Surveys

A screener is a multiple-choice question in an online/self-administered survey. However, differing from a standard behavioral or attitudinal question, a screener asks the respondent to complete a task, which is complicated by the presence of misleading information. Generally speaking, a screener can be subdivided into four sections, as exemplified in Figure 1. The first section presents an introduction that provides information concerning the survey topic, but not relevant to the accomplishment of the task. The second section of the screener describes the task that the respondent must do to pass the screener. In the example, the task consists of selecting a combination of non-consistent answer categories which would be improbably chosen if one does not read (or reads shallowly) the instructions. The third part is the trap-question, aimed at diverting the respondent from passing the screener. The trap-question, indeed, is semantically connected to the available answer options, but passing the screener is independent of the content of the question. Finally, the fourth part is constituted by the answer categories.

Respondents pass the screener if they perform the task described in the second part and fail it otherwise. In the example presented in Figure 1 (which represents the screener employed in the experiment, see below), to pass the screener, the respondent must select

Part 1. Introduction	Previous research shows that the large majority of people who gather information online prefer a site or portal that they perceive as more trustworthy than others.		
Part 2. Task	In this case, however, we are interested to know whether people take the time they need to follow carefully instructions in interviews. To show that you have read this much, please ignore the question and select only the options “Local newspaper websites” and “I never consult websites” as your two answers, no matter of the websites you actually visit.		
Part 3. Trap question	When there is a breaking news story, which is the news website you would visit more frequently? (Maximum three answers)		
Part 4. Answer categories	□ La Repubblica	□ Il Giornale	□ La Stampa
	□ Corriere	□ Dagospia	□ Press association websites
	□ Huffington post	□ Il Fatto Quotidiano	□ Other
	□ Libero quotidiano	□ Local newspapers websites	□ I never consult websites

Figure 1. Base screener wording employed in the experiment (Source: CAWI ITANES UniMi 2015)

both the “Local newspapers” and “I never consult websites” options, a clearly contradictory combination from a semantic point of view. To pass the screener, the respondent is expected to carefully read the instructions of part 2. People who only read the introduction, the trap question or, even, the answer categories fail the test, being deceived by the apparent semantic consistency of the answer. The cognitive strain required to pass the test varies according to the amount of misleading text that must be read to correctly follow the instructions, and identify the presence/absence of the trap-question.

The main aim of the screener is to distinguish between the so-called “workers”, who read the questions carefully and are attentive in answering the survey, and “shirkers” (Berinsky et al., 2014), namely subjects who do not pay enough attention and answer shallowly to the survey questions[3]. The assumption here is that people who pay more attention to screeners are those who pay, in general, more attention to every question in a survey and answer consistently. According to this assumption, people who pass the screeners can be defined as “attentive respondents”. Several authors suggest that screeners should be included in online surveys to allow researchers to exclude *ex-post* inattentive respondents (Goodman et al., 2013; Oppenheimer et al., 2009), or, at least, to stratify analyses by levels of respondents’ attentiveness (Berinsky et al., 2014).[4]

So far, the literature studying screeners has shown different gaps. Among these, we can identify the three most important. First of all, previous works analyzing screeners showed inconsistent evidence concerning the relationship between socio-demographic characteristics and the ability to pass a screener. On one hand, Berinsky and colleagues (2014) show that older, female and more educated respondents are more likely to pass screener questions. On the other, Anduiza and Galais (2017) find that the educational level might lead to peculiar results, with the higher and lower educated being less able to pass the screener compared to the medium educated. Thus, one can see knowledge of screeners' empirical functioning is still embryonic and the potential of the instrument to improve quality of survey response is far from clear. The first aim of the article is thus to bring additional descriptive evidence to assess which types of respondents are more likely to pass a screener.

In addition, a substantial number of screeners proposed in the literature present a very complex wording and a length that exceeds several times the average length of a survey question. This choice largely reflects the aim for which screeners have been proposed: that is, distinguishing attentive respondents from inattentive and shallow ones. In some situations, however, choosing screeners characterized by excessive cognitive strain might turn out to be a questionable choice. If on one hand, hard screeners can undoubtedly identify more attentive respondents, on the other, they risk identifying as shirkers respondents who are actually *not so shallow* (and thus produce good quality answers), but fail the screener because the task is too complicated.[5] To better explain the problem, it seems appropriate to remember that a screener, as every question in a survey, needs the respondent to apply a certain amount of cognitive effort to be correctly understood. That effort varies according to the question's wording (Kahneman, 1973; Kool et al., 2010). Psycholinguistic literature (Tourangeau et al., 2000; Lenzner et al., 2010) shows that the length and the syntactic complexity of a question may lead to difficulties in understanding the question; this, in turn, can have a significant impact on data quality (Christian et al., 2007). Previous methodological literature has partially confirmed these results: by using a survey experiment conducted with a relatively small sample (a few hundred cases overall), Liu and Wronski (2018) demonstrate that the length of a screener is negatively associated with the possibility of passing it (see also Anduiza and Galais, 2017). Also, the authors show that people who are able to pass more difficult screeners present a quality of the responses which is not significantly different from those who pass the easier ones. Their final suggestion is thus that an optimal screener should be short, because it gives us a correct proportion of people who are reasonably attentive to survey questions, without minimizing that proportion with a task too difficult even for an attentive quota of respondents. Our article aims at analyzing this aspect by employing a survey inserted in a CAWI mass election survey, with stronger statistical power. Employing such a survey minimizes the possibilities of type-II error which, in this case, would fail to assess a significant difference in the quality of responses between people who are exposed to hard and easy screeners.

In addition to *post hoc* discrimination between respondents, according to their attentiveness, it has been suggested that screeners can be employed to increase answer quality. In a certain manner, screeners can "wake up" subjects by activating the attention of those who are answering the survey. The idea is that when a respondent realizes that some questions are actually traps, he/she will be more attentive to avoiding errors in subsequent questions. Oppenheimer et al. (2009) propose to insert screeners that do not allow respondents to continue the survey until they have successfully completed the test. In this way, it is guaranteed that the respondent has correctly understood the screener wording (Guess, 2015) and that he/she understands the need of answering subsequent questions with an increased level of attentiveness (a strategy that would be particularly useful for people who are initially shirkers). The idea that screeners can be employed as an activation tool also guided an *ad hoc* experiment carried out with Amazon Mechanical Turk, in which Hauser and Schwartz (2015) show that the introduction of a screener placed before a task enhances the likelihood of passing it. The effect of the exposure to a screener on the quality of subsequent questions, however, has shown a rather short latency. The instrument showed to be not very effective in maintaining respondents' attentiveness in answering questions that are not placed immediately after the screener (Berinsky et al., 2016). By means of an experiment which randomizes the position of a screener's position, our study aims to verify whether or not screeners can also act as a tool of activation (or intervention, Hauser and Schwartz, 2015) in the realistic context of an online survey. In other words, we hypothesize that respondents exposed to a screener will produce better quality answers with respect to those who have not been subjected to it and, in particular, that this holds for those who have passed the screener successfully.

Hypotheses

The first hypothesis concerns the relation between cognitive load and the likelihood of correctly accomplishing the task requested in the screener. As underlined above, the complexity of a screener should influence the cognitive load requested of the respondent and, thus, the likelihood of accomplishing the task (Anduiza and Galais, 2017; Liu and Wronski, 2018). Starting with this consideration, it is possible to present the hypothesis as follows:

> Hyp1. The higher the cognitive load (in terms of complexity and length of the screener's wording), the lower the success rate.

The main aim of a screener, however, is to identify respondents who pay attention to the questions of the survey and answer in a more accurate way. The second hypothesis, thus, focuses on the relation between the outcome of the screener and the quality of answers, and can be formulated as follows:

> Hyp2. Respondents who pass the screener produce answers of better quality with respect to those who do not pass it.

It has been pointed out (Hauser and Schwartz, 2015) that a screener can be considered an instrument of respondent activation by hypothesizing that the former it enhances respondent attentiveness after exposure to the instrument and, thus, he/she answers in a more accurate way the questions that follow. We can formulate the third hypothesis as follows:

> Hyp3. Exposure to a screener enhances the quality of answers to the questions that follow.

Finally, once shown that a screener is actually able to distinguish between respondents according to the quality of their answers, it is possible to find the best compromise between cognitive load and the capacity to discriminate between workers and shirkers. In the literature, very complex screeners have been proposed, implicitly assuming that this solution guarantees a more transparent identification of attentive respondents. However, more recent research suggests that short screeners can also correctly identify reasonably attentive respondents (Liu and Wronski, 2018). We can thus formulate the last hypothesis as follows:

> Hyp4. Respondents who pass the screener present different quality of answers, according to the complexity of the screener.

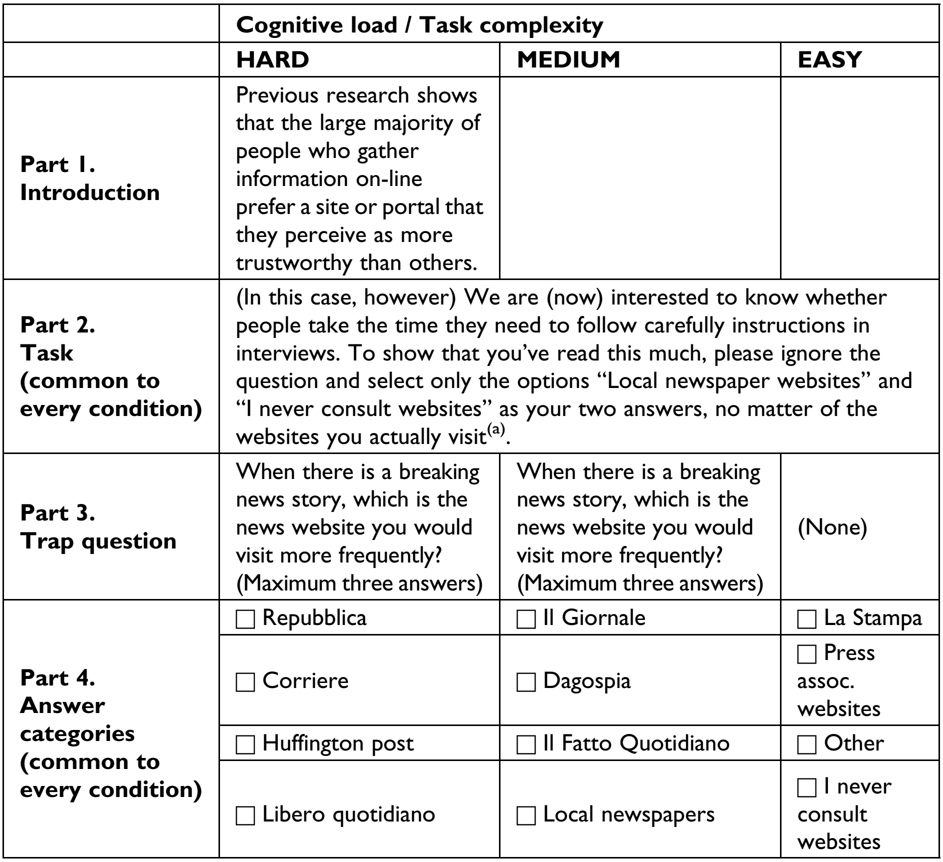
The Experiment

In a CAWI self-administered survey of 3,000 individuals (see below), a screener was inserted, based on the model of Berinsky and colleagues (2014: 740), focusing on the websites that one consults after learning of breaking news. The complete version of the screener is presented in Figure 1. Following Oppenheimer and colleagues (2009), the content of the screener does not vary much with respect to the topic of the survey, which is aimed at measuring Italians' political opinions and includes specific questions on media consumption. By means of a randomized procedure, the cognitive load – that is, the complexity of the task (hard, medium, easy) – and the position (whether before or after a battery of items) have been manipulated. Thanks to the randomization, we can reasonably argue that the differences detected among the various treatment groups will be independent of other omitted variables. Concerning the cognitive load, the three experimental conditions are presented in Figure 2.

Indications to pass the task (Figure 2, part 2) and answer categories (part 4 of the screener) are identical for each of the three groups. In the medium version, we erased the introductory misleading information (part 1), and in the easy version the trap question (part 3) has also been removed. Concerning the position of the screener in the survey, it has been randomly located before or after a battery of questions regarding attitudes toward democracy (see below), which are employed to evaluate the attentiveness of respondents. The experimental design and the sample size of the groups are shown in Table 1.

Data and Methods

Data come from the ITANES (Italian National Election Study) – University of Milan panel 2013-18. The multipurpose study contains repeated measurements on the same group of respondents during the most relevant elections of the electoral cycle 2013-18. Interviews were carried out by means of an online method (CAWI). Data analyzed here concern the sixth wave of the panel, which took place about a month after the Italian regional elections of 2015 and involved 3,000 respondents. The respondents were randomly selected from a starting sample of panel participants (N=8,723), originally drawn by quota sampling (according to gender, age, and educational level) from an opt-in community group of a private research company (SWG). The dataset contains respondent socio-demographic information (age, gender, educational level subdivided in "Primary", "Secondary", and "Tertiary"), and information about respondent socio-political behavior and attitudes, such as interest in politics (a 4-point scale from "Not at all interested" to "Very interested in politics").



	Cognitive load / Task complexity		
	HARD	**MEDIUM**	**EASY**
Part 1. Introduction	Previous research shows that the large majority of people who gather information on-line prefer a site or portal that they perceive as more trustworthy than others.		
Part 2. Task (common to every condition)	(In this case, however) We are (now) interested to know whether people take the time they need to follow carefully instructions in interviews. To show that you've read this much, please ignore the question and select only the options "Local newspaper websites" and "I never consult websites" as your two answers, no matter of the websites you actually visit$^{(a)}$.		
Part 3. Trap question	When there is a breaking news story, which is the news website you would visit more frequently? (Maximum three answers)	When there is a breaking news story, which is the news website you would visit more frequently? (Maximum three answers)	(None)
Part 4. Answer categories (common to every condition)	□ Repubblica	□ Il Giornale	□ La Stampa
	□ Corriere	□ Dagospia	□ Press assoc. websites
	□ Huffington post	□ Il Fatto Quotidiano	□ Other
	□ Libero quotidiano	□ Local newspapers	□ I never consult websites

Figure 2. Different experimental conditions (Source: CAWI ITANES UniMi 2015).
Note. In the easy and medium versions, the first sentence of the task was slightly modified to make it coherent. "In this case however, we are interested" becomes "We are now interested"

Table 1. Factorial design and experimental groups size (N = 3,000)

	Position	
Cognitive load	Before	After
Hard	515	510
Medium	502	492
Easy	479	502

In addition to variables concerning manipulated factors of the experiment, the outcome of the screener (positive when the task is correctly accomplished, negative otherwise) and the attentiveness of respondents are also considered.

1. Compromises in politics are really just selling out on one's principles
2. Parties are necessary to defend special interests of groups and social classes
3. Parties criticize one another, but they are actually all the same
4. Parties guarantee that people can participate to politics in Italy
5. Without parties there cannot be democracy
6. Politicians would help the country more if they would stop talking and just take action on important problems

Figure 3. Items on attitudes toward democracy

The attentiveness of respondents has been assessed by analyzing their answers on a battery of 6 items placed immediately before or after the screener. The items, partially from a battery on stealth democracy proposed by Hibbing and Theiss-Morse (2002, see also Vezzoni, 2014), measure attitudes toward democracy (see Figure 3). The respondent is asked for degree of agreement on every item using a scale from 0 (totally disagree) to 10 (totally agree).

The items' semantic polarity varies, as three items (1, 3 and 6) express a negative attitude toward democracy and the other three (item 2, 4 and 5) express a positive attitude. This choice is aimed at minimizing possible response set effects.

The first measure of answer quality is defined at the individual level, considering the so-called *straight-line response set*, which indicates whether or not one answered with the same category on every item of the battery (see Liu and Wronski, 2018). In the case of a battery in which the items have an inverted semantic polarity, a set of identical answers would very likely show that the respondent did not adequately consider the meaning of the questions. The measure is dichotomous: the variable is equal to 1 if respondents answer all the questions of the battery[6] in the same way and 0 otherwise.

The second measure of quality is defined at the aggregate level by calculating the internal consistency (Cronbach's Alpha) of the 6-item scale, adequately recoded so all items have the same semantic polarity[7]. Higher values of the coefficient indicate higher coherence among the answers of the battery, and thus the higher the value of the coefficient, the higher the attentiveness of the group of respondents on which the coefficient has been calculated. We compute Alpha's confidence intervals at the 95% level, obtained by means of a bootstrap procedure on original data (Padilla et al., 2012)[8].

Results

Descriptive Results

As stressed above, our first task is to assess descriptively which respondents are more likely to pass a screener. Table 2 shows coefficients of a logistic regression model in which the dependent variable is 1 when the respondent passes the trap-question and 0 otherwise, and the independent factors are age, gender, educational level and interest in politics.

As shown in the table, gender and age effects do not seem to cause any change in the likelihood of passing the screener. However, education and interest in politics increase the

Table 2. Logistic regression model to analyze screener passage rate

Independent variables	Coef.	S.E.
Age	-0.00	(0.00)
Gender (ref. Male)	0.07	(0.08)
Educational level: Secondary (ref. Primary)	0.26**	(0.11)
Educational level: Tertiary	0.35***	(0.13)
Interest in politics	0.23***	(0.05)
Constant	-1.38***	(0.22)
Log-likelihood	-1897.5	
N	2,949	

*** p<0.01, ** p<0.05, * p<0.1

Table 3. Success rate by cognitive load and position

	Position			
Cognitive load	Pre	Post	Success rates	N
Hard	21	22	22	1,025
Medium	36	33	35	994
Easy	48	51	50	981
Success rates	35	35	35	
N	1,496	1,504		3,000

individual probability of passing the trap-question. In particular, by calculating average marginal effects, higher educated respondents are 8 percentage points more likely to pass the screener than lower educated ones (for the medium educated, the positive difference is 6 percentage points), while very interested people present an approximate difference of 20 percentage points compared to not-at-all interested in politics. These results are only partially similar to those found in previous literature - for instance, Anduiza and Galais (2017) find that the high-educated tend to present the same passage rate as low-educated respondents. However, no significant differences are detected between men and women.

Hypotheses Testing

The second analysis focuses on the success rates of the screeners, testing the first hypothesis. Table 3 shows the percentage of respondents who pass the screener for each of the six experimental groups. Results show that the complexity of the task to be accomplished is proportional to the screener difficulty. In the easy version, the screener is passed by the 50% of subjects, while only 22% pass the screener in the hard form. The first hypothesis is thus supported by the evidence. This latter result is particularly relevant for our aims, since the most complex screener is largely similar to the ones proposed in the literature. If the aim of the instrument is to distinguish between workers and shirkers, it is clear that such a level of complexity makes the screener extremely selective, and only a small part of the sample (1 out of 5) is able to pass the test. Also, we

Table 4. Quality of the answers on the battery of attitudes toward democracy by screener outcome

Outcome	% Response Set[(a)]	N	Alpha I.C. 95%	N[(b)]
Positive	1.3	1,054	.72 - .77	978
Negative	13.2	1,946	.57 - .64	1,683

Notes: [(a)] Chi2(1) = 116.8; p < .01; [(b)] Results obtained with listwise deletion.

Table 5. Quality of answers to the attitudes toward democracy battery by screener position (whole sample and positive outcome results)

	% Response Set		Alpha I.C. 95%	
Position	Whole sample[(a)]	Only positive outcome[(b)]	Whole sample	Only positive outcome
Pre	9.2	1.7	.63 - .70	.69 - .77
Post	8.8	0.9	.63 - .70	.72 - .79
N	3,000	1,054	2,661[(c)]	978[(c)]

Notes: [(a)] Chi2(1) = 0.2; p = .67; [(b)] Chi2(1) = 1.2; p = .27; [(c)] Listwise deletion.

can underline another element of particular interest: success rates are not influenced by the position of the screener. This assures us that the battery on attitudes toward democracy does not influence the task described in the screener.

Concerning the ability of the instrument to distinguish between workers and shirkers, Table 4 shows that respondents who passed the screener present, according to our measures, higher quality answers. Among these respondents, the straight-line response set's prevalence is practically non-existent, even it involves 13% of respondents who did not pass the screener (the difference is significant, $p < .001$). The same result emerges for values of the Cronbach's alpha for the two groups. Regarding individuals who passed the screener, the lower bound of the confidence interval is higher than .70, a value considered the minimum for that measure in research where non-validated scales are employed (Peterson, 1994). The other respondents, identified by the screener as not attentive in responding to questions, present a confidence interval entirely located under this threshold. We can thus confirm our second hypothesis.

Results from the experimental manipulation of the screener lead us to reject our third hypothesis since no activation effect was detected. Table 5 shows clearly that the quality of the answers on attitudes towards democracy, both in terms of response set and alpha, does not vary according to the position of the screener. This outcome is strengthened by the absence of the relation even when analyzing only those respondents who passed the screener and consequently realized the presence of the trap-questions.

Finally, we should evaluate the relationship between cognitive load and attentiveness of the respondents who pass the test. Our expectation was that those who pass screeners present a different quality of answers, depending on the difficulty of the test. Results obtained by means of the manipulation of the cognitive load of the screener, and

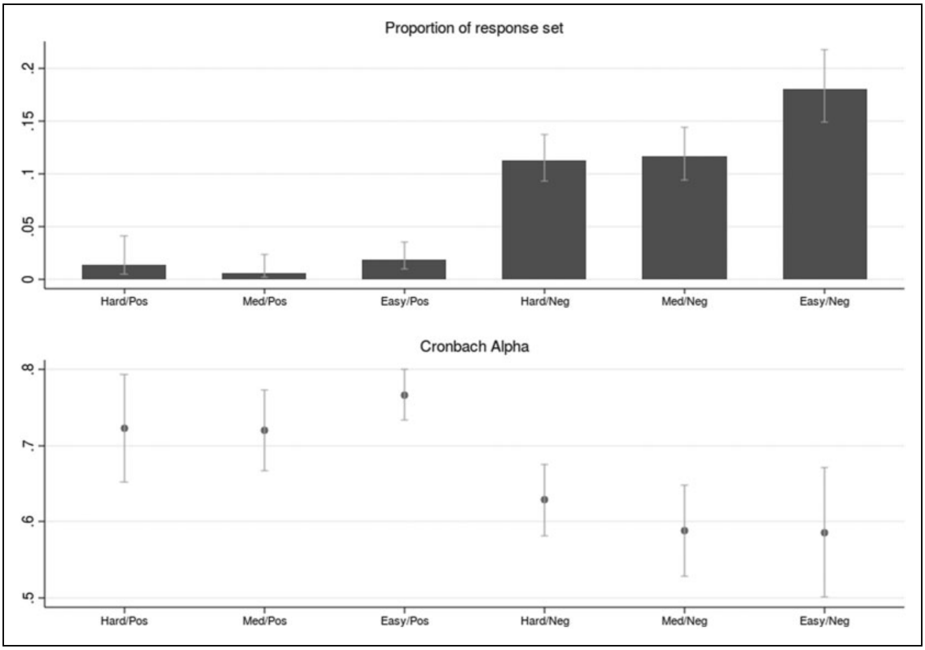


Figure 4. Confidence intervals at the 95% level of Cronbach's alpha (boostrapped) and straight-line response set of the battery on attitudes toward democracy, by cognitive load (Hard/Medium/Easy) and screener's outcome (Positive/Negative)

presented in Figure 4, go against this and lead us to argue that our hypothesis is not confirmed by the evidence. Regarding the straight-line response set (top panel of Figure 4), the attentiveness of respondents who pass the screener is the same, irrespective of the cognitive load. Similar results are given by the analysis of Cronbach's alpha. Once we distinguish between the respondents according to the outcome of the task, the overlap of confidence intervals (see Figure 4 – bottom panel) confirms the absence of a relationship between cognitive load and quality of the answers[9]. In other words, people who pass a difficult or an easy screener tend to be equally attentive to the battery, producing a roughly similar quality of answers – a finding consistent with evidence produced by Liu and Wronski (2018). This latter result is particularly relevant when we have to define and calibrate our instrument, since the increase in cognitive load does not seem to improve the discrimination in terms of the attentiveness of respondents. By employing too difficult screeners, on the contrary, we obtain a sub-optimal result: part of the attentive respondents fail the test, as seen in Table 3, and the number of people identified as attentive respondents decreases considerably[10].

Conclusion and Discussions

Online surveys represent one of the primary tools for collecting data in social and political research (Callegaro et al., 2015: 4), because they allow obtaining large datasets in a fast and cheap way (Loosveldt and Sonck, 2008: 96). This advantage has a cost: the absence of an interviewer in controlling the answering process leads to the risk of a reduction of the quality of the collected data. For this reason, it has been suggested to introduce survey instruments aimed at assessing respondents' attention when answering a survey. These instruments, known as screeners, are tests in which the respondent must complete a task by following the instructions hidden in the wording of a question, which in turn contains a variable amount of misleading information. The screener is also known as a "trap question" because misleading information is aimed at diverting respondents from their task.

The employment of screeners is increasingly broad (Berinsky et al., 2014), but knowledge concerning the empirical working of these instruments is still scarce and almost completely limited to experimental designs in *ad hoc* surveys (see Liu and Wronski, 2018). The experiment proposed in this article aims at investigating, in the context of a real survey, how a screener actually works and how to calibrate the cognitive load of the task according to its capacity to identify attentive respondents. The experimental design, thus, randomizes two factors: the cognitive load and the position of the screener.

The first result of the study is that only a limited number of respondents seem to read the wording of the question carefully. Generally speaking, less than half of the sample passes the screener successfully. We need to underline that this result is in line with previous research in the European context: Anduiza and Galais (2017: 508) present to a Spanish representative sample a screener not so different with respect to the medium screener here employed, and measure a 42% success rate. If we analyze the success rates of our screeners, we notice that the rate is of 50% for the easiest version and drops to 22% when the wording is more complicated and the trap more insidious. This result leads one to reflect on the quality of answers that people produce in online surveys, especially when respondents have to answer questions with more complex wording with respect to the rest of the survey. The result represents an alarm bell for the quality of non-conventional questions - such, for instance, vignette studies (Atzmüller and Steiner, 2010; Mutz, 2011: 54-67), which necessitate a careful and accurate reading because their long and complex wording varies in a systematic way. Answers to these questions can be subjected to significant error, given the limited attention of the respondent and could be prone to significant biases. Therefore, our evidence shows that the might not be suitable for this type of designs.

Also, we have been able to assess a difference in the socio-demographic characteristics of those who pass a screener. People with a higher level of education and more interested in the topic of the survey (which is focused on socio-political matters) are more likely to pass the screener. This suggests that merely erasing from our analysis inattentive respondents, as suggested in previous contributions (Goodman et al., 2013), might lead to severe selection issues. In the light of this results, our suggestion is thus to keep inattentive voters in the analysis, by taking into account that they produce lower quality answers - namely, by adding screeners' answers as control variables or including them as interactions (see Berinsky et al., 2016).

Our results are even more significant if we consider that passing the task is associated with the quality of answers that respondents give to other questions in the survey. This means that the screener is actually able to distinguish between workers and shirkers and that it can be thus employed if we want to maintain only better respondents in the analysis.

We have shown that effective screeners need a calibration of their cognitive load so not to be too selective. Regarding this aspect, the experiment presented in this article stressed that the quality of the answers of people who correctly passed the easy screener is substantially identical with respect to those who passed the medium, as well as the most complex one. To differentiate between our respondents, it is thus sufficient to introduce tasks characterized by a limited cognitive load, with the advantage of being able to identify a higher number of attentive respondents. This result is consistent with previous research focused on this aspect (Liu and Wronski, 2018), but opposite to general practice of research, where proposed screeners are generally very complex and require more cognitive strain compared to the other questions in the survey (see Goodman et al., 2013; Oppenheimer et al., 2009; Berinsky et al., 2014). Thus, our article suggests rethinking calibration and the empirical working of the screeners, which in the most common format are not effective tools. The employment of brief and relatively simple screener, less invasive and more readily applicable to different contexts, turns out to be the preferred choice. Starting from this work, more research on this topic will be needed to test further the calibration of the instrument, to get to an optimal wording able to distinguish efficiently among respondents.

Finally, results of the manipulation of the screener's position show that in the context of a real survey, the instrument does not act as a respondent activator. Indeed exposure to the screener does not affect the quality of the answers to questions that immediately follow it. This result seems to disprove previous studies, which argued that screeners are able to activate the attention of respondent in completing a task. This divergence with respect to the activation of attention by screeners could be due to the different conditions in which experiments have been conducted. In particular, Hauser and Schwartz (2015) refer to an *ad hoc* experimental study, with a relatively small sample (N < 400), with paid respondents and a brief questionnaire, while our study is carried out on a large sample, in real survey conditions. The question remains however open, and further work manipulating screener's position is necessary to confirm results of our experiment.

Acknowledgements

We would like to thank anonymous reviewers and Karl van Meter for their help, which greatly improved the quality of the manuscript.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Henceforth, we will use the label screener to define this kind of questions.
2. Berinsky and colleagues (2014) identify about 40 studies employing screeners between January 2006 and July 2013.
3. In the literature, this tendency is defined as satisficing (Krosnick, 1991; Oppenheimer et al., 2009).
4. Stratification is a less onerous procedure than filtering subjects who do not pass the screener. Excluding these subjects could indeed lead to external validity issues by unbalancing the composition of the sample with respect to several individual properties, such as sociodemographic characteristics, which are associated with the successful outcome of the test.
5. A complementary risk could be present. A too easy screener will not be able to operate an adequate distinction between workers and shirkers, leading to coding as "attentive" people who are not so attentive. If we consider the screener examples presented in the literature, however, usually characterized by a high level of complexity, this issue seems less relevant.
6. Value 1 is also attributed to respondents who have always answered "Don't know".
7. The measures have been calculated with listwise deletion. This led to a deletion of the 11% of the original sample.
8. Synthetically, bootstrap is a technique that allows inferring standard errors and distribution of an estimate in cases in which the underlying distribution is unknown. The technique is based on a set of *n* random resamples that produce a distribution of n parameters, by means of which it is thus possible to infer the estimates of interest. In our case, the bootstrapped estimate of confidence intervals was calculated through 400 resamples.
9. As stressed above, the scale is largely inspired by Hibbing and Theiss-Morse (2002). However, we have further investigated the reliability of the scale by analysing it with Cronbach's alpha with item deletion. By means of this additional analysis, it seems that items 1 and 6 lower the overall reliability of the scale. This, per se, does not represent a particular issue, since our main aim is to compare the internal consistency of the scale among different groups. However, one might argue that a more reliable scale might produce different patterns among those groups. For this reason, we have re-run all the analyses by computing the Cronbach's alpha of a scale based on items 2 to 5. Results (available under request) lead us to substantially similar conclusions to the ones presented in the article – as one would expect, the main difference concerns the general levels of the alpha, which are always higher than in the tables and figures here presented.
10. For what regards the response set, relatively easy screeners have the charming property of selecting better shirkers, as it is possible to see in Table 5. 18% of those who fail the test with the easy screener answered with a straight-line response set to the battery questions, while this happens only for the 12% who did not pass the medium and difficult tasks. This relation, however, is not confirmed by the Cronbach's alpha.

References

Anduiza E and Galais C (2017) Answering Without Reading: IMCs and Strong Satisficing in Online Surveys. *International Journal of Public Opinion Research* 29(3): 497-519.

Atzmüller C and Steiner PM (2010) Experimental Vignette Studies in Survey Research. *European Journal of Research Methods for the Behavioral and Social Sciences* 6(3): 128-138.

Berinsky AJ, Margolis MF and Sances MW (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.

Berinsky AJ, Margolis MF and Sances MW (2016) Can We Turn Shirkers into Workers? *Journal of Experimental Social Psychology* 66(1): 20-28.

Callegaro M, Lozar Manfreda K and Vehovar V (2015) *Web Survey Methodology*. London: Sage.

Christian LM, Dillman DA and Smyth JD (2007) Helping Respondents get it Right the First Time: the Influence of Words, Symbols, and Graphics in Web Surveys. *Public Opinion Quarterly* 71(1): 113-125.

Corbetta P (2003) Social Research: Theory, *Methods and Techniques*. Los Angeles: Sage.

Goodman JK, Cryder CE and 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.

Guess AM (2015) Measure for Measure: An Experimental Test of Online Political Media Exposure. *Political Analysis* 23(1): 59-75.

Groves RM (1989) *Survey Errors and Survey Costs*. New York: Wiley-Interscience.

Hauser DJ and Schwarz N (2015) It's a Trap!: Instructional Manipulation Checks Prompt Systematic Thinking on "Tricky" Tasks. *SAGE Open* 5: 1-6.

Hibbing JR and Theiss-Morse E (2002) *Stealth Democracy: Americans' Beliefs About How Government Should Work*. Cambridge: Cambridge University Press.

Kahneman D (1973) *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kool W, McGuire JT, Rosen ZB et al. (2010) Decision Making and the Avoidance of Cognitive Demand. *Journal of Experimental Psychology: General* 139(4): 665-682.

Krosnick JA (1991) Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys. *Applied cognitive psychology* 5(3): 213-236.

Lenzner T, Kaczmirek L and Lenzner A (2010) Cognitive Burden of Survey Questions and Response Times: A Psycholinguistic Experiment. *Applied Cognitive Psychology* 24(7): 1003-1020.

Liu M and Wronski L (2018). Trap Questions in Online Surveys: Results from Three Web Survey Experiments. *International Journal of Market Research* 60(1): 32-49.

Loosveldt G and Sonck N (2008) An Evaluation of the Weighting Procedures for an Online Access Panel Survey. *Survey Research Methods* 2(2): 93-105.

Meade AW and Craig SB (2012) Identifying Careless Responses in Survey Data. *Psychological Methods* 17(3): 437-455 .

Mutz DC (2011) *Population-Based Survey Experiments*. Princeton: Princeton University Press.

Oppenheimer DM, Meyvis T and Davidenko N (2009) Instructional Manipulation Checks: Detecting Satisficing to Increase Statistical Power. *Journal of Experimental Social Psychology* 45(4): 867-872.

Padilla MA, Diverse J and Newton M (2012) Coefficient Alpha Bootstrap Confidence Interval Under Non Normality. *Applied Psychological Measurement* 36(5): 331-348.

Peterson RA (1994) A Meta-Analysis of Cronbach's Coefficient Alpha. *Journal of Consumer Research* 21(2): 381-391.

Tourangeau R, Rips LJ and Rasinski K (2000) *The Psychology of Survey Response*. Cambridge: Cambridge University Press.

Vezzoni C (2014) Italian National Election Survey 2013: a Further Step in a Consolidating Tradition. *Rivista italiana di scienza politica* 44(1): 81-108.