Survey Research Methods (2020)
Vol. 14, No. 1, pp. 55-77
doi:10.18148/srm/2020.v14i1.7374

ISSN 1864-3361
http://www.surveymethods.org

Attention Check Items and Instructions in Online Surveys with Incentivized and Non-Incentivized Samples: Boon or Bane for Data Quality?

Hawal Shamon
Forschungszentrum Jülich
Germany

Carl Berning
Johannes Gutenberg Universität
Mainz, Germany

In this paper, we examine rates of careless responding and reactions to detection methods (i.e., attention check items and instructions) in an experimental setting based on two different samples. First, we use a quota sample (with monetary incentive), a central data source for internet-based surveys in sociological and political research. Second, we include a voluntary opt-in panel (without monetary incentive) well suited for conducting survey experiments (e.g., factorial surveys). Respondents' reactions to the detection items are analyzed by objective, nonreactive indicators (i.e., break-off, item nonresponse, and measurement quality), and two self-report scales. Our reaction analyses reveal that the detection methods we applied are not only well suited for identifying careless respondents, but also exert a motivational rather than a demotivating influence on respondents' answer behavior and, hence, contribute to data quality. Furthermore, we find that break-off behavior differs across both samples suggesting that results from methodological online studies on the basis of incentivized samples do not necessarily transfer to online studies in general.

Keywords: careless responding; inattentive responding; satisficing; detection method; measurement quality; data quality; online survey; true-score-models; multi-trait-multi-method (MTMM); justice; decent life; societal participation

1 Introduction

1.1 Motivation

In recent years, online surveys have become a popular survey mode in the social sciences, and they are even used in probability-based online panels such as the German GESIS Panel or the Dutch LISS Panel. Compared to the more traditional modes, such as paper-and-pencil, face-to-face, or telephone interviews, online surveys have the advantage of lower costs and short data collection periods. However, due to the physical distance and the concomitant anonymity during the measurement process (Johnson, 2005; Meade & Craig, 2012), data quality—in terms of measurement quality—of online surveys is impaired by respondents who "randomly" respond to a survey item (cf. Beach, 1989; Kurtz & Parrish, 2001), provide an answer to a survey item without considering the content of the item (cf. Meade & Craig, 2012, p. 438), or answer a survey measure with low or little motivation to correctly interpret the content of the item, comply with survey instructions, or provide accurate responses (Huang, Curran, Keeney, Poposki, & Deshon, 2012).

Researchers from the field of psychology, who administered their surveys almost exclusively to student samples participating in exchange for course credits or monetary incentives (cf. Niessen, Meijer, & Tendeiro, 2016), have already been concerned for some time about the consequences of these behaviors which impair measurement quality since they "could result in largely meaningless data, primarily adding error variance (i.e., noise) to analyses" (Maniaci & Rogge, 2014, p. 80). Therefore, they designed various detection methods to identify careless responses.

Detection-methods for careless responses can be categorized into ex ante methods and post hoc methods. Ex ante methods, which represent the focus of this study, require "special items or scales to be inserted into a survey prior to its administration" (cf. Meade & Craig, 2012, p. 439), such as bogus items (cf. Beach, 1989), explicitly instructed response items (cf. Huang et al., 2012; Kam & Meyer, 2015; Meade & Craig, 2012), item manipulation checks (cf. Oppenheimer, Meyvis, & Davidenko, 2009), and self-reported measures (cf. Maniaci & Rogge, 2014). Bogus items use the same answer scale as the substantial items of a survey. However, they have only one correct answer, such that respondents who conscientiously pass all four steps of the question-answer process (Tourangeau, Rips, & Rasinski, 2000)—i.e.,

Contact information: Hawal Shamon, Forschungszentrum Jülich Germany, Wilhelm-Johnen-Straße, 52425 Jülich (email: h.shamon@fz-juelich.de)

comprehension, retrieval, judgment, and response—should all provide the same response to that item.[1] Beach (1989) used for example "I was born on February 30th" as a bogus item and labeled it random response scale. Other researchers (cf. Huang et al., 2012; Maniaci & Rogge, 2014) prefer the term "infrequency scale" for such items. Explicitly instructed response items are items asking the respondents to choose a specific answer option from the answer scale. In doing so, respondents are not required to pass steps two and three of the question-answer process (Tourangeau et al., 2000). Hence, these items are designed to identify carelessly responding subjects who make no effort to attend to the survey question (step 1 of Tourangeau et al., 2000 question-answer process). In a similar vein, the item manipulation check consists of a question that resembles other questions in length and response format. It explicitly asks respondents in the instructions of the item to provide confirmation by ignoring the standard response format. Thus, the item manipulation check allows researchers to conclude whether respondents have paid attention to the item instructions or not (step 1 of Tourangeau et al., 2000 question-answer process). Last but not least, self-report measures have been used which assess respondents' tendency to respond carelessly by asking them on the final page of the survey to assess, for example, their engagement with the study or attention to the study (cf. Maniaci & Rogge, 2014; Meade & Craig, 2012). In contrast to ex ante detection methods, post hoc "methods do not require specialized items that are apparent to respondents but instead involve special analyses after data collection is complete" (cf. Meade & Craig, 2012, p. 440) such as methods examining response consistency, (multivariate) outlier, speeding or straightlining (Conrad, Tourangeau, Couper, & Zhang, 2017; Greszki, Meyer, & Schoen, 2014; Leiner, 2019; Zhang & Conrad, 2014).[2]

Compared to post hoc methods, ex ante detection methods (hereinafter also referred to as attention checks) allow survey designers to conclude with relatively high confidence whether or not respondents have carefully processed and responded to a survey question (e.g. Beach, 1989; Johnson, 2005; Kurtz & Parrish, 2001).[3] The extent to which studies in the field of psychology are affected by careless responses varies widely across the studies examining the issue. For instance, Johnson (2005) reports a rate of 3.5%, Kurtz and Parrish (2001) 10.6%, Kam and Meyer (2015) 16.8%, and Oppenheimer et al. (2009) a rate of as much as 46%. The differences in these results may be due to sample specifics, i.e., organizational samples vs. student samples, or detection methods employed (cf. Kam & Meyer, 2015; Maniaci & Rogge, 2014).[4] Huang, Bowling, Liu, and Li (2015, p. 308) stressed that the survey context in which attention checks are utilized should be taken into account, and Maniaci and Rogge (2014, p. 82) called on studies to investigate inattention "in more diverse samples to ensure that the effects of inattention remain relatively consistent across demographic groups".

In recent years, the issue of careless responding has been increasingly investigated in the field of social and political science where online surveys are usually administered to heterogeneous respondent samples gathered from voluntary opt-in panels of commercial survey institutes (e.g. Anduiza & Galais, 2016; Berinsky, Margolis, & Sances, 2014; Gummer et al., 2018; Hauser & Schwarz, 2015; Hauser, Sunderrajan, Natarajan, & Schwarz, 2017; Mancosu, Ladini, & Vezzoni, 2019; Miller & Baker-Prewitt, 2009), or probability-based online access panels recruited offline (Study 3 Gummer et al., 2018), which are attributed higher data quality than non-probability samples (e.g. Yeager et al., 2011). All these studies identified subjects engaged in careless responding (e.g., 8% in Study 1 by Hauser and Schwarz, 2015, 24% in Study 1 by Gummer et al., 2018, up to as much as 78% in Mancosu et al., 2019[5]) and, in doing so, demonstrate that careless responding is not only a problem for surveys administered to specific population samples (i.e., students), but also in the case of heterogeneous respondent samples that are used to examine research questions in the social and political sciences.

[1] Each of the four steps consists of several specific processes (cf. Tourangeau et al., 2000, p. 8). Step 1 (i.e., comprehension) includes among other things the requirement that respondents attend to questions and instructions, step 2 (i.e., retrieval) refers among other things to the retrieval of relevant information, step 3 (i.e., judgment) requires among other things that respondents integrate the retrieved information, and step 4 (i.e., response) stipulates among other things that respondents map the judgment onto the response category.

[2] Speeding is considered problematic for data quality, since conscientious answers are unlikely with only marginal response time (Gummer, Roßmann, & Silber, 2018).

[3] Attention checks flag respondents who do not pay sufficient attention to a survey item or instruction at a specific point (if only a single attention check is integrated into the questionnaire) or at several points (if multiple attention checks are integrated into the questionnaire) in the survey and, hence, indicate respondents who are selectively (if only one of multiple attention checks is failed) or globally (if all multiple attention checks are failed) unmotivated to participate appropriately in the survey.

[4] For a more detailed description of the response rates, (see Meade & Craig, 2012).

[5] Mancosu et al. (2019) applied screener questions. A screener question is a multiple-choice question where the survey instruction consists of three components (Mancosu et al., 2019): a) an introduction, b) a task that instructs respondents on how to answer the question correctly (e.g., choose option D and F to show that you have read this much), and c) a trap question that resembles a conventional question the survey designer is allegedly interested in. Hence, the item instruction of screener questions contains a relatively large number of words (cf. also the screener question of Berinsky et al., 2014).

Taken together, these findings from different disciplines which gather participants in different ways (i.e., student courses, voluntary opt-in samples from commercial survey institutes, and probability-based online access panels recruited offline) evoke the impression that careless responding is a general phenomenon of online surveys and, hence, a necessary consequence of the lack of interviewers to supervise the interview situation. This impression may become all the more established as monetary incentives do not seem to make a difference. Anduiza and Galais [2016] find evidence that respondents' (self-reported) motivation for participation due to paid material incentives in their longitudinal correlational study (six-wave panel) did not increase the likelihood of failing attention checks. These authors take this result as "additional support for the idea that material incentives are not a problem, nor (.) 'professional' respondents" (Anduiza & Galais, 2016, p. 514). However, all of the studies on careless responding mentioned above share the detail that they incentivized their participants.[6] Concluding that the prospect of remuneration for respondents' participation does not contribute to (higher rates of) careless responses in a study might be premature given the research designs to date and previous research findings on survey incentives.

Research on survey incentives has shown that incentives significantly increased response rates in online (e.g. Becker, Möser, & Glauser, 2019; Göritz, 2006) and in mail surveys (e.g. Church, 1993; Edwards, Cooper, Roberts, & Frost, 2005), another form of self-completion survey. Empirical evidence also points to the fact that incentives in web surveys increase response rates particularly of less-motivated respondents (Ernst Stähli & Joye, 2016) and that subjects who access online surveys for whatever reason are more likely to finish them when an incentive is offered than when not (Göritz, 2006). At the same time, previous studies on careless responding found that attention checks failures are promoted by situational factors affecting respondents' motivation to respond adequately, such as their lack of interest in the survey topic (e.g. Anduiza & Galais, 2016; Gummer et al., 2018; Maniaci & Rogge, 2014). These findings suggest that the anonymous interview situation is likely to be used to provide a careless response only by those who are not motivated to respond adequately at a specific point in the questionnaire.[7] Hence, given the findings of both research fields, attention checks might be particularly useful in questionnaires that are administered to target persons who will be remunerated for their survey participation.

However, despite their ability to identify respondents who do not pay sufficient attention to a survey item, attention checks might be more than mere measures of attention and act as interventions that affect how respondents approach subsequent survey items (Hauser & Schwarz, 2015). Some researchers fear that attention checks have a negative effect on respondents' answer behavior (e.g. Miller & Baker-Prewitt, 2009; Niessen et al., 2016), while other researchers expect positive effects on answers to subsequent survey items due to attention checks (e.g. Hauser & Schwarz, 2015; Huang et al., 2015). A number of empirical studies find either a positive (e.g. Hauser & Schwarz, 2015; Miller & Baker-Prewitt, 2009) or no effect (Berinsky et al., 2014; Gummer et al., 2018; Hauser et al., 2017; Mancosu et al., 2019) due to different attention checks on subsequent answer behavior. Hence, these findings challenge the concern that the application of attention checks affects subsequent answer behavior in a negative manner. However, at the same time, empirical evidence regarding positive effects of attention checks is ambiguous and requires further investigation.

1.2 Contribution

We aim to contribute to existing research by examining whether or not careless responding is a general problem of online surveys or rather a specific problem of incentivizing respondents. Taking the results of studies on survey incentives and careless responding into consideration, it is reasonable to assume that incentives promote the participation of less-motivated respondents and, hence, contribute to higher careless response rates in a survey. To the extent that this is true, attention checks are ceteris paribus particularly necessary in surveys where respondents are remunerated for survey participation, as is usually the case in opt-in panels of commercial panel operators which are used to build quota samples. Whether or not careless responding is a particular problem of samples in which respondents are incentivized for survey participation has not yet been examined in depth, to the best of our knowledge. The investigation of the role of incentives on careless responding helps to assess more effectively whether or not careless responses are a general problem of online surveys or rather of incentivized samples. Hence, this investigation also helps to better understand the implications of offering target persons (monetary) incentives.

At the same time, empirical clarification is still required on whether or not attention checks, which could be included in online surveys by default to account for careless responding, exert a positive or negative influence on subsequent answer behavior. The reported mixed evidence regarding the positive effects might be the consequence of operationalization strategies that are partly too rough to capture the positive effects of attention checks. Against this background, it is appropriate to (re)examine this issue by applying, among other indicators, an objective and nonreactive indicator, i.e., measurement quality. Measurement quality is equally effective in

[6] Participants in the probability-based online access panel recruited offline (study 3 by Gummer et al., 2018) were also incentivized (Rattinger et al., 2014).

[7] Careless responding is also distinct from social desirability (cf. Maniaci & Rogge, 2014, p. 65).

sensitively measuring both motivational and demotivational effects of attention checks.

1.3 Research design

To address all of these aspects, we conducted a split-ballot experiment in two studies simultaneously. In Study 1, participants were gathered from an access panel provided by a commercial service institute (commercial access panel) that recruits its panel members, among other things, by promising monetary incentives for survey participation to prospective respondents in the registration procedure.[8] In Study 2, participants were gathered from a non-commercial panel operator (non-commercial access panel) where in the registration procedure prospective respondents are asked to provide support and are offered insights into scientific research rather than monetary incentives.[9] Thus, both panels are distinct in an important detail: the self-selection process of panel members. It is reasonable to assume that panel members of the non-commercial access panel show on average a higher degree of intrinsic motivation to participate in (scientific) surveys (hereinafter referred to as general intrinsic motivation) than commercial access panel members, who might also participate for extrinsically motivated reasons, e.g., monetary incentives for survey participation, and, hence, can be assumed to participate for mixed reasons in the survey (hereinafter referred to as mixed motivation).[10]

In both studies, respondents were randomly assigned to different experimental settings (i.e., settings with or without attention checks). In the settings with ex ante detection methods, we applied two attention check items (i.e., an explicitly instructed response item and a careless response scale) and two attention check instructions. For the investigation of reactions to the attention check items and attention check instructions, we compared three objective and nonreactive indicators across the different settings, i.e., measurement quality scores that are suitable for capturing positive and negative effects, as well as item nonresponse and break-off rates, which are suitable for capturing negative rather than positive effects in split-ballot experiments on careless responding. Measurement quality was calculated based on reliability and validity coefficients obtained from a true score model (cf. Saris, 1990; Saris & Andrews, 1991; Saris et al., 2011) for three different concepts (i.e., a just earning level, earning level for a decent life, earning level for adequate societal participation). Each concept was measured by three different methods (i.e., open answer format, endpoint verbalized eleven-point scale, endpoint verbalized slider bar), reflecting a 3x3 Multi-Trait-Multi-Method design (MTMM) (cf. Campbell & Fiske, 1959).

2 Theoretical background

2.1 Careless responding

Careless responding refers to answering a survey item by ignoring the item content (cf. Meade & Craig, 2012, p. 438) or paying insufficient attention to the item content (cf. Kam & Meyer, 2015, p. 513).[11] Other authors prefer different labels for very similar concepts, such as random response (cf. Beach, 1989; Kurtz & Parrish, 2001), mental coin flipping (Converse, 1964; Krosnick, 1991), inattentive responding (Maniaci & Rogge, 2014) or insufficient effort responding (Huang et al., 2012), which is defined as "a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses" (p.100). All of these concepts have in common that respondents (perfidiously) provide a formally valid answer that does not necessarily reflect their true score on the measured attribute. Hence, these conceptualizations can be categorized as focusing primarily on the aspect of careless responses that undermines measurement quality. However, they differ regarding the steps in the question-answer process (Tourangeau et al., 2000) that are omitted and, hence, that constitute a careless response. While according to some definitions (e.g. Meade & Craig, 2012) a response is careless because respondents omit step one, i.e., comprehending the survey item by attending to the question and instruction, according to other definitions (e.g. Huang et al., 2012) a response is careless because respondents omit one or more steps in the question-answer process.

Another source of concept variation is introduced by authors who use the label of careless responding (cf. e.g. Niessen et al., 2016) by drawing upon Johnson (2005) definition of "random responding". Random responding is caused by carelessness and inattentiveness on the part of the respondent and is characterized by "leaving many answers blank, misreading items, answering in the wrong areas of the answer sheet, and/or using the same response category repeat-

[8] Panel members obtain points for survey participation that they can exchange for vouchers, cash or donations.

[9] In Germany, besides commercially operated access panels, non-commercial access panels exist that are accessible for scientific research only (e.g. SosciPanel, PsyWeb).

[10] The general intrinsic motivation to participate in scientific surveys is not to be confused with target persons' specific intrinsic motivation to participate in a particular survey that might arise, for example, due to their interest in a survey topic.

[11] In contrast to all the above-mentioned authors, Schmitt and Stuits (1985) define a careless respondent, rather than a response, as a subject, who "is not responding randomly. He/she is simply reading a few of the items in a measuring instrument, inferring what it is the items are asking of the respondent, and then responding in like manner to the remainder of the items in the instrument." (Schmitt & Stuits, 1985, p. 367).

edly without reading the item" (Johnson, 2005, pp. 104–105). That is to say, these authors also consider item nonresponse as a manifestation of careless responding, and, thus, address its representation-undermining aspect.[12]

We propose to reduce the conceptual heterogeneity regarding careless responses and similar, but not identical, concepts to a common denominator by distinguishing between perfidious careless responses and apparent careless responses. Both are seen as the consequence of omitting at least one of the steps in the question-answer process but they manifest themselves differently. Perfidious careless responses refer to providing a formally valid answer that harms measurement quality in the first instance because a respondent's reported score is likely to deviate from his or her true score. Apparent careless response refers to all answers by means of which a survey designer can conclude with a certain confidence that respondents have answered carelessly—e.g., leaving an item blank, straightlining in the case of reverse-keyed items, speeding—such that the survey designer has to or may decide to disregard these answers in the analysis, which immediately impairs representation quality.[13]

A major determinant of careless responses is the respondents' lack of motivation to pass all the steps of the question-answer-process (cf. also Huang et al., 2015; Krosnick, 1991).[14] Previous studies have pointed to the fact that failing attention checks is promoted by situational factors that affect a respondent's motivation to respond adequately (e.g. Anduiza & Galais, 2016; Gummer et al., 2018; Huang et al., 2012; Maniaci & Rogge, 2014) such as his or her lack of interest in the survey topic (e.g. Gummer et al., 2018).[15] This might explain why careless responding is understood as a "transitory (state) phenomenon, allowing for the possibility that the same individual might provide high levels of attention in one study (e.g., a short and particularly interesting study) but insufficient levels of attention in other studies" (Maniaci & Rogge, 2014, p. 62) and leads us to expect that excluding respondents from the analyses who are identified by attention checks as being engaged in careless response behavior will increase measurement quality (effectivity hypothesis).

2.2 Effects of incentives on answer behavior

Whether or not target persons participate in a survey depends on their anticipated costs of survey participation (which arise, among other things, from cognitive efforts that are required for answering questions) as well as their subjectively expected utility (in the following referred to as expected utility) of survey participation (D. A. Dillman, 1978). Target persons will participate if the expected utility of survey participation exceeds the anticipated costs.

Respondents' expected utility can be assumed to be shaped by two internal sources of (intrinsic) motivation (i.e., general intrinsic motivation such as a respondent's interest in participating in surveys or desire to contribute to scientific research by his or her answers and specific intrinsic motivation that is activated by aspects of a particular survey such as the survey topic) and sources of extrinsic motivation such as survey incentives. As a consequence, target persons whose level of intrinsic motivation does not outweigh the anticipated cost of survey participation might participate in an incentivized survey because the incentive increases the expected utility decisively such that a respondent's expected utility exceeds his or her anticipated costs of participation. This expectation is in line with findings from research on survey incentives and its increasing effects on response rates in online surveys (e.g. Becker et al., 2019; Göritz, 2006), particularly on the part of less motivated respondents (Ernst Stähli & Joye, 2016). In this respect, it is reasonable to assume that samples where respondents are incentivized show a lower level of intrinsic motivation for survey participation on average compared to samples where target persons are not promised any incentive for participation (incentive-motivation hypothesis). As careless responses are assumed to be triggered by a lack of respondents' motivation and we assume intrinsic motivation to be more important than external motivation in pass-

[12] The apparent absence of the necessity of conducting "representative" research in psychological research (cf. Henrich, Heine, & Norenzayan, 2010) might explain why item nonresponse has been considered by so few authors as a manifestation of a careless response.

[13] Perfidious careless response is distinct from mental coin flipping as the latter concept is defined as the consequence of omitting step two and/or step three of the question-answer-process only [Krosnick, 1991]. Hence, marking a wrong response option of an explicitly instructed item which is specially designed to check whether or not respondents have read the survey question could not be categorized as mental coin flipping but rather as perfidious careless response. In a similar vein, compared to saying "don't know" (Krosnick, 1991) apparent careless response allows researchers to define those responses as careless where respondents mark the "don't know" option of an attention check rather than a response option along the answer scale.

[14] Careless responding is also distinct from social desirability (cf. Maniaci & Rogge, 2014, p. 65).

[15] Previous studies have also examined the effect of respondent's gender, age, and education (as a proxy for ability) on attention check failures (Anduiza & Galais, 2016; Berinsky et al., 2014; Gummer et al., 2018; Mancosu et al., 2019). In all studies, the signs of the effects suggest that female respondents, older respondents and respondents with high education, respectively, are less likely to fail attention checks compared to male respondents, younger respondents and respondents with low education, respectively. However, the effects of the three variables are not statistically significant across all studies. Gender has a significant effect on attention check failure in one study only (i.e. Berinsky et al., 2014), age is significant in two studies only (i.e. Berinsky et al., 2014; Gummer et al., 2018), and education is significant in three studies only (i.e. Anduiza & Galais, 2016; Gummer et al., 2018; Mancosu et al., 2019).

ing all the cognitive steps of the question-answer-process, we expect that failure rates of attention checks will be higher in samples where respondents are incentivized for participation compared to samples where respondents are not incentivized (incentive careless-response hypothesis).

Apart from the target persons' participation decision, the prospect of incentives is also likely to affect respondents' break-off decision. Respondents' motivation to carefully complete all the cognitive the steps in the question-answer process can be assumed to decrease over time (Krosnick, 1991). Incentives may compensate for the decline in respondents' motivation such that respondents will remain in the survey rather than break off even if intrinsic motivation decreases below the anticipated participation costs. Thus, incentives may even work as catalysts. In line with Göritz (2006), who finds that respondents are more likely to complete a survey when an incentive is offered than when not, we expect break-off rates to be lower in incentivized samples compared to nonincentivized samples (incentive-break-off hypothesis).

2.3 Effects of attention checks on response behavior

Attention checks might act as interventions that affect how respondents approach subsequent survey items and, hence, be more than mere measures of attention (Hauser & Schwarz, 2015). Some researchers fear that attention checks demonstrate a lack of respect for the survey respondent (Miller & Baker-Prewitt, 2009) or signal respondents distrust in their behavior (e.g. Niessen et al., 2016, p. 8). In this respect, attention checks might elicit adverse reactions among respondents and, thus, negatively affect answer behavior (demotivation hypothesis). In contrast, other researchers postulate that attention checks might have a positive effect on answers to subsequent survey items (e.g. Hauser & Schwarz, 2015; Huang et al., 2015) (motivation hypothesis). From this point of view, respondents might presume that paying close attention to the survey item is important and apparently highly valued by the survey designer (Hauser & Schwarz, 2015) such that "participants will empathize with the survey administrator's desire to obtain accurate data" (e.g. Huang et al., 2015, p. 305) and attentive respondents "may view their survey completion as more meaningful" when attention checks are applied to screen for perfidious careless responses.

3 Data, measures, and methods

3.1 Data

We administered our questionnaire online to two different samples in December 2014. In both studies, target persons were defined as persons with a (principal) residential address in Germany at the time of the survey. For Study 1, respondents were recruited from the access panel of a commercial research institute to build up a quota sample[16] on the three crossed characteristics: gender, age, and school education.[17] The commercial research institute paid participants a monetary incentive of € 1 for completing the survey. Hence, in sample 1 we cannot exclude the fact that respondents participated in the survey not only due to general and/or specific intrinsic but also due to extrinsic motivation.[18] 613 (= n_{started}) respondents started the survey, of whom 544 (= $n_{\text{completed}}$) completed it. For Study 2, participants were recruited from a non-commercial access panel, which is accessible to scientific researchers only. This sample is an ordinary non-probability sample.[19] In the registration procedure for this noncommercial access panel, members of the panel declare that they are prepared to participate in a maximum of four scientific surveys a year, whereas monetary incentives are not promised by the panel operator. This allows researchers to decide themselves whether to use incentives, such as lotteries, or merely distribute the results of the survey. In our case, we did not pay any incentive nor did we use any other form of incentive. Hence, it is reasonable to assume that the average level of general intrinsic motivation among respondents in Study 2 is higher compared to respondents in Study 1. 424 (= n_{started}) respondents started the survey, while 338 (= $n_{\text{completed}}$) participants completed it.

The survey topic was "adequate earning levels"; nonetheless, the questionnaire was exclusively designed to study methodological aspects. We had no epistemological interest in substantial questions on adequate earning levels, and communicated the character of the survey (i.e., method experiment) to our participants after completion.

In case of the quota sample it was unavoidable to ask re-

[16] Quota samples for online surveys reflect the first choice for social scientists who want to administer a complete questionnaire online to a broad range of demographic groups and do not have the time and/or money to set up a probability-based online panel. Among the nonprobability-based samples, quota samples promise the highest degree of external validity with regard to analysis results because the sample distribution is equivalent to the distribution of the selected demographic variables, which serve as control characteristics in the target population. However at the same time, they should be used with caution (Yeager et al., 2011; Zack, Kennedy, & Long, 2019).

[17] As a member of the European Society for Opinion and Market Research (ESOMAR), the market research institute applies ESOMAR guidelines.

[18] An incentive of € 1 might seem very little external motivation. However, some of the members of the access panel of the commercial survey institute try to participate in as many surveys as possible to increase their total incentive.

[19] In contrast to commercial access panels, non-commercial access panels do not allow quota samples to be built but are still well suited for conducting online survey experiments (e.g., factorial surveys) and have been used in an increasing number of publications in recent years (cf. e.g. Czymara & Schmidt-Catran, 2017; Goerres & Rabuza, 2014; Hörstermann & Andreß, 2015; Shamon, 2014; Shamon & Dülmer, 2014).

spondents on their sociodemographic characteristics at the beginning of a survey, as socio-demographic characteristics, at least those required for screening respondents, were needed for the steering process during the fielding. In case of the non-commercial panel, it would have been unconventional to ask respondents at the beginning of a survey on socio-demographic characteristics. Instead, respondents were asked at the end of the survey, as also proposed in general for self-administered surveys (cf. e.g. Bourque & Fielder, 2003; D. Dillman, 2007; Jackson, 2008). If we had asked respondents of the non-commercial panel at the beginning of the survey on their sociodemographic characteristics, we could not have ruled out that respondentsâĂŹ motivation (and hence their reactions) are (at least partly) influenced by this unconventional placement of the sociodemographic questions.

Table 1 shows the distribution of sociodemographic characteristics in both samples. The average age of respondents is higher in Study 1 than in Study 2, and the proportion of females is much higher in Study 2 than in the quota sample. Finally, the ordinary nonprobability sample is skewed regarding the respondents' education. That is, lower ISCED categories hardly participated at all whereas highly educated respondents made up half of the sample.[20] In summary, compared to Study 1 and the general German population, Study 2 is biased towards highly educated, young, and females.

3.2 Settings & ex-ante detection methods

For the examination of reactions to the investigated attention checks, we designed and integrated different experimental settings into our surveys. In Study 1, participants were randomly assigned to one of three settings (cf. Figure 1, Setting 1: no ex ante detection methods, Setting 2: ex ante detection methods with explanations, Setting 3: ex ante detection methods without explanation). In Study 2, participants were randomly assigned to either Setting 1 (no ex ante detection methods) or to Setting 2 (ex ante detection methods with explanation). In each of the respective settings with attention checks (i.e., Setting 2 and Setting 3), we integrated two attention check instructions and two attention check items that reflect objective ex ante methods of detecting perfidious careless responses. In the following, all the detection methods will be described in the order in which they appeared in the survey.

Attention check instruction 1. We started with an ex ante detection method that is suitable for assessing how many respondents read the instructions of the survey items properly. Starting with this detection method was promising because we assumed he highest failure rates among the ex ante detection methods investigated in this study to be found here. High failure rates imply that this attention check item draws the attention of only a "small" proportion of respondents to the use of ex ante detection methods in the survey.

In an item battery of the study comprising six individual items, we asked the respondents to state for six different jobs (i.e., shop assistant, doctor in general practice, unskilled factory worker, minister in the national government, chairman of an executive board, and facility manager) what they think the gross monthly earnings of such people actually are.[21] In Setting 1, we asked our respondents in the instruction to estimate the average gross earnings for each job and we provided them with an open text field for their answers. In Setting 2, we additionally instructed the respondents to answer "0" for a facility manager and explained that this is necessary for us to see whether or not they had read the instruction. The instruction in Setting 3 was identical to the instruction in Setting 2, except that we did not provide any explanation for our request (no explanation). This attention check instruction will be referred to as facility manager instruction.

Attention check item 1: careless response scale. In an item battery comprising three items, we replaced the last substantial item by the following careless response scale: "I was born before 1920". This wording asks for a realistic event that can be expected to prevent situations in which careful respondents endorse a bogus item because they find it amusing (cf. Meade & Craig, 2012). The realistic event can be specified for each survey in such a way that it should not be applicable to survey participants. Furthermore, surveys usually ask for respondents' age and year of birth so that it is post hoc possible to rule out that choosing one of the other options is a substantial answer rather than a careless response to this item. In Setting 2 of both studies, this statement was complemented by the following additional explanation: "With the help of your response to this statement, you show us that you have read the statement".[22] Respondents could answer on a seven-point scale with the endpoints being verbalized as "does not apply at all" and "applies completely." In our survey, the only answer option that was very likely to apply to all respondents was "not at all" for the born before 1920 item. Exit options were not offered for the items of this item battery. This attention check item will be referred to as born before 1920 item.

[20] ISCED is an abbreviation for the International Standard Classification of Education (cf. OECD, Eurostat, & UNESCO, 2015). For coding education, an 8-point scheme ranging from 1 to 8 was used. The 8-point scheme is a slightly modified ISCED-97 classification, which was developed for comparison purposes within the framework of the European Values Study 2008/2009 by Dülmer, Jagodzinski, and Siegers (2008).

[21] This item battery was based on the idea of the items V26 to V36 in the basic questionnaire of the ISSP survey on social inequality (GESIS Data Archiv, 1999).

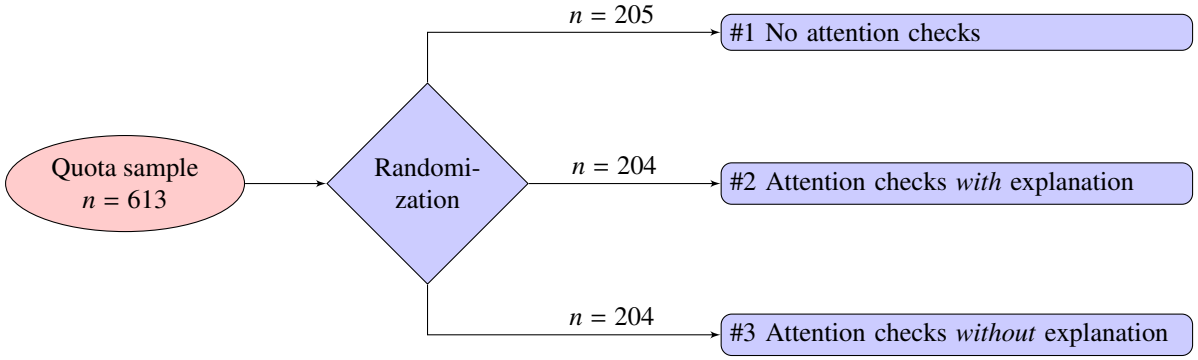
[22] The wording of the careless response scale in Setting 1 of both studies in German is: "Ich bin vor 1920 geboren. Mit Hilfe Ihrer Antwort auf diese Aussage zeigen Sie uns, dass Sie die Aussage auch gelesen haben."

Table 1
Distribution of sociodemographic characteristics in both studies

	Study 1				Study 2[a]	
	started		completed		completed	
	%	*n*	%	*n*	%	*n*
Gender						
Male	50	306	48	281	41	139
Female	50	307	52	263	59	198
Else	-[b]	-[b]	-[b]	-[b]	0	1
ISCED						
1 & 2	9	56	8	44	0	1
3 & 4	23	357	24	130	1	3
5 & 6	48	133	48	259	39	131
7 & 8	20	124	20	111	59	199
Missing	0	0	0	0	1	4
Mean age	44		44		38	
Median age	45		45		34	
Observations	613		544		338	

[a] Sociodemographic questions were placed at the end of the questionnaire, which is why we cannot report the distributions of sociodemographic characteristics for started. [b] Answer option was not presented, because the gender question was part of the screening questions and (crossed) quota were defined for male and female only.

(a) Study 1 (with remuneration)



(b) Study 1 (No remuneration)

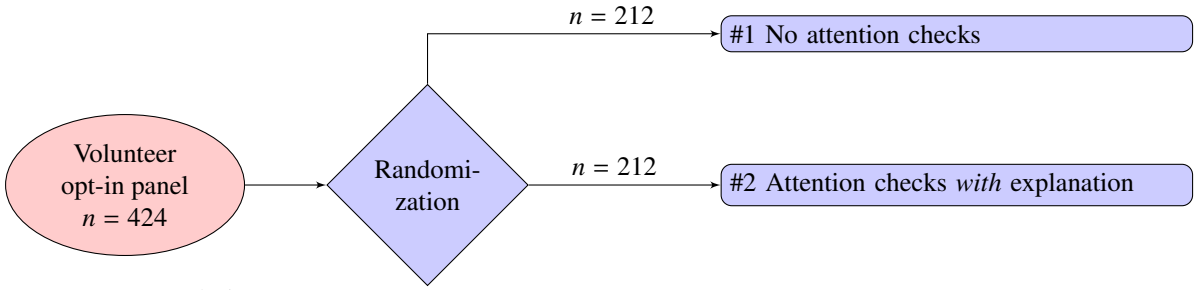


Figure 1. Survey designs

Attention check item 2: explicit instructed response item. In another (subsequent) item battery comprising three items, we replaced the last substantial item in Setting 1 by explicit instructed response items without explanations in Setting 3 of Study 1. The mark "applies completely" item was formulated as follows: "In this line, we ask you to mark the right option 'applies completely' on the answer scale". In Setting 2 of both studies, we explained our request in a subordinate clause as follows: "to show that you have read this sentence".[23] Exit options were not offered for the items of this item battery. This attention check item will be referred to as mark applies completely item.

Attention check instruction 2. We also used an explicit instructed response instruction for the tenth vignette rating task in the respective settings of our study. This helped us to examine how many respondents paid attention to the last instruction in a situation of repeated vignette rating tasks. In Setting 3 of Study 1, we asked our respondents to mark the right option of the 11-point scale "unjustified too high". In Setting 2 of both studies, we additionally explained that they should mark the option "to show that you have read this instruction". To be consistent with the nine previous vignette rating tasks, we offered respondents an exit option for the final rating task. This attention check instruction will be referred to as unjustified too high instruction.

3.3 Data quality measurements

As mentioned in the introduction, we used questions regarding adequate earning levels to assess data quality. In the questionnaire, we presented respondents with ten identical vignettes taken from a factorial survey by Shamon and Dülmer (2014). These vignettes described fictitious persons using a combination of different attributes (e.g., high work effort, no children) of different dimensions (work effort, number of children) (cf. Table A1 in the appendix). We integrated an MTMM initially proposed by Campbell and Fiske (1959) in our questionnaire, by systematically varying the concept of interest among the first nine vignette rating tasks, and the answer formats. Respondents were questioned about their idea of a just earnings level, what they think is needed for a decent life, and for adequate social participation. Specifically, respondents were asked: "The person described above has a gross monthly income of € 3000. Would you say that this person's income is unfairly low, fair, or unfairly high?"[24] (trait 1) and whether "this income is too low, sufficient, or too high for all family members to lead a decent life?"[25] (trait 2). Lastly, the respondents should indicate whether "the income of the person described above is too low, sufficient, or too high for all family members to adequately participate in social activities?"[26] (trait 3). All three traits have in common the fact that respondents' answers are expected to depend on their preferences for a distribution rule according to which earnings ought to be allocated among individuals (ideal standards) as well as on conditions of the social context that allow for social comparisons (existential standards), such as average earnings or pay inequality in a reference group (Shamon & Dülmer, 2014; Shepelak & Alwin, 1986). We not only varied the constructs, but also the scale, i.e., the answer options. Each concept was measured once by each scale (method). First, we used an open answer format (method 1), second, an endpoint verbalized eleven-point scale (method 2), and, third, an endpoint verbalized slider bar (method 3). Respondents had an exit option in all three methods.

Open-ended answer formats are applied in social justice research for the measurement of people's normative beliefs on just earnings (cf. Hysom & Fişek, 2011; Jasso, 1978; Shamon, 2014; Shamon & Dülmer, 2014), while closed (scale-based) answer formats are used to measure people's evaluations of actual earnings (cf. Cohn, White, & Sanders,

[23] The wording of the attention check item 2 in Setting 1 of both studies in German is: "In dieser Zeile bitten wir Sie die rechte Option 'Trifft voll und ganz zu' zu markieren, um zu zeigen, dass Sie diesen Satz gelesen haben."

[24] Original wording in the German questionnaire is: "Wir sind an Ihrer Gerechtigkeitsvorstellung interessiert. Die oben geschilderte Person verdient € 3000 Brutto im Monat. Würden Sie sagen, dass das Einkommen der Person ungerechterweise zu niedrig, gerecht oder ungerechterweise zu hoch ist?", which translates to: "We are interested in your idea of justice. The person described above has a gross monthly income of € 3000. Would you say that this person's income is unfairly low, fair, or unfairly high?" in English.

[25] Original wording in the German questionnaire is: "Wir sind an Ihrer Vorstellung über ein menschenwürdiges Leben interessiert. Die oben geschilderte Person verdient € 3000 Brutto im Monat und ist Alleinverdiener bzw. Alleinverdienerin in ihrem Haushalt. Würden Sie sagen, dass das Brutto-Einkommen der Person weitaus zu gering, ausreichend oder weitaus höher als erforderlich ist, damit alle Familienmitglieder ein menschenwürdiges Leben führen können?", which translates to: "We are interested in your idea of a decent life. The person described above has a gross monthly income of € 3000 and is the sole wage earner in his or her household. Would you say that this person's gross income is much lower, sufficient, or much higher than necessary?" in English.

[26] Original wording in the German questionnaire: "Wir sind an Ihrer Vorstellung über eine angemessene Teilhabe am gesellschaftlichen Leben interessiert. Die oben geschilderte Person verdient € 3000 Brutto im Monat und ist Alleinverdiener bzw. Alleinverdienerin in ihrem Haushalt. Würden Sie sagen, dass das Brutto-Einkommen der Person weitaus zu gering, ausreichend oder weitaus höher als erforderlich ist, damit alle Familienmitglieder in angemessener Weise am gesellschaftlichen Leben teilhaben können?", which translates to: "We are interested in your idea of adequate participation in social activities. The person described above has a gross monthly income of € 3000 and is the sole wage earner in his or her household. Would you say that this person's gross income is much lower, sufficient, or much higher than necessary for all family members to adequately participate in social activities?" in English.

2000; Gatskova, 2013; Jasso & Webster, 1997). Jasso (1978) showed that a person's evaluation of justice (J), that is their attitude towards the earnings of a specific rewardee can be derived by applying a logarithmic transformation of the rewardee's actual earnings (A) and the person's normative belief about just earnings for the rewardee (C); see Equation 1. The basic idea behind the logarithmic transformation of the two values is that deficiencies of the absolute value $Z = A - C < 0$ evoke a stronger sense of injustice among the evaluating persons than any surplus of the same absolute value $Z = A - C > 0$:

$$J = \ln\left(\frac{A}{C}\right) \quad . \tag{1}$$

We think that the underlying idea of Jasso (1978) logarithmic transformation is not merely confined to feelings about justice. Kanouse and Hanson (1972) pointed to the general tendency of persons "to weigh negative aspects of an object more heavily than positive ones" (p. 47) when forming an overall evaluation of the object. This tendency has become known as negativity bias. According to Rozin and Royzman (2001), negativity bias is manifested in four ways. Among others, it is manifested in the potency that "negative events are more potent with respect to their objective magnitude than are positive events" (298 Rozin & Royzman, 2001), which is also described in the prospect function (Kahneman & Tversky, 1979; Tversky & Kahneman, 1991) and is at the basis of the loss aversion phenomenon (cf. Rozin & Royzman, 2001). In this respect, an individual's final impression is determined to a greater extent by negative than by positive information (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). Therefore, we also apply the logic of Jasso (1978) logarithmic transformation to the other two concepts.

At the end of the survey, we explained to our respondents the scientific character of our study and the necessity of carefully processing the survey questions and set two self-report scales. We asked our respondents how carefully they had read the survey items (hereinafter referred to as careful-read item scale) and how carefully they had read the instructions of the survey items (hereinafter referred to as careful-read instruction scale). Respondents could answer on a 7-point scale from 0—"not read carefully at all" to 6—"read very carefully" as endpoint verbalizations.

3.4 Analytical strategy

Rates of perfidious careless responding. Analyses of rates of perfidious careless responding will utilize two observational bases. First, perfidious careless response rates will be analyzed according to the gross sample of each setting ($n_{\text{started},s}$). Using n_{started} as the observational basis tends to deflate perfidious careless response rates because respondents who broke off the survey at the respective page or at any of the previous pages will also be considered in the analysis even though they did not provide valid answers to the survey items. This deflation tendency can be expected to be the stronger, the higher the break-off rate of a survey is and the further back the items are placed in the survey. As a countermeasure, we will also report perfidious careless response rates on the basis of those respondents who did not break off the survey up to the appearance of the attention check item (n_{survived}). This figure is informative in the sense that it indicates the proportion of respondents who failed the attention check at the respective place in a specific survey among those respondents who participated in the survey up to the appearance of the attention check. However, using n_{survived} as the observational basis tends to inflate perfidious careless response rates. The inflation tendency can be expected to be the stronger, the higher the break-off rate of a survey is and the less correlated break-off and perfidious careless responding are. Hence, both rates are informative from a specific point of view but each has drawbacks that can be overcome by reporting both of them.

Reactions to attention check items and instructions. In order to investigate potential reactions to the use of attention check instructions and attention check items, we draw on three objective, nonreactive indicators as well as on two self-report scales.

Break-off: For each study, we estimate logistic regressions (break-off as dependent variable) with dummies for the settings to examine the main effects (hereinafter referred to as main-effect model) on the basis of n_{started}. Additionally, logistic regression models will be estimated based on n_{started} in which we account for a potential interaction effect between settings and interest in the survey topic as a source of specific intrinsic motivation for participation (hereinafter referred to as interaction-effect model).

Item nonresponse: Among the respondents who completed the survey ($n_{\text{completed}}$), we counted the number of items left blank by each person and added up this number over all the persons of a setting (total number of unanswered items per setting). In order to account for the differences in the observational basis of the settings, we divided the total number of unanswered items per setting by the number of items (= 36) that were posed to all respondents of a setting who completed the survey ($=n_{\text{completed},s}$), that is to say, by the product of items and respondents per setting. In doing so, we obtained the proportion of unanswered items in a setting, which are comparable across settings. The proportion of unanswered items in a setting is used as the dependent variable of an ordinary least squares estimation in which we examine the main effects between the settings of a study by accounting for dummies for the settings and controlling for age, gender, and ISCED.

Measurement quality: To evaluate measurement quality, we use an MTMM, initially proposed by Campbell and Fiske

(1959). Based on at least three traits (j) and three methods (i), this design allows us to calculate reliability and validity scores for each trait-method combination with the data of respondents who completed the survey ($n_{\text{completed}}$). The design is based on a true score measurement model (Saris & Andrews, 1991):

$$Y_{ij} = h_{ij}T_{ij} + e_{ij} \tag{2}$$

$$T_{ij} = v_{ij}F_j + m_{ij}M_i \tag{3}$$

Y represents the observed score for the j^{th} trait and the ith method. The measure consists of a true score T, i.e., a systematic component and an error term e. F is the latent factor for the j^{th} trait and M is the common variance due to the i^{th} method. The standardized factor loading h_{ij} is the reliability coefficient of the item and v_{ij} is the validity coefficient of the measure for the j^{th} trait and the i^{th} method. The standardized factor loading m is the method coefficient of the item. Figure 2 shows the estimated structural equation models. For the sake of clarity, the observed variables Y_{ij} and the error terms e_{ij} are not shown. Ellipses are latent variables, arrows represent factor-loadings, and double-headed arrows show covariance.

Combining reliability and validity allows us to calculate quality coefficients q_{ij}, i.e. measurement quality depends on validity and reliability (cf. Equation 3):

$$q_{ij}^2 = v_{ij}^2 h_{ij}^2 \tag{4}$$

To compare the results across settings, we calculated the average measurement quality q_{ij}^2 for each setting. As is common in related research, we use point estimates to assess measurement quality (Revilla, Saris, & Krosnick, 2014). The comparison of average quality scores across the settings of a study shows whether reactions are in favor of the motivation or demotivation hypothesis. Furthermore, it is possible to compare the average quality score within a setting between all respondents of a setting and those respondents of the setting who passed the ex ante detection methods. This comparison allows us to assess the effectivity of attention checks in terms of measurement quality scores. All MTMM models were estimated in Mplus 7 based on a robust full-information maximum likelihood estimator with missing data. We assessed the goodness of fit with the following commonly used indices: X^2/df, CFI, and RMSEA. Models with a $X^2/df < 5$, CFI > 0.95, and RMSEA < 0.06 are considered to have a good fit to the data (Bentler, 1990; Boomsma, 2000; Hu & Bentler, 1999).

Self-reported carefulness: Finally, we will estimate OLS regressions for both the careful-read item scale and careful-read instruction scale based on $n_{\text{completed}}$ to examine differences between the experimental settings. This analysis allows us to obtain insights into the mode of operation of attention check items and instructions.

4 Results

4.1 Rates of perfidious careless Responses

Study 1. In Study 1 the attention check instructions yielded higher perfidious careless response rates compared to the attention check items (cf. Table 2).[27] In settings 1 and 2 of Study 1, 69.12 percent of the participants did not comply with the facility manager instruction, while 3.92 percent (or 5.88 percent) of the respondents in Setting 2 (or Setting 3) left the respective item blank. The perfidious careless response rate was even higher for the unjustified too high instruction, which was embedded in the tenth vignette rating task. The discrepancy in the perfidious careless response rates between the two attention check instructions of 11.27 percentage points in Setting 2 and 14.21 percentage points in Setting 3 might be explained by the respondents' decreasing motivation to read instructions during the survey, particularly in the context of the large number of repetitions of the vignette rating task. Repeatedly asking respondents for the same (vignette rating) task might have led them to believe that re-reading of the instructions related to the vignette rating task is superfluous, and, therefore, may lead to a higher perfidious careless response rate in the unjustified too high instruction.

According to the born before 1920 item, in Setting 2 of the first study, 7.35 percent (in Setting 3 6.37 percent) of the respondents were identified as being engaged in careless responding. These rates are just as low as the perfidious careless response rate identified by the mark applies completely item. According to the mark applies completely item, 8.33 percent of participants in Setting 2 and 9.80 percent of participants in Setting 3 engaged in careless responding in the respective item battery, while the proportion of persons who left the respective item blank does not differ substantially between the two attention check items in a setting. By implication, we can conclude that the vast majority of respondents in Setting 2 and Setting 3, respectively, noticed the usage of attention check items. Respondents' awareness of the usage of ex ante detection methods in surveys is a sine qua non for the potential emergence of reactions to these instruments.

Study 2. The rates of each of the four objective ex ante detection methods are substantially lower in Study 2 than in Study 1, which corroborates the incentive careless response hypothesis. The difference in the facility manager instruction between Setting 2 in Study 1 and Study 2 amounts to 14.4 percentage points while the difference in the unjustified too high instruction between Setting 2 in Study 1 and Study 2 amounts to 11.99 percentage points. However, in Study 2 the

[27] As can be seen in Table 3, calculating perfidious careless response rates on the basis of both n_{started} and n_{survived} does not alter the pattern of the results. Thus, although we report both rates in Table 3, we will only refer to perfidious careless response rates on the basis of n_{started} in the discussion of the results.

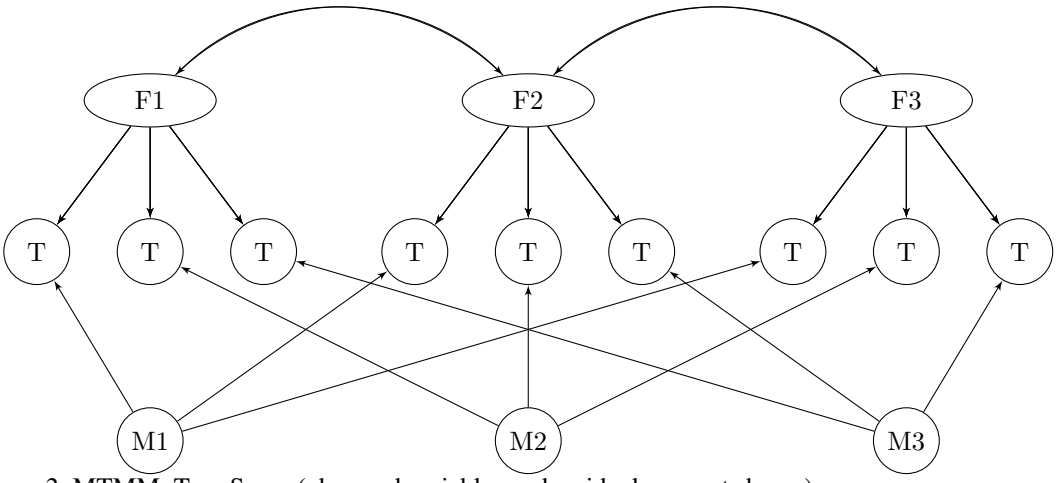


Figure 2. MTMM: True Score (observed variables and residuals are not shown)

difference between the facility manager instruction and the unjustified too high instruction amounts to 13.68 percentage points in Setting 2, which is 2.41 percentage points higher compared to the difference between the attention check instructions in Setting 2 of Study 1. In Setting 2 of Study 2, none of the respondents was identified as being engaged in careless responding using the born before 1920 item, while one respondent failed to answer the mark applies completely item correctly. That is to say, we can conclude with high confidence that almost all respondents noticed the attention check items in Setting 2.

In both studies, results regarding the attention check instructions show that most respondents did not read the instructions properly. In each study, attention check instruction 1 reveals that more than 50 percent of the respondents engaged in careless responding. Attention check instruction 2 (i.e., the attention check that was embedded in the instruction of the last item of a set of 10 subsequently presented rating tasks) identifies an even higher proportion of careless responses compared to attention check instruction 1. Taken together, these results do not provide confidence that a "better" placement of the attention check instructions (e.g., placing attention check instruction 2 in the instruction of the first of ten subsequently presented rating tasks) substantially decreases the number of respondents identified as careless responders.[28] Hence, the placement of an attention check in the instruction of an item does not appear effective for identifying careless responses. The placement of attention checks in instructions might be theoretically more meaningful if a researcher intends to communicate very important information related to the survey item and if he/she wants to check how many respondents noticed the information provided. However, at the same time, our results suggest that survey designers might be on a safer side if they embed the information in the survey item text than in its instruction if it is important, because only a minority of respondents read the instruction. Due to their limited effectiveness in identifying careless responses, we will ignore both attention check instructions in the following reaction analysis of measurement quality (Section 3.2.2).

While the results of the attention check instructions were similar in both studies, results concerning the attention check items differed substantially across the two samples. In the generally intrinsically motivated sample of Study 2, (almost) none of the respondents failed both attention check items. Up to this point, the results suggest that attention check items are superfluous in exclusively motivated samples as used in Study 2.

4.2 Reactions

Break-off. Even though participants in Study 2 are expected to show on average a higher (general) intrinsic motivation for participating in surveys, break-off is higher in Study 2 compared to Study 1. In total, 69 of the 613 (11.26 percent) participants broke off the survey in Study 1, while in Study 2, 86 of the 424 (20.28 percent) participants broke off the survey in total (cf. Table 3). This result is in line with our incentive-breakoff hypothesis and might be explained by the fact that respondents in Study 2 are more motivated to contribute to substantial than to methodological research questions. As our MTMM design is characterized by repeated questions that might have particularly revealed the methodological character of our study but also have bored

[28]The results of Oppenheimer et al. (2009) on item manipulation check are another source of evidence for the possible ineffectiveness of attention check instructions for detecting careless responding. As mentioned in the introduction, the item manipulation check which resembles the attention check instruction regarding the placement of the attention check in the survey instrument identified 46% of the respondents as careless responders (Oppenheimer et al., 2009).

Table 2
Descriptive results of detection methods - failure rate (in %)

	Failed Respondents		Break-off	Item nonresponse
	% of $n_{started}$	% of $n_{survived}$	%	%
Study 1				
Setting 2 (n=204)[a,b]				
Facility manager instruction	69.12	71.94	3.92	-
Born before 1920 item	7.35	8.02	8.33	-
Mark applies completely item	8.33	9.14	8.82	-
Unjustified too high instruction	80.39	93.18	9.31	4.41
Setting 3 (n=204)[c]				
Facility manager instruction	69.12	73.44	5.88	-
Born before 1920 item	6.37	7.08	9.31	-
Mark applies completely item	9.80	10.81	9.31	-
Unjustified too high instruction	83.33	97.14	10.29	3.92
Study 2				
Setting 2 (n=212)[a,b]				
Facility manager instruction	54.72	54.72	0.00	-
Born before 1920 item	0.00	0.00	8.96	-
Mark applies completely item	0.47	0.53	11.32	-
Unjustified too high instruction	68.40	79.67	11.79	2.36

Reading example: In Setting 2 of Study 1, 69.12% of the respondents who started the survey or 71.94% of the respondents who did not break-off the survey at the respective page or at any of the previous pages failed the facility manager instruction.
[a] Ex ante detection methods were not used in Setting 1 [b] Ex ante detection methods with explanation
[c] Ex ante detection methods without explanation

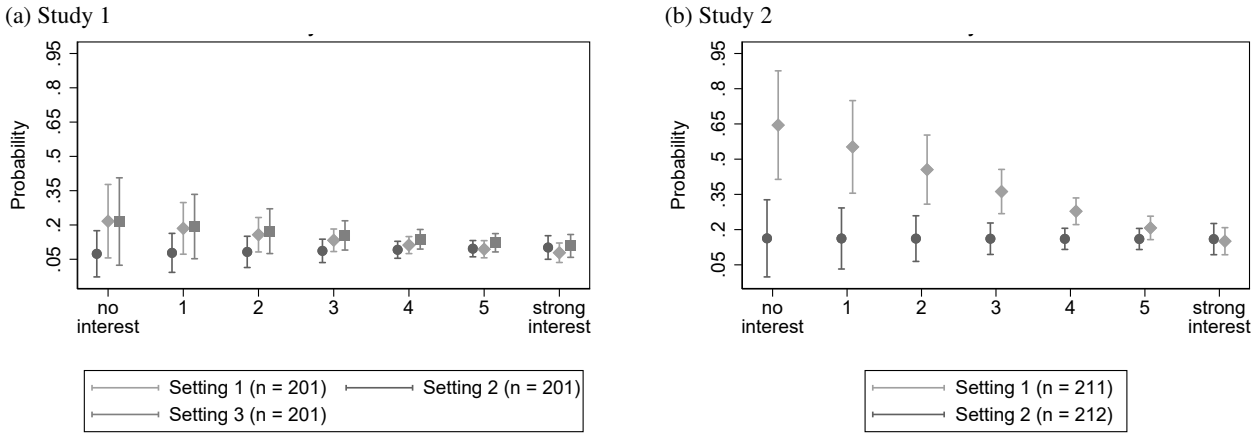


Figure 3. Break-off probabilities based on interaction effect model in Studies 1 and 2. Notes: Forecasts are based on logistic regression (n = 423); 1 person broke off leaving interest in topic blank; Dependent variable is break-off (=1); Setting dummies and interaction terms with interest in topic.

respondents, participants in Study 2 might simply have broken off the survey in the absence of incentives as an external source of motivation. In contrast, in Study 1 respondents who started the survey might also have had great interest in completing the survey, presumably because they would otherwise not be paid the incentive (hereinafter referred to as incentive effect).

Turning to the analysis of break-off rates as reaction measures for attention checks, the observable pattern is the same in both studies. In Study 1, the break-off rate in Setting 1 is higher than in Setting 2, while the break-off rate in Setting 3 is in between that of the two other groups. In Study 1, break-off rates vary across the settings between 9.8 percent and 12.68 percent, and between 16.04 percent and 24.53 percent in Study 2. However in Study 1, we found neither a significant main level effect between Setting 1 and one of the other two settings in the main effect model nor significant conditional effects in the interaction effect model ($\alpha = 0.10$) (cf. also Figure 3b). This finding is in line with other studies (i.e. Berinsky et al., 2014; Gummer et al., 2018) which examined break-off rates as measures of reaction to attention checks on the basis of incentivized samples.

In contrast to Study 1, results for Study 2 showed that the risk of a respondent's break-off in Setting 1 at 8.49 percentage points is significantly higher than in Setting 2 ($\alpha = 0.05$). This finding supports the motivation hypothesis in the noncommercial access panel. Furthermore, the interaction effect model reveals a significant interaction effect between setting and respondents' interest in the topic ($\alpha = 0.05$) As can be seen from Figure 3 (on the right-hand side), the risk of breaking off was higher among respondents in Setting 1 with a low specific intrinsic motivation (i.e., interest in the survey topic) than comparable respondents in Setting 2. That is to say, in Study 2 attention checks exerted a motivational influence particularly on respondents with low specific intrinsic motivation for participation in the survey.

In summary, while we neither found empirical evidence for the motivation nor for the demotivation hypothesis in Study 1, the results of Study 2 speak in favor of the motivation hypothesis rather than the demotivation hypothesis, particularly if respondents have little interest in the survey topic. These different results across both studies can be explained in the context of the incentive effect that compensates for any potential decrease in intrinsic participation motivation and, hence, moderates the motivating effect of attention check instructions and attention check items. This raises the question of the applicability of break-off rates as measures of reaction in incentivized samples.

Item nonresponse. The proportion of the total number of unanswered items is higher in Study 1 (2.27 percent) than in Study 2 (0.75 percent). Considering item nonresponse in absolute term as a manifestation of apparent careless responses, this pattern corroborates our incentive careless response hypothesis. While in Study 1, the proportion is highest in Setting 1 (2.56 percent) and lowest in Setting 2 (2.04 percent), in Study 2, the proportion of items left blank is lower in Setting 1 (0.63 percent) than in Setting 2 (0.86 percent) (cf. Table 4). However, the difference between Setting 1 and any of the settings is not significant in Study 1 or in Study 2.[29] Hence, regarding item nonresponse, we do not find any support for the motivation or the demotivation hypothesis. While the results of Study 1 are in line with other studies using incentivized samples, the results of Study 2 might be explained by the fact that demotivated respondents dropped out rather than complete the survey with little intrinsic motivation to pass all four steps of the question-answer process.

Table 3
Break-off rates in Study 1 and Study 2

	Break-off		
	%	$n_{\text{break-off}}$	n
Study 1			
Setting 1	12.68	26	205
Setting 2[a]	9.80	20	204
Setting 3[b]	11.27	23	204
Study 2			
Setting 1	24.53	52	212
Setting 2[a]	16.04	34	212

Analyses based on n_{started}; percentages refer to sample sizes of settings.
[a] Ex ante detection methods with explanation. [b] Ex ante detection methods without explanation.

Measurement quality. Before we turn to the assessment of measurement quality, we will evaluate the fit indices of our true score models. We estimated a series of measurement models on the basis of all respondents who completed the survey ($n_{\text{completed}}$) and found that all specifications show an adequate fit to the data. The fit indices are summarized in Table 5.[30]

As described in Section 3, the true score models were used to calculate average measurement quality scores (cf. Table 6). The second column of Table 6 reports the average measurement quality in both studies, while the third column (un-

[28] As participants in Study 2 were asked about sociodemographic characteristics at the end of the survey, we could not examine the effect of sociodemographic characteristics on break-off nor control for sociodemographic characteristics in the break-off analysis.

[29] Results of the regression models are reported in the appendix in Table A9 and A10, respectively.

[30] The detailed results of factor loadings (reliability and validity) for each setting are provided in the appendix in Table A2 to Table A8.

Table 4
Proportion of item nonresponse in Study 1 and Study 2

	Item nonresponse	
	%	n
Study 1		
Setting 1	2.56	179
Setting 2[a]	2.04	184
Setting 3[b]	2.38	181
Study 2		
Setting 1	0.63	160
Setting 2[a]	0.86	178

Analyses based on $n_{completed}$; percentages reflect proportion of items left blank by respondents who completed the survey.
[a] Ex ante detection methods with explanations.
[b] Ex ante detection methods without explanations.

filtered) of Table 6 depicts the average measurement quality for all respondents in each setting, i.e., who passed and failed the attention check items. The last column (filtered) of Table 6 shows the average measurement quality exclusively for those respondents who passed the attention check items in Setting 1 and Setting 3, respectively.[31]

In line with the incentive-motivation hypothesis, the measurement quality of Study 2 exceeds the measurement quality of Study 1 (cf. Table 6). We further find that all scores based on Setting 2 (or Setting 3) are higher than in Setting 1 (cf. Table 6, column unfiltered and filtered). That is to say, ex ante detection methods consistently increase measurement quality across studies. Furthermore, in Study 1 we find that measurement quality is higher in Setting 2 (with explanation) than in Setting 3 (without explanation). Overall, these findings support the motivation hypothesis, i.e., seeing ex ante attention check items increases average measurement quality, especially when these items are presented with explanations.

The comparison of respondents who passed or failed both attention check items (cf. Table 6, column unfiltered) and respondents who passed both attention check items (cf. Table 6, column filtered) in Study 1 shows that omitting respondents who failed the attention checks from the sample increases measurement quality. This pattern is consistent for settings with and without explanations and, hence, supports the effectivity hypothesis.

Self-reported carefulness. The results presented in this section are based on all respondents who completed the survey and provided valid answers to the self-reported carefulness questions. In Study 1, 20 out of 544 respondents did not answer the careful-read item scale, while in Study 2 all 338 participants provided a valid answer to this scale. Regarding the careful-read instruction scale, 23 out of 544 participants in Study 1 and 4 out of 338 participants in Study 2, respectively, left this scale blank.

Study 1. The average carefulness in the careful-read item scale varies between 4.98 in Setting 1 and 5.35 in Setting 2. That is to say, respondents reported retrospectively that they had read the survey items almost very carefully. Even though there is a difference of nearly one scale point between the careful-read item scale and the careful-read instruction scale in each setting, the reported average carefulness with respect to the instructions of the survey items in the careful-read instruction scale is still very high. The average carefulness in the careful-read instruction scale varies between 4.17 in Setting 2 and 4.05 in Setting 3. This difference of nearly one scale point between the careful-read item scale and the careful-read instruction scale does not seem to reflect the substantial differences in failure rates between the attention check instructions and the attention check items found in our study.[32] Nevertheless, we can observe for both self-report scales that the average carefulness is highest in Setting 2 compared to the other two settings. For each of the two self-report scales, we estimated study-specific regression models to assess whether differences between the settings are significant while controlling for age, gender and ISCED (cf. Table A9 in the appendix). The difference between Setting 2 and each of the other two groups is significant ($\alpha = 0.01$) with respect to the careful-read item scale, but not regarding the careful-read instruction scale ($\alpha = 0.10$) The motivation hypothesis is thus partly corroborated.

Study 2. In Study 2, we observe a similar pattern to that in Study 1regarding both self-report scales. The average carefulness in the careful-read item scale and the careful-read instruction scale, respectively, is higher in Setting 2 than in Setting 3.[33] However, the difference is significant only in the case of the careful-read item scale ($\alpha = 0.01$) but not in the case of the careful-read instruction scale ($\alpha = 0.10$) as

[31] We did not filter on the basis of attention check instructions, due to their limited practical utility as a measure of perfidious careless responses in our study (cf. Section 3.1).

[32] In Study 1, those who passed the attention check items in Setting 2 and Setting 3, respectively, display significantly higher scores on the careful-read item scale but not on the careful-read instruction scale (biserial correlations). However, the careful-read instruction scale only correlates significantly with attention check instructions among respondents of Setting 2.

[33] In Setting 2 of Study 2, the careful-read instruction scale correlates significantly with both attention check instructions (biserial correlation), but not with the attention check items. As only one person failed an attention check item, correlations between the self-report scales and attention check items are meaningless.

Table 5
True score models: model fit indices

	Study 1					Study 2	
	Passed and failed			Only passed		not appl.[c]	
Setting	1	2[a]	3[b]	2[a]	3[b]	1[a]	2
X2/df	1.81	1.12	1.50	1.48	1.67	1.92	0.92
CFI	0.975	0.994	0.983	0.980	0.973	0.968	1.000
RMSEA	0.068	0.027	0.050	0.054	0.066	0.076	0.000
N	174	180	179	161	155	160	177

Analyses based on $n_{\text{completed}}$.
[a] Ex ante detection methods with explanations;
[b] Ex-ante detection methods without explanations; [c] not applicable because (almost) no one failed the attention check items in Study 2.

Table 6
Average measurement quality in both studies

	All		Unfiltered		Filtered	
Setting	Avg. q	*n*	Avg. q.	*n*	Avg. q.	*n*
Study 1						
Setting 1	-	-	0.565	174	-	-
Setting 2[a]	0.571	544	0.577	180	0.631	161
Setting 3[b]	-	-	0.570	179	0.595	155
Study 2						
Setting 1	-	-	0.632	160	-	-
Setting 2[a]	0.650	337	0.667	177	-[c]	-[c]

Analyses based on $n_{\text{completed}}$;
[a] Ex ante detection methods with explanations;
[b] Ex ante detection methods without explanations;
[c] Not applicable because (almost) no one failed the attention check items in Study 2.

can be seen in Table A10 in the appendix.[34] Again, we find evidence for the motivation hypothesis only on the basis of the careful-read item scale.[35]

5 General discussion

Online surveys are cost-efficient, fast, and easy to implement. However, the many reasons for using online surveys raise questions about data and measurement quality in this survey mode. Respondents of online surveys obviously miss the social control component of face-to-face interviews. Previous research points to problems of careless responding in online surveys and suggests using different detection methods for the identification of careless responses, e.g., items that check whether the respondents are paying attention when they answer survey items (e.g. Berinsky et al., 2014; Hauser & Schwarz, 2015; Johnson, 2005; Kam & Meyer, 2015; Kurtz & Parrish, 2001; Oppenheimer et al., 2009). In this context, survey designers might wish to routinely use attention checks to identify respondents who are engaged in careless responding. However, while these detection methods are effective to a certain extent in increasing data quality by allowing researchers to filter for perfidious careless responses, it is still unclear whether ex- detection methods negatively affect the answer behavior of respondents who are not engaged in careless responding and, thus, do more harm than good. The scientific evidence on attention check items is solely based on surveys with incentivized respondents. Previous research shows that survey incentives affect response behavior (e.g., higher response rates Becker et al., 2019; Göritz, 2006, lower drop-out rates Göritz, 2006), but there was still no evidence as to whether incentives might contribute to perfidious careless responding.

In this study, we examined rates of and reactions to attention check items (i.e., explicitly instructed response item and careless response scale) and instructions. To answer the question whether these ex ante detection methods exert a motivating or demotivating influence on respondents, we examined respondents' reactions, among other things, by objective

[34] As in Study 1, we estimated a regression model with the careful-read item scale and the careful-read instruction scale, respectively, as the dependent variable and accounted for age, gender, and ISCED in addition to a dummy for the control group.

[35] Subsequent analyses show that self-reported carefulness on the careful-read item scale and careful-read instruction scale is in most settings negatively related to the failure of attention checks items and attention check instructions, respectively (cf. Table A11 and Table A11 in the appendix). This might be taken as a proof of validity of the self-report scale to distinguish between poor and strong performing respondents. However, at the same time, we cannot exclude that this significant difference is driven by the fact that respondents who passed the attention check items rated higher on the scale, because they remembered that they had successfully passed the attention check item(s) before. In this latter case, the biserial correlations might be taken as an indicator for the motivational effect of attention check items on response behavior that we found in our study.

Table 7
Self-reported average carefulness

	Careful-read item scale		Careful-read instruction scale	
Setting	Avg.	*n*	Avg.	*n*
Study 1				
Setting 1	4.98	168	4.08	166
Setting 2[a]	5.35	181	4.17	180
Setting 3[b]	5.02	175	4.05	175
Study 2				
Setting 1	4.73	160	3.36	159
Setting 2[a]	5.05	178	3.50	175

Analyses based on $n_{completed}$; Numbers of observations refer to the proportion of respondents per setting who did not answer a control item.
[a] Ex ante detection methods with explanation;
[b] Ex ante detection methods without explanation;

and nonreactive indicators (i.e. measurement quality, item nonresponse, and break-off rates). Respondents' reactions were analyzed with data from two sources that differ regarding the self-selection process of the participants. In Study 1, participants were gathered from a commercial access panel and monetary incentives were paid to respondents who completed the survey by the survey institute. In Study 2, participants were obtained from a noncommercial access panel, in which respondents were not promised any incentive for their participation. Hence, the two samples differed regarding respondents' motivation for participating in our survey (mixed motivation vs. general intrinsic motivation). For this reason, our results provide interesting insights that are relevant for survey designers.

Results on the attention check items differed substantially across the two samples. While in Study 1 (mixed motivation) about 7 percent (or 9 percent) of the respondents in Setting 2 and Setting 3 failed the first (or second) attention check item, in Study 2 (general intrinsic motivation) hardly any of the respondents failed the attention check items. In the context of rates reported in other studies, e.g., Johnson (2005), Kam and Meyer (2015), Kurtz and Parrish (2001), Oppenheimer et al. (2009), careless response rates seem rather moderate in our study. Most of the respondents in both studies failed to answer the attention check instructions, indicating that item instructions appear to be ineffective for identifying careless responding. However, this finding does not imply that survey item instructions are generally redundant. According to the attention check items applied in our study, a vast majority of respondents in Study 1 (mixed motivation) and almost all respondents in Study 2 (general intrinsic motivation) read the question text. Hence, our findings suggest that when exposed to a survey item with an instruction respondents attach different degrees of importance to these two components. From the vantage point of a respondent, reading an instruction might not be a precondition for providing an answer, but rather an option that is utilized when respondents need additional clarification. The higher rates of perfidious careless responses in Study 1 compared to Study 2 may be explained by the evidence that incentives in web surveys increase response rates particularly of less-motivated respondents (Ernst Stähli & Joye, 2016) and complements previous research on survey incentives that examined other consequences (e.g., response rates) by suggesting that incentivizing survey respondents yields higher careless response rates.

Regarding the reaction analysis, in line with previous research, we did not find any demotivating influence of attention checks on answer behavior. Instead, we found in both studies evidence for motivational effects on respondents' answers. We found measurement quality to be higher in settings with attention checks than in the respective control group (i.e., Setting 1). In Study 1, measurement quality increased by 2 percent, while in Study 2 measurement quality increased by 5.5 percent. Beside these purely motivational effects of our ex ante detection methods, measurement quality additionally increased by 9.3 percent when respondents engaged in careless responding were filtered on the basis of our two attention check items in Setting 2 of Study 1. Furthermore, with respect to break-off rates, we found in Study 2 (but not in Study 1) that attention checks exerted a motivational influence on answer behavior, particularly on those respondents who reported low interest in the survey topic (i.e., low specific intrinsic motivation for survey participation). The results of the careful-read item scale support the idea that increasing motivation on the part of respondents to answer survey items explains the effects we found in our study among the objective, non-reactive indicators. In both studies, respondents in settings with ex ante detection methods on average reported greater carefulness after reading the question texts than in the respective Setting 1. Hence, our reaction analyses justify the utilization of attention check items with explanations in online surveys even if hardly any of the respondents can be expected to be engaged in careless responding.

Additionally, we found a motivational effect of attention checks on break-off behavior in Study 2 that is not replicated in Study 1 where break-off rates are generally lower than in Study 2. This pattern is in line with research on survey incentives which found that subjects who access online surveys for whatever reason are more likely to finish them when an incentive is offered (Göritz, 2006). At the same time, this raises the question of whether break-off rates are an appropriate operationalization strategy when it comes to investigating the motivational effects of attention checks on answer behavior in incentivized samples, as has also been examined by Berinsky et al. (2014) and by Gummer et al. (2018). Fur-

thermore, this result may call for caution in not generalizing results from methodological online studies that are based on incentivized samples to online studies in general.

Future research on this issue is necessary as the results presented in this study rely on two different access panels that may not appropriately represent commercial and noncommercial access panels, respectively. In the same vein, Study 2 is potentially biased by an overrepresentation of highly educated, young, and female respondents, in such as these characteristics might relate to attention check failures. Nonetheless, while female and highly educated respondents are less likely to fail attention checks, one can argue that younger respondents are more likely to fail attention checks, which should in turn reduce the bias of education and gender in Study 1. In addition, the effects of these sociodemographic variables on attention checks items are not consistent across previous studies (Anduiza & Galais, 2016; Berinsky et al., 2014; Gummer et al., 2018; Mancosu et al., 2019).

In summary, ex ante detection methods do not only increase data quality by filtering for perfidious careless responses according to unambiguous indicators, but also exert a motivational influence on respondents who notice the use of ex ante detection methods by increasing their cognitive effort in answering survey items. Furthermore, in online surveys based on nonprobability samples, survey designers could benefit from the advantage of attention check items in comparison to post hoc measures for data quality and thus filter respondents engaged in careless responding during the fielding of the survey so that they can achieve both the initially targeted sample size and a high measurement quality in the data.

Acknowledgements

The authors thank the Research Training Group Social Order and Life Chances in Cross-National Comparison (SOCLIFE) at the University of Cologne, which is funded by the German Science Foundation, (DFG) for funding Study 1. We thank Dominik Leiner for constructive comments and suggestions on the questionnaire. Data are available upon request from the authors.

References

Anduiza, E. & Galais, C. (2016). Answering without reading: Imcs and strong satisficing in online surveys. *International Journal of Public Opinion Research*, *29*(3), 497–519.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology*, *123*(1), 101–103.

Becker, R., Möser, S., & Glauser, D. (2019). Cash vs. vouchers vs. gifts in web surveys of a mature panel study — main effects in a long-term incentives experiment across three panel waves. *Social Science Research*, (221–234).

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246.

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.

Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling: A Multidisciplinary Journal*, *7*(3), 461–483.

Bourque, L. B. & Fielder, E. P. (2003). *How to conduct self-administered and mail surveys.* Thousand Oaks: Sage.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Church, A. H. (1993). Estimating the effect of incentives on mail survey response rates: A meta-analysis. *Public Opinion Quarterly*, *57*(1), 62–79.

Cohn, E. S., White, S. O., & Sanders, J. (2000). Distributive and procedural justice in seven nations. *Law and Human Behavior*, *24*(5), 553–579.

Conrad, F. G., Tourangeau, R., Couper, M. P., & Zhang, C. (2017). Reducing speeding in web surveys by providing immediate feedback. *Survey Research Methods*, *11*(1), 45–61.

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

Czymara, C. S. & Schmidt-Catran, A. W. (2017). Refugees unwelcome? changes in the public acceptance of immigrants and refugees in germany in the course of europe's 'immigration crisis'. *European Sociological Review*, *33*(6), 735–751.

Dillman, D. (2007). *Mail and internet surveys the tailored design method.* Hoboken, NJ: Wiley.

Dillman, D. A. (1978). *Mail and telephone surveys: The total design method.* New York: Wiley-Interscience.

Dülmer, H., Jagodzinski, W., & Siegers, P. (2008). *Classification scheme for the modified education variable of the EVS 1999/2000*. Cologne: Mimeo.

Edwards, P., Cooper, R., Roberts, I., & Frost, C. (2005). Meta-analysis of randomised trials of monetary incentives and response to mailed questionnaires. *Journal of Epidemiology and Community Health*, *59*(11), 987–999.

Ernst Stähli, M. & Joye, D. (2016). Incentives as a possible measure to increase response rates. In Y.-C. Fu, C. Wolf, D. Joye, & T. W. Smith (Eds.), *The sage hand-*

book of survey methodology. (pp. 425–440). Los Angeles: SAGE.

Gatskova, K. (2013). Distributive justice attitudes in ukraine: Need, desert or social minimum? *Communist and Post-Communist Studies*, *46*(2), 227–241.

GESIS Data Archiv. (1999). International social survey programme: Social inequality - issp 1999 - basic questionnaire. Cologne. ZA3430 Data file Version 1.0.0. doi:10.4232/1.3430

Goerres, A. & Rabuza, F. (2014). The social opportunity costs of voting: A factorial vignette survey of a most-likely-to-vote population. Retrieved from https://ssrn.com/abstract=2462471

Göritz, A. S. (2006). Incentives in web studies: Methodological issues and a review. *International Journal of Internet Science*, *1*(1), 58–70.

Greszki, R., Meyer, M., & Schoen, H. (2014). The impact of speeding on data quality in nonprobability and freshly recruited probability-based online panels. In M. Callegaro, B. Reg, B. Jelke, A. S. Göritz, J. A. Krosnick, & P. J. Lavrakas (Eds.), *Online panel research: A data quality perspective.* (pp. 238–262). Chichester: Wiley.

Gummer, T., Roßmann, J., & Silber, H. (2018). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research. Online first.*

Hauser, D. J. & Schwarz, N. (2015). It's a trap! instructional manipulation checks prompt systematic thinking on "tricky" tasks. *SAGE Open*, *5*, 1–15.

Hauser, D. J., Sunderrajan, A., Natarajan, M., & Schwarz, N. (2017). 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.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–83.

Hörstermann, K. & Andreß, H.-J. (2015). ''wer nicht arbeitet, soll auch nicht essen!" eine vignettenanalyse zur bestimmung eines einkommensmindestbedarfs. *Zeitschrift für Sozialreform*, *61*(2), 171.

Hu, L.-T. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55.

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299–311.

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.

Hysom, S. J. & Fişek, M. H. (2011). Situational determinants of reward allocation: The equity–equality equilibrium model. *Social Science Research*, *40*(4), 1263–1285.

Jackson, S. L. (2008). *Research methods and statistics : A critical thinking approach*. Belmont, Calif.: Wadsworth.

Jasso, G. (1978). On the justice of earnings: A new specification of the justice evaluation function. *American Journal of Sociology*, *83*(6), 1398–1419.

Jasso, G. & Webster, M. (1997). Double standards in just earnings for male and female workers. *Social Psychology Quarterly*, *60*(1), 66–78.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–291.

Kam, C. C. S. & Meyer, J. P. (2015). How careless responding and acquiescence response bias can influence construct dimensionality: The case of job satisfaction. *Organizational Research Methods*, *18*(3), 512–541.

Kanouse, D. E. & Hanson, L. (Eds.). (1972). *Negativity in evaluations*. Morristown, NJ: General Learning Press.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Kurtz, J. E. & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the neo-pi-r. *Journal of Personality Assessment*, *76*(2), 315–332.

Leiner, D. (2019). Too fast, too straight, too weired: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 220–248. doi:10.18148/srm/2019.v13i3.7403

Mancosu, M., Ladini, R., & Vezzoni, C. (2019). 'short is better'. evaluating the attentiveness of online respondents through screener questions in a real survey environment. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *141*(1), 30–45.

Maniaci, M. R. & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83.

Meade, A. W. & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*(3), 437–455.

Miller, J. & Baker-Prewitt, J. (2009). *Beyond 'trapping' the undesirable panelist: The use of red herrings to reduce satisficing.* Paper presented at the CASRO Panel Quality Conference. New Orleans, LA.

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.

OECD, Eurostat, & UNESCO. (2015). *Isced 2011 operational manual: Guidelines for classifying national education programmes and related qualifications.* Paris: OECD Publishing.

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.

Rattinger, H., Roßteutscher, S., Schmitt-Beck, R., Weßels, B., Wolf, C., Bieber, I., & Scherer, P. (2014). *Langfrist-online-tracking, T19 (GLES).* GESIS Datenarchiv, Köln. ZA5719 Datenfile Version 2.0.0.

Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research*, *43*(1), 73–97.

Rozin, P. & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320.

Saris, W. E. (1990). The choice of a model for evaluation of measurement instruments. In W. E. Saris & A. V. Meurs (Eds.), *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies.* (pp. 118–133). Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Saris, W. E. & Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement errors in surveys.* (pp. 575–599). New York: Wiley.

Saris, W. E., Oberski, D., Revilla, M. A., Zavala, D., Lilleoja, L., Gallhofer, I., & Gruner, T. (2011). The development of the program sqp 2.0 for the prediction of the quality of survey questions. RECSM Working Paper Number 24. Universitat Pompeu Fabra - Research and Expertise Centre for Survey Methodology. Retrieved from https://www.upf.edu/documents/3966940/3986764/RECSM%5C_wp024.pdf

Schmitt, N. & Stuits, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373.

Shamon, H. (2014). Ist mein Einkommen gerecht? *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *66*(3), 397.

Shamon, H. & Dülmer, H. (2014). Raising the question on 'who should get what?' again: On the importance of ideal and existential standards. *Social Justice Research*, *27*(3), 340–368.

Shepelak, N. J. & Alwin, D. F. (1986). Beliefs about inequality and perceptions of distributive justice. *American Sociological Review*, *51*(1), 30–46.

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

Tversky, A. & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. *The Quarterly Journal of Economics*, *106*(4), 1039–1061.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of rdd telephone surveys and internet surveys conducted with probability and nonprobability samples. *Public Opinion Quarterly*, *75*(4), 709–747.

Zack, S., Elizabeth, Kennedy, J., & Long, J. S. (2019). Can nonprobability samples be used for social science research? a cautionary tale. *Survey Research Methods*, *13*(2), 215–227.

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
Tables

Table A1
Example vignette of the study

The person H. O. (industrial sales representative, aged 35 years, married, spouse does not work) has the further characteristics	
Gender	Male
Own children	2
Job experience	5 years
Effort at work	High
In the region in which H. O. works, the following is true of industrial sales representatives	
Highest gross salary	€ 5,400
Average gross salary	€ 3,009
Lowest gross salary	€ 1,600

Table A2
Study 1, Setting 1

	Reliability r^2			Validity v^2			Quality q^2		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Method 1	0.677	0.939	0.823	0.354	0.261	0.224	0.240	0.245	0.184
Method 2	0.817	1.000	0.856	0.998	0.696	0.945	0.816	0.696	0.808
Method 3	0.865	0.916	0.769	0.814	0.752	0.922	0.704	0.688	0.709

T1= just earning level, T2= earning level for a decent life, T3= earning level for adequate societal participation.

Table A3
Study 1, Setting 2 (with explanation)

	Reliability r^2			Validity v^2			Quality q^2		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Method 1	0.572	0.837	0.554	0.563	0.420	0.564	0.321	0.352	0.312
Method 2	0.706	1.000	0.843	0.992	0.531	0.972	0.700	0.531	0.819
Method 3	0.787	0.755	0.972	0.927	0.872	0.789	0.730	0.659	0.767

T1= just earning level, T2= earning level for a decent life, T3= earning level for adequate societal participation.

Table A4
Study 1, Setting 3 (without explanation)

	Reliability r^2			Validity v^2			Quality q^2		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Method 1	0.440	0.731	0.805	0.469	0.425	0.246	0.206	0.311	0.198
Method 2	0.681	0.885	0.978	0.918	0.986	0.857	0.625	0.873	0.839
Method 3	0.850	1.000	0.876	0.933	0.605	0.774	0.793	0.605	0.678

T1= just earning level, T2= earning level for a decent life, T3= earning level for adequate societal participation.

Table A5
Study 2, Setting 1

	Reliability r^2			Validity v^2			Quality q^2		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Method 1	0.510	0.666	0.510	0.729	0.664	0.490	0.372	0.442	0.250
Method 2	0.658	1.000	0.716	0.976	0.826	0.990	0.642	0.826	0.709
Method 3	0.852	0.861	0.978	0.951	0.859	0.922	0.810	0.740	0.901

T1= just earning level, T2= earning level for a decent life, T3= earning level for adequate societal participation.

Table A6
Study 2, Setting 2 (with explanation)

	Reliability r^2			Validity v^2			Quality q^2		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Method 1	0.526	0.746	0.968	0.834	0.533	0.469	0.438	0.398	0.454
Method 2	0.897	1.000	0.814	0.990	0.801	0.992	0.888	0.801	0.807
Method 3	0.937	0.914	0.872	0.889	0.731	0.821	0.833	0.668	0.716

T1= just earning level, T2= earning level for a decent life, T3= earning level for adequate societal participation.

Table A7
Study 1, Setting 2 (with explanation) filtered

	Reliability r^2			Validity v^2			Quality q^2		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Method 1	0.599	0.773	0.506	0.783	0.729	0.701	0.469	0.563	0.354
Method 2	0.773	1.000	0.817	0.996	0.457	0.931	0.770	0.457	0.761
Method 3	0.876	0.876	0.903	0.815	0.908	0.878	0.714	0.796	0.792

T1= just earning level, T2= earning level for a decent life, T3= earning level for adequate societal participation.

Table A8
Study 1, Setting 3 (without explanation) filtered

	Reliability r^2			Validity v^2			Quality q^2		
	T1	T2	T3	T1	T2	T3	T1	T2	T3
Method 1	0.392	0.723	0.728	0.643	0.646	0.383	0.252	0.467	0.279
Method 2	0.672	0.884	0.994	0.914	0.986	0.839	0.615	0.871	0.834
Method 3	0.835	1.000	0.872	0.941	0.601	0.741	0.786	0.601	0.647

T1= just earning level, T2= earning level for a decent life, T3= earning level for adequate societal participation.

Table A9
OLS estimation results on proportion item nonresponse, careful-read -item scale, and on careful-read instruction scale in Study 1

	Proportion item nonresponse	Careful-read item scale	Careful-read instruction scale
Setting 1 (ref. categ.)	-	-	-
Setting 2	−0.004	0.355***	0.090
	(0.007)	(0.105)	(0.188)
Setting 3	−0.002	0.002	−0.093
	(0.007)	(0.117)	(0.189)
18 to 29 (ref. categ.)	-	-	-
30 to 39 years	0.011	0.135	0.345
	(0.009)	(0.169)	(0.249)
40 to 49 years	−0.005	0.433**	0.294
	(0.008)	(0.14)	(0.227)
50 to 59 years	−0.003	0.633***	0.736**
	(0.008)	(0.143)	(0.240)
> 60 years	−0.008	0.646***	0.931***
	(0.009)	(0.157)	(0.257)
Male (ref. categ.)	-	-	-
Female	−0.001	0.126	−0.013
	(0.005)	(0.089)	(0.153)
ISCED 1 & 2 & 3	0.013	−0.018	0.218
	(0.007)	(0.114)	(0.191)
ISCED 4 & 5 (ref. categ.)	-	-	-
ISCED 6 & 7 & 8	−0.006	−0.104	−0.162
	(0.007)	(0.106)	(0.185)
Constant	0.024**	4.613***	3.659***
	(0.008)	(0.151)	(0.233)
R^2	2.3%	9.2%	4.9%
n	544	524	521

Directed hypotheses tested with a one-tailed test only for Setting 3 and Setting 1. Standard errors in parentheses. Test on heteroscedasticity was performed. Heteroscedastic robust standard errors for careful-read item scale. Differences in observational basis are due to missing values on dependent variables.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table A10
OLS estimation results on proportion item nonresponse, careful-read item scale, and on careful-read instruction scale in Study 2

	Proportion item nonresponse	Careful-read item scale	Careful-read instruction scale
Setting 1 (ref. categ.)	-	-	-
Setting 2	0.003	0.312**	0.154
	(0.003)	(0.110)	(0.229)
18 to 29 (ref. categ.)	-	-	-
30 to 39 years	0.01*	−0.077	−0.256
	(0.004)	(0.151)	(0.306)
40 to 49 years	0.000	−0.133	−0.001
(0.005)	(0.165)	(0.351)	
50 to 59 years	0.011*	0.124	0.091
	(0.005)	(0.162)	(0.349)
> 60 years	0.003	0.173	0.105
	(0.006)	(0.193)	(0.432)
Male (ref. categ.)	-	-	-
Female	0.007*	−0.109	0.241
	(0.003)	(0.111)	(0.230)
else	−0.002	0.729***	−3.493***
	(0.03)	(0.175)	(0.419)
ISCED 1 & 2 & 3	−0.008	0.105	−0.475
	(0.016)	(0.457)	(1.050)
ISCED 4 & 5 (ref. categ.)	-	-	-
ISCED 6 & 7 & 8	−0.004	−0.129	−0.121
	(0.005)	(0.161)	(0.354)
Constant	0	4.914***	3.354***
	(0.006)	(0.196)	(0.424)
R^2	4.3%	4.3%	1.7%
n	334	334	330

Directed hypotheses tested with a one-tailed test only for Setting 3 and Setting 1. Standard errors in parentheses. Test on heteroscedasticity was performed. Heteroscedastic robust standard errors for careful-read item scale. Differences in observational basis are due to missing values on dependent variables.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table A11
Biserial correlations between failures of respective attention checks and "careful-read-item-scale"

	Correlation with "careful-read-item-scale"			
	Study 1 & Study 2	Study 1 Setting 2	Study 1 Setting 3	Study 1 Setting 2
Attention check instruction 1 failed	−0.02	−0.02	−0.02	−0.05
Attention check item 1 failed	−0.21***	−0.33***	−0.22**	-
Attention check item 2 failed	−0.28***	−0.30***	−0.39***	0.00[a]
Attention check instruction 2 failed	−0.06	−0.10	−0.06	−0.10

Directed hypotheses tested with a one-tailed test.
[a] Not reported because only one person failed attention check item 2.
* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$

Table A12

Biserial correlations between failures of respective attention checks and "careful-read-instruction-scale"

	Correlation with "careful-read-instruction scale"			
	Study 1 & Study 2	Study 1 Setting 2	Study 1 Setting 3	Study 1 Setting 2
Attention check instruction 1 failed	−0.02	0.09	−0.03	−0.17*
Attention check item 1 failed	0.03	0.07	−0.05	-
Attention check item 2 failed	0.00	0.05	−0.11	0.06[a]
Attention check instruction 2 failed	−0.10*	−0.14*	−0.02	−0.19**

Directed hypotheses tested with a one-tailed test.

[a] Not reported because only one person failed attention check item 2.

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$