Methods in Psychology 2 (2020) 100022

Contents lists available at ScienceDirect

Methods in Psychology

journal homepage: www.journals.elsevier.com/methods-in-psychology

The quality of data collected online: An investigation of careless responding in a crowdsourced sample

Florian Brühlmann *, Serge Petralito, Lena F. Aeschbach, Klaus Opwis

Center for General Psychology and Methodology, University of Basel, Missionsstrasse 62a, CH-4055, Basel, Switzerland

ARTICLE INFO

Keywords:
Careless responding
Crowdsourcing
Survey
Response patterns
Inattentive responding
Open answer
Latent profile analysis

ABSTRACT

Despite recent concerns about data quality, various academic fields rely increasingly on crowdsourced samples. Thus, the goal of this study was to systematically assess carelessness in a crowdsourced sample (N = 394) by applying various measures and detection methods. A Latent Profile Analysis revealed that 45.9% of the participants showed some form of careless behavior. Excluding these participants increased the effect size in an experiment included in the survey. Based on our findings, several recommendations of easy to apply measures for assessing data quality are given.

1. Introduction

Online surveys have become a standard method of data collection in various fields, such as in recent psychological research (Gosling and Mason, 2015) and market research (Comley, 2015). Whereas in 2003 and 2004 only 1.6% of articles published in APA journals used the Internet (Skitka and Sargis, 2006), Gosling and Mason (2015) stated just a few years later that "studies that use the Internet in one way or another have become so pervasive that reviewing them all would be impossible" (p. 879). Moreover, this method covers virtually all areas of psychology.

One of the most popular recruitment methods for participants in online studies for psychological research is the use of crowdsourcing services, such as Amazon's Mechanical Turk (MTurk) or FigureEight (formerly known as CrowdFlower). Regarding MTurk, approximately 15′000 published articles used this crowdsourcing platform between 2006 and 2014 for their data collection (J. Chandler and Shapiro, 2016; Kan and Drummey, 2018). On these platforms, various small tasks are offered in exchange for money to "crowd workers".

As a primary advantage, crowdsourcing platforms offer a more diverse population compared to typically homogenous samples from psychological studies (Kan and Drummey, 2018): In the case of MTurk, these workers are composed of a demographic containing more than 500′000 individuals from 190 countries (Paolacci and Chandler, 2014). While concerns considering the generalizability and validity of crowdsourced online samples have been discussed (Kan and Drummey, 2018), Gosling and Mason (2015) also reported that the mean and range of ages from an MTurk-sample are more representative of the general US population than a sample merely consisting of undergraduate students. Moreover, in comparison to online samples recruited on social media platforms, some crowdsourced samples were found to have a higher diversity in terms of age, cultural, and socioeconomic factors (Casler et al., 2013), and more balanced gender ratios (de Winter et al., 2015). Online data collection, further, has numerous advantages over laboratory studies in convenience: lower infrastructure costs faster and cheaper data collection (Casler et al., 2013; de Winter et al., 2015), more extensive distribution of the study, and lower hurdles for participation (Kan and Drummey, 2018).

Given the increased distance between researchers and participants in online studies, and the possible influence of distractions in an uncontrolled setting, data collected online may suffer from bad quality stemming from carelessness or different forms of deceptive behavior (Cheung et al., 2017; Fleischer et al., 2015). Participants may further have their deceptive tendencies exacerbated by the incentive-structure of crowdsourcing platforms (J. Chandler, Mueller and Paolacci, 2014; J. Chandler, Paolacci, Peer, Mueller and Ratliff, 2015; Kan and Drummey, 2018; Peer et al., 2017; Stewart et al., 2017). However, in surveys with Likert-type items, for example measuring attitudes and personality attributes, it may be challenging to identify clearly intentional deceptive behavior due to the lack of an identifiable 'correct' answer. Therefore, we decided on participant carelessness as the major influencing factor of the quality of

* Corresponding author. Tel.: + 41 (0)61 207 06 66.
E-mail addresses: florian.bruehlmann@unibas.ch (F. Brühlmann), s.petralito@unibas.ch (S. Petralito), lena.aeschbach@unibas.ch (L.F. Aeschbach), klaus.opwis@unibas.ch (K. Opwis).

https://doi.org/10.1016/j.metip.2020.100022
Received 23 April 2019; Received in revised form 18 September 2019; Accepted 23 March 2020
Available online 6 April 2020

data collected in online surveys on crowdsourcing services as the primary focus of this study, in line with previous research (Kam and Meyer, 2015; Niessen et al., 2016). Participant carelessness in this context refers to "content nonresponsivity", meaning we assume careless respondents give answers of bad data quality regardless of the content of the questions (Nichols et al., 1989).

Participant carelessness in surveys (Meade and Craig, 2012), has recently received increased attention from various researchers regarding their reasons, effects, detection, and prevention (Maniaci and Rogge, 2014; Meade and Craig, 2012; Niessen et al., 2016), partly due to the aforementioned increased distance between researchers and participants. Indeed, within psychology the clear and accelerated trend to bring small scale experiments from the controlled context of a laboratory into the crowd brings into questions what challenges researchers will face and how they may do so. Gadiraju et al. (2017) gives a systematic overview of challenges, risks and opportunities to take into account in this transformation of data acquisition. Methods of data quality are mentioned as a means to exert control over the crowdworkers, but not discussed in further detail. This paper will examine these methods as part of the recommended careful design process in detail to give further recommendations for research.

1.1. Causes and effects of carelessness

In the present study we primarily focus on participant careless responding. In line with other work (e.g., Meade and Craig, 2012), we include possibly short-term inattention in our definition of carelessness. Other forms of invalid responding and the aforementioned deceptive behavior (such as social desirability and faking responses), also decrease data quality, but may have different causes and effects, which make it difficult for assessment in Likert-type questionnaires (Maniaci and Rogge, 2014; McKay et al., 2018). Participant inattention might have many sources, one possibility being the anonymity of computer-based surveys, which can result in a lack of accountability (Douglas and McGarty, 2001; Lee, 2006; Meade and Craig, 2012). Further important factors affecting carelessness in survey data are respondent motivation and interest, length of survey, social contact, and environmental distraction (Meade and Craig, 2012). Extrinsic motivation might also account for carelessness, such as when participants are paid for their answers. Gadiraju, Kawase, Dietze, and Demartini (2015) found that some participants recruited via crowdsourcing services employ strategies to minimize their invested time or effort in return for participation compensation. In these cases, careless responding and subsequent poor data quality emerge from crowdworkers who are solely interested in receiving their payment as fast as possible without providing valid data for the researcher. Furthermore, Niessen et al. (2016) observed that students also strove to complete surveys as quickly as possible in exchange for course credits. Aside from these external factors, a study conducted by Kazai et al. (2011) showed that the quality of the crowd workers' responses is related to personality traits. Furthermore, McKay et al. (2018) found that careless responding is strongly related to malevolent personality traits. This indicates that careless responding is influenced by a participant's personality and self-interest motivation. In summary, we will use carelessness as general higher-level category that include different and more specific phenomena like inattentiveness but also amotivation.

Recently, most of the attention of empirical research has been given to the discovery of said carelessness (see Curran, 2016, for a review). The screening methods can be divided into two groups. The first group is the planned implementation of special items or scales to screen carelessness. For example, Bogus Items (Meade and Craig, 2012), Instructed Response Item (IRI) (Curran, 2016), and Instructional Manipulation Checks (Oppenheimer et al., 2009). The second group of detection methods can be described as post hoc measures. These include the examination of response time, multivariate outliers, and (in-) consistency indices. These do not require special items, but an elaborate analysis after data collection. There are a variety of different methods available, but we will focus on those recommended in the recent literature (Curran, 2016).

None of the above-mentioned detection methods are applicable to data which is not quantitative or if measures differ from Likert-type scales. Conducting surveys that collect quantitative and qualitative data, such as when applying critical incident technique, are common in many research areas (e.g., user experience research; Tuch et al., 2016). Qualitative data, such as experience reports, is comparatively easy to screen for careless responding, because low-effort responses are easier to spot than in responses to Likert-type scales and offer a high face validity. Thus, including data quality measures based on qualitative responses offers a different perspective on the phenomenon of carelessness. The possibility to analyze the relationship between quantitative and qualitative indicators of carelessness allow for some kind of convergent validation of different methodological approaches.

Base rate estimates for bad online data quality stem from different concepts of invalid responding and different sources for online data collection: Recent research has estimated that, depending on the method, between 10% and 12% of participants in an online survey exhibit an answering behavior described as insufficient effort responding or careless responding (Meade and Craig, 2012). In a more heterogeneous sample, Maniaci and Rogge (2014) found that between 3% and 9% of participants respond carelessly. In another study assessing carelessness in a crowdsourced sample, Peer et al. (2017) found that only 27% of all participants in a FigureEight-sample passed all attention checks, and approximately 18% failed in all of them. While these numbers provide some valuable insights for assessing careless behavior in a crowdsourced sample, the study did not include other carelessness measures. Therefore, it only identified one behavioral form of inattention or carelessness.

Consequently, these alarmingly high estimates for bad data quality stemming from carelessness or other deceptive forms of behavior vary greatly between studies, methods, and recruitment methods. However, even a seemingly small number of careless responses can have serious consequences, such as failed replications (Oppenheimer et al., 2009) or false-positives (Huang et al., 2015). Furthermore, careless responding may cause failed manipulations when instructions are not carefully read (Maniaci and Rogge, 2014), lower internal consistency of validated scales (Maniaci and Rogge, 2014), and problems in questionnaire development and item analysis (Johnson, 2005). Additionally, it can lead to problems in investigating questionnaire dimensionality (Kam and Meyer, 2015). All the aforementioned research examined academic participant pools or mixed types of online data (e.g., Maniaci and Rogge, 2014; Meade and Craig, 2012), or the studies only applied one measure to determine carelessness in a crowdsourced sample (Dogan, 2018; Hauser and Schwarz, 2016; Peer et al., 2017). An integrated analysis of careless behavior on crowdsourced platforms using various carelessness detection methods is still missing.

2. Aim of the present study

The aim of the present study was to analyze the data quality of a crowdsourced online sample, based on various recommended methods for assessing careless behavior. This addresses the limited variety of methods used in existing research about carelessness in surveys on crowdsourcing platforms.

2.1. Research question 1

How are the different planned and post hoc measures of carelessness related?

Many detection methods of careless responding are still not fully understood in their performance and their relationship with each other. Further, in this study we utilized Resampled Individual Reliability (RIR) and Person-Total Correlation (PTC), which are described in Curran, 2016, but, to our knowledge, have never before been implemented and assessed empirically. We also implemented an open-answer questions of

which we assessed the respondents answer quality and compared this to the other methods of detecting careless respondents. To gain a further understanding of the performance of the various measures of careless respondents we assessed the correlation between all measures.

2.2. Research question 2

How prevalent is careless responding in samples from crowdsourcing platforms based on various detection methods for carelessness?

An important factor in planning out surveys is the prevalence of careless respondents, as researchers will have to account for surplus respondents to avoid ending up with insufficient sample size for analysis. Therefore, the second aim of this study was to gain an estimate of the prevalence of careless respondents in a crowdsourced sample. For a comprehensive understanding of the rate of carelessness, we aimed to examine detection methods of careless responding independently from, as well as in combination with, each other. Additionally, we followed the approach by Meade and Craig, 2012 in applying Latent Profile Analysis (LPA) with the goal to identify different types of respondents and gain an estimate based on a statistically meaningful combination of carelessness detection methods.

2.3. Research question 3

What are the consequences of the exclusion of the respondents flagged as careless?

We aim to understand how excluding careless respondents would affect the quality of the data we analyze. We therefore included an experiment which has been shown to produce significant results in previous research (Rieser and Bernhard, 2016) and analyzed the consequences (e.g. results of significance tests as well as effect sizes) between inclusion and exclusion of participants identified as careless.

2.4. Research question 4

Based on our sample, which are efficient detection methods of careless respondents to recommend?

As we identify the amount of careless respondents in our sample based on multiple indicators, we aim to understand how these measures are interrelated. Therefore we assessed the correlation between all measures. As using many different measures of careless respondents is not economical in the long run and could introduce false flags at a higher rate, we aim to find the most valid detection methods for a recommendation for future researchers. Therefore we followed the approach by Meade and Craig (2012) in applying Latent Profile Analysis (LPA) with the goal to identify different types of respondents and gain an estimate based on a statistically meaningful combination of carelessness detection methods.

3. Method

To investigate the aforementioned research questions we designed a between-subject design with two experimental groups. We collected both qualitative and quantitative data, this was done to test different methods to detect careless respondents and to compare the different methods.

3.1. Data collection

The present study was conducted using a crowdsourced sample from FigureEight (Crowdflower). Especially outside the U.S., FigureEight is a viable choice for crowdsourcing, as Amazon's MTurk has for a long time required requesters to have a US-address. FigureEight is accessible from Europe and other regions outside the USA, and provides access to millions of contributors (Van Pelt and Sorokin, 2012). The crowdsourcing platform is a well-established tool to gather participants for online-surveys, as shown by over 5600 hits on Google Scholar (August 23, 2019, Keyword: CrowdFlower).

The eligibility requirements on FigureEight were set to only include workers from the US with a qualification level of at least level 1, meaning the crowdworkers only had to pass a short exam before participating in the study, but not build any reputation with other researchers by providing high quality data. The default settings for the contributor channels were retained and participants were given a compensation of $0.60 with an estimated completion time of 10 min. All participants received this compensation regardless of their answer quality or values on the different carelessness methods.

Data and analysis code used in this study is available at https://osf.io/9vjur/.

3.2. Participants

A total of 394 participants completed all parts of the study and were included in the analysis. Most participants were female (36.5% male and 1% non-binary or not specified), employed (56%) and on average 36.5 years old ($Md = 34$, $SD = 12.66$, Range $18 - 78$). Although only Americans were allowed to participate, a few participants had their primary residence outside the USA (12; 3%) and a mother tongue other than English (21; 5.3%, most frequently Spanish).

3.3. Procedure

After providing consent, participants were asked to recall a recent negative experience with an online store. In particular, participants were asked to respond to two questions 1) what exactly caused this experience to be negative and 2) how this affected their online shopping habits. We instructed participants to respond in free text with as much detail as possible, with complete sentences, and with at least 50 words. Next, 10 items of the positive and negative affect schedule (PANAS; Watson et al., 1988), 23 items of the AttrakDiff2 (Hassenzahl et al., 2003), and 24 items measuring psychological need satisfaction adapted from Sheldon et al. (2001) were presented. This type of critical incident method is a common procedure in user experience research (e.g., Tuch et al., 2016). After this first block of questions, participants were randomly allocated to be shown either a high trust or low trust mockup of a website. The website was manipulated according to the trust supporting elements identified by Seckler et al. (2015). This setting was chosen to conduct a plausible experiment in user experience research that was thematically related to the rest of the study. After this, participants were asked to complete 16 items of a Likert-type scale for trust in websites (Flavián et al., 2006). The goal of this section was to examine the effects of excluding data from careless participants on effect sizes and p-values in a group comparison. On the next page, participants rated the visual aesthetics of the website mock-up with 18 items (VisAWI, Moshagen and Thielsch, 2010)). Following this section, the big five personality types were assessed with 44 items of the Big Five Inventory (BFI) (John and Srivastava, 1999). All post hoc detection methods of carelessness were investigated using this scale. On the last page of the survey, participants completed demographic information and a scale on self-reported careless responding (as in Maniaci and Rogge, 2014) and a self-reported single item (SRSI UseMe) (Meade and Craig, 2012). Finally, all participants were given a completion code.

3.4. Measures

All post hoc detection methods of carelessness were applied to the 44 items of the BFI in the last part of the questionnaire. We decided to focus on the BFI because it is multidimensional with a sufficient length to calculate various indices, and it is comparable with other studies in this field (Johnson, 2005; Maniaci and Rogge, 2014; Meade and Craig, 2012). The data of the other questionnaires used in this study were not subject to further analysis except for the trust scale by Flavián et al. (2006), which was used as a dependent variable in the experiment.

3.5. Planned detection methods

The wording of the self-reported responding tendencies scale, the Bogus Item, and the Instructed Response Item (IRI) incorporated in the study is presented in Table 1.

3.5.1. Self-reported responding tendencies

Following demographic questions, ten items based on Maniaci and Rogge (2014) were used to assess general tendencies in responding. Although excluding participants based on self-reported responding tendencies has been found to improve data quality significantly (Aust et al., 2013), these items are also easily detected and prone to manipulation and dishonest answers. Three items were used to measure self-reported careless responding ($\alpha = .84$), two items to measure self-reported patterned responding ($\alpha = .88$), three items to assess self-reported rushed responding ($\alpha = .83$), and two items assessing self-reported skipping of instructions ($\alpha = .68$). All items are presented in Table 1. Items were rated on a 7-point scale (1 = never, 4 = approximately half the time, 7 = all of the time), and responses were averaged ensuring high scores reflected more problematic responding.

Applying the cutoff used by Maniaci and Rogge (2014), answers of 4 or higher were flagged. Additionally, self-indicated data usage was assessed using the SRSI UseMe.

Table 1
The items of the self-reported responding tendencies scale (Maniaci and Rogge, 2014) and planned detection items included in the study. Self-report answer options ranged from 1 (never), over 4 (about half of the time) to 7 (all the time). The Bogus Item was included in the BFI where answers between 1 (disagree strongly) and 5 (agree strongly) were possible. The IRI was included in the trust scale that was used as the dependent variable of the experiment.

Measure		Item
Self-report		[How often do you …]
	Careless responding	
		1. Read each question
		2. Pay attention to every question
		3. Take as much time as you need to answer the questions honestly
	Patterned responding	
		4. Make patterns with the responses to a block of questions
		5. Use the same answer for a block of questions one the same topic [rather than reading each question]
	Rushed responding	
		6. Answer quickly without thinking
		7. Answer impulsively without thinking
		8. Rush through the survey
	Skipping of instructions	
		9. Skim the instructions quickly
		10. Skip over parts of the instruction
SRSI UseMe		In your honest opinion, should we use your data in our analyses in this study? (Do not worry, this will not affect your payment, you will receive the payment code either way.)
Bogus Item		[I see myself as someone who …] Did not read this statement
Instructed Response Item		I read instructions carefully. To show that you are reading these instructions, please leave this question blank.

3.5.2. Attention checks

We employed two attention check items in the questionnaire following the Infrequency Approach (Huang et al., 2012), which entails including items to which all careful respondents should respond to in the same, or similar, fashion. One measure we applied was the Bogus Item similar to Meade and Craig (2012), which are items that are very unlikely for participants to agree with. The Bogus Item was located within the BFI (see Table 1). Participants who did not select "strongly disagree" or "slightly disagree" were thus flagged as inattentive. The other attention check item was an Instructed Response Item (IRI) similar to Meade and Craig (2012) and Curran (2016). According to Meade and Craig (2012) the IRI has several advantages over Bogus Items, as they are easier to create, have a singular correct answer, and therefore provide an obvious metric for scoring. Furthermore, they offer a clear interpretation and are not prone to humorous answers, which is a problem with the Bogus Item. The IRI (see Table 1) was placed within the items of the trust scale by Flavián et al. (2006). Participants who nevertheless answered this question were flagged.

3.6. Post hoc detection methods

3.6.1. Response time

One simple post hoc measure to assess careless responding is to measure participant overall response time. The concept is that inattentive or careless respondents will be noticeable through unusually short or long completion times. Although this measure is easily applicable in any online survey, the issue of what constitutes an acceptable range of completion times must be decided individually for each question (Curran, 2016).

3.6.2. LongString index

The LongString Index acts as an invariability measure, which assesses the number of same answers given in sequence. Careless participants who might select the same answer for equal or greater than half the length of the total scale will be excluded from the sample (Curran, 2016; Huang et al., 2012). Curran (2016) recommended LongString analysis to identify some of the worst respondents that would otherwise be missed, although the measure can easily be deceived. The LongString Index in this study was calculated for the BFI following the procedure described in Meade and Craig (2012).

3.6.3. Odd-Even Consistency

To assess the Odd-Even Consistency (OEC), each individual's responses on each unidimensional subscale are split into responses to even and to uneven items (Curran, 2016). In the present case, this was implemented for each of the five dimensions of the BFI (Openness, Extraversion, Agreeableness, Conscientiousness, and Neuroticism). Reverse coded items must be recorded before calculating this measure. The responses to the even and uneven items are then averaged separately, ensuring each participant receives a score based on the even and the uneven items for each subscale of one larger scale. The individual correlation of these two vectors acts as a score of consistency. An important limitation is that this correlation is constrained by the number of subscales and the number of items in a scale. The OEC in this study was assessed for the BFI based on the procedure described by Meade and Craig (2012). Following Curran's (2016) recommendation, any correlation below 0 or undefined (such as the same answer for all items) was flagged.

3.6.4. Resampled Individual Reliability

Curran (2016) proposed a more general conceptualization of the OEC measure – Resampled Individual Reliability (RIR). Here, the basic concept is that items that should measure the same construct should correlate positively within individuals. However, instead of limiting this idea to odd and even items, Curran (2016) suggests creating two halves of each subscale randomly without replacement. The individual

correlation of these two vectors acts as a score of consistency. This process is then repeated several times (resampling). This is a new measure that was included in the present study and, to the best of our knowledge, has never been empirically examined. Following Curran's (2016) recommendation for the OEC, any correlation below 0 or undefined was flagged.

3.6.5. Person-Total Correlation

The measure of Person-Total Correlation (PTC) describes the correlation of a participant's answers to each of the items of a scale, with the means of these items based on the whole sample (Curran, 2016). This measure relies on the assumption that a large majority of the sample responded attentively, thus this measure may be problematic in situations where a large number of careless respondents are expected. Because this measure has currently not been empirically examined, no widely accepted cutoff value for this correlation exists. However, as recommended by Curran (2016), participants with a negative or undefined PTC were flagged.

3.7. Open answer quality

A priori criteria for the quality rating of open answers predominantly originates from the studies conducted by Holland and Christian (2009) and Smyth et al. (2009). The following indicators for calculating an open answer quality index were taken into consideration: 1. Whether participants provided a thematically substantive response. 2. If a minimum of 50 words was provided as instructed. 3. If participants provided answers in complete sentences as instructed. 4. The number of subquestions answered as instructed. 5. The number of subquestions further elaborated. A detailed description of how the open answer quality index was created is presented in the Appendix. The third author coded all experience reports. To ensure inter-rater reliability, the second author coded a random subset of 100 open-ended answers. Because two fixed raters rated a randomly selected subset, ICC3 was used (Koo and Li, 2016). Inter-rater agreement of each category was between moderate (Complete Sentences, $ICC3 = .80$), good (Substantive Response, $ICC3 = .78$; Number of Subquestions Elaborated, $ICC = .84$) and excellent (Number of Subquestions Answered, $ICC3 = .94$). Inter-rater agreement for the overall answer quality index was excellent $ICC3 = .96$, with a 95% confidence interval from 0.94 to 0.97 ($F(99,99) = 50.63$, $p < .001$).

4. Results

In this section, we first report on each group of carelessness detection methods separately, and then investigate how they relate to answer quality. Table 2 presents an overview of the number of participants flagged by each method.

Table 2
Descriptive statistics for all detection methods used in the study. Self-report includes problematic responding tendencies as well as the SRSI UseMe item.

	Mean	SD	Min	Max	No. Flagged	%
Planned detection						
Self-report					106	26.90
Bogus Item					92	23.35
Instructed Response Item					96	24.37
Response time	16.71	9.22	3.93	61.15		
Post hoc detection						
LongString	6.63	9.15	0	44	25	6.35
Odd-Even Consistency	.61	.43	−1	1	63	15.99
Resampled Individual Reliability	.56	.39	-.82	.99	63	15.99
Person-Total Correlation	.38	.32	-.47	.88	74	18.78
Answer quality					100	25.38
Total (flagged by at least one method)					233	59.14

Note. Total.$N = 394$

4.1. Planned detection methods

4.1.1. Self-reported responding tendencies

Participants relatively frequently indicated that they engaged in problematic responding tendencies. Applying the cutoff used by Maniaci and Rogge (2014), answers with 4 or higher were flagged. Thus, we flagged 25 careless respondents (6.6%), 50 pattern-respondents (12.7%), 44 rushed-respondents (11.2%), and 65 (16.5%) participants for skipping instructions. As depicted in Fig. 1, skipping instructions was admitted most frequently ($M = 2.27$) followed by rushed responding ($M = 2.07$). However, there were fewer values of 4 and above for the rushed responding than for the patterned responding scale. Only 9 participants were flagged in every scale, 17 in 3 scales, 24 in 2 scales, and a majority of 49 participants were flagged in only 1 of the 4 self-reported scales. In total, the 4 scales flagged 99 (25.1%) participants as conspicuous.

The SRSI UseMe, indicating whether we should use the data provided by the participant or not, was negated by 22 participants (5.6%). Thus, these participants were also flagged as self-reported careless. It was then decided to aggregate these self-reported measures into one variable for self-reported carelessness.

Aggregating the self-reported measures of carelessness, patterned responding, rushed responding (*flagged* >= 4) and the SRSI UseMe, 106 participants (26.9%) were flagged as self-reported low-quality responses (see Table 2).

4.1.2. Attention checks

The IRI and the Bogus Item were missed by 96 participants (24.4%) and 92 participants (23.3%), respectively. Because there was no clear cutoff for the Bogus Item, we decided to code all answers with an agreement of 4 or higher to the item "*I see myself as someone who did not read this statement*" as failing to answer the Bogus Item correctly. Fig. 2 demonstrates that the majority of respondents (258, 65.5%) answered both items correctly, while 40 (10.2%) failed only at the Bogus Item, and 44 (11.2%) only at the IRI. However, a large number of participants who were flagged as inattentive missed both questions (52, 13.2%).

4.2. Post hoc detection methods

The boxplots and individual values of each post hoc detection method are presented in Fig. 3. Where applicable, cutoffs are indicated by a vertical line and flagged participants are marked red ("fail") and inconspicuous participants are marked blue ("pass").

4.2.1. Response time

The distribution of overall response time presented in Fig. 3 did not show a cluster or conspicuous responses below a certain value. Therefore, no suspiciously fast respondents were flagged.

4.2.2. LongString index

Results from the LongString analysis, with the recommended cutoff from Curran (2016) (More than half of the items, that is 22 for the BFI), reveal that 25 (6.3%) of the participants were flagged by this method. The distribution depicted in Fig. 3 displays that the vast majority of participants were significantly below this threshold, and 18 (4.6%) suspicious respondents with a LongString Index of 44 were identified with this method. These 18 participants provided the same answer for all 44 items of the BFI.

4.2.3. Odd-Even Consistency

The distribution in Fig. 3 is left-skewed with a long tail, and only a few suspicious correlations are close to −1. Curran (2016) recommended removing all correlations below 0, which in this case would flag 63 (16.0%) of participants as responding too inconsistently.

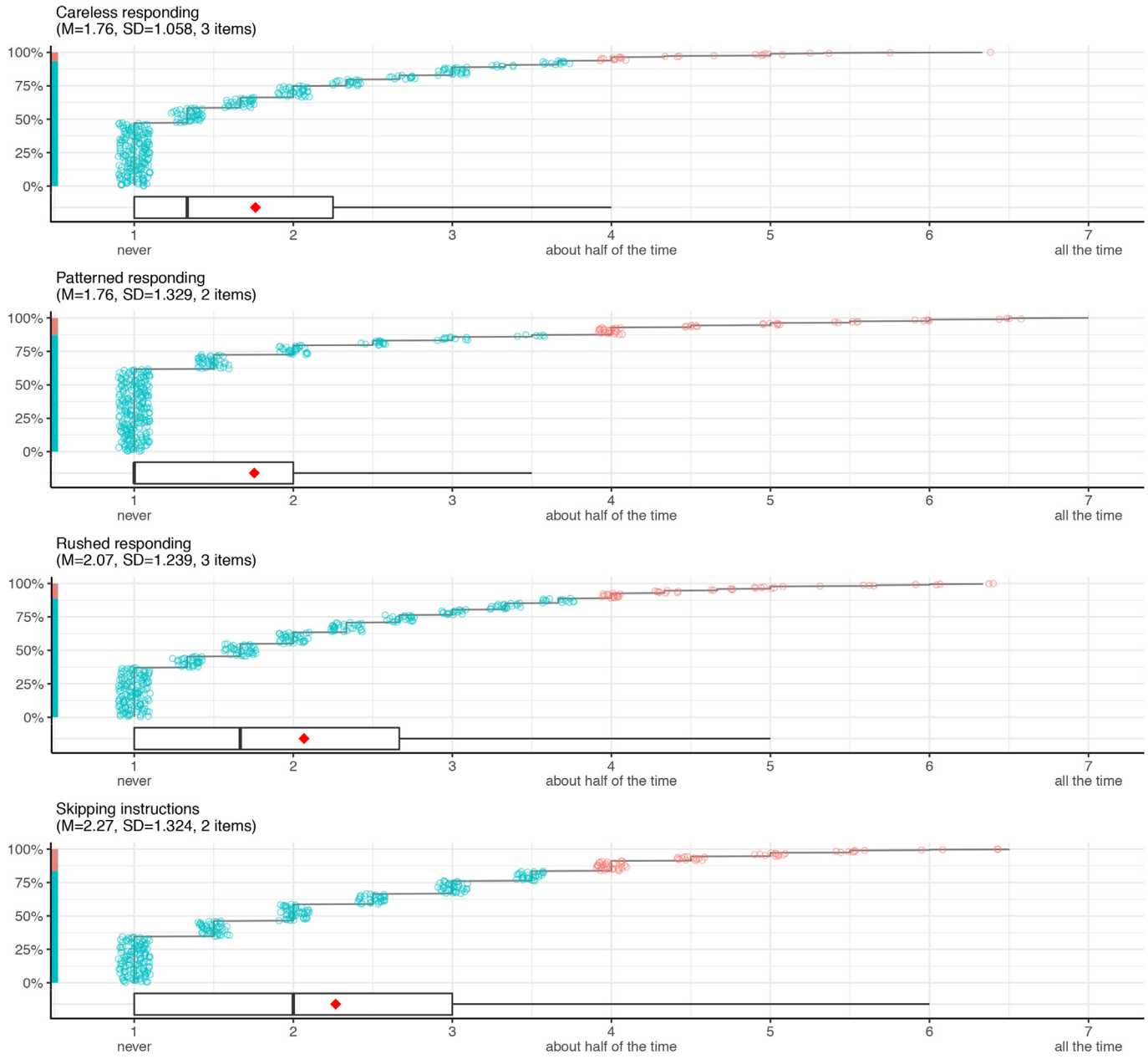

Fig. 1. Distributions of self-reported responding tendency scales. A random value was added to individual points to reduce overplotting.

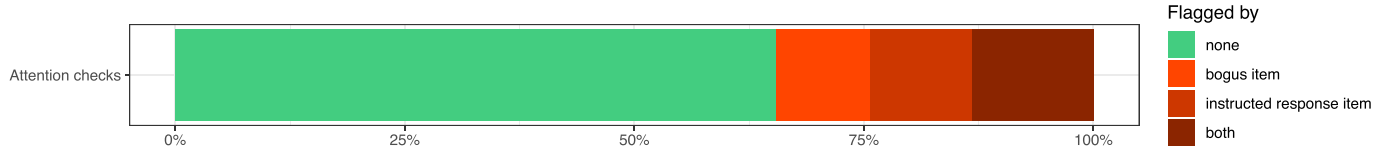

Fig. 2. Number of participants flagged by one or both attention check items.

4.2.4. *Resampled Individual Reliability*

As a more general approach to consistency than the OEC, RIR was calculated with 100 times randomly selected two halves of each subscale of the BFI. These two vectors were then correlated for each individual, giving a more general (resampled) reliability. As with the OEC, the distribution is left-skewed with a long tail (see Fig. 3). Although, it has less extreme negative values, the same amount of respondents were identified as careless with this method (63, 16.0%).

4.2.5. *Person-Total Correlation*

Correlations of individual answers with the mean of answers from the whole sample exhibited a comparatively narrow distribution of values (see Fig. 3). This method flagged 74 (18.8%) of participants as careless, indicated by a correlation of less than 0.

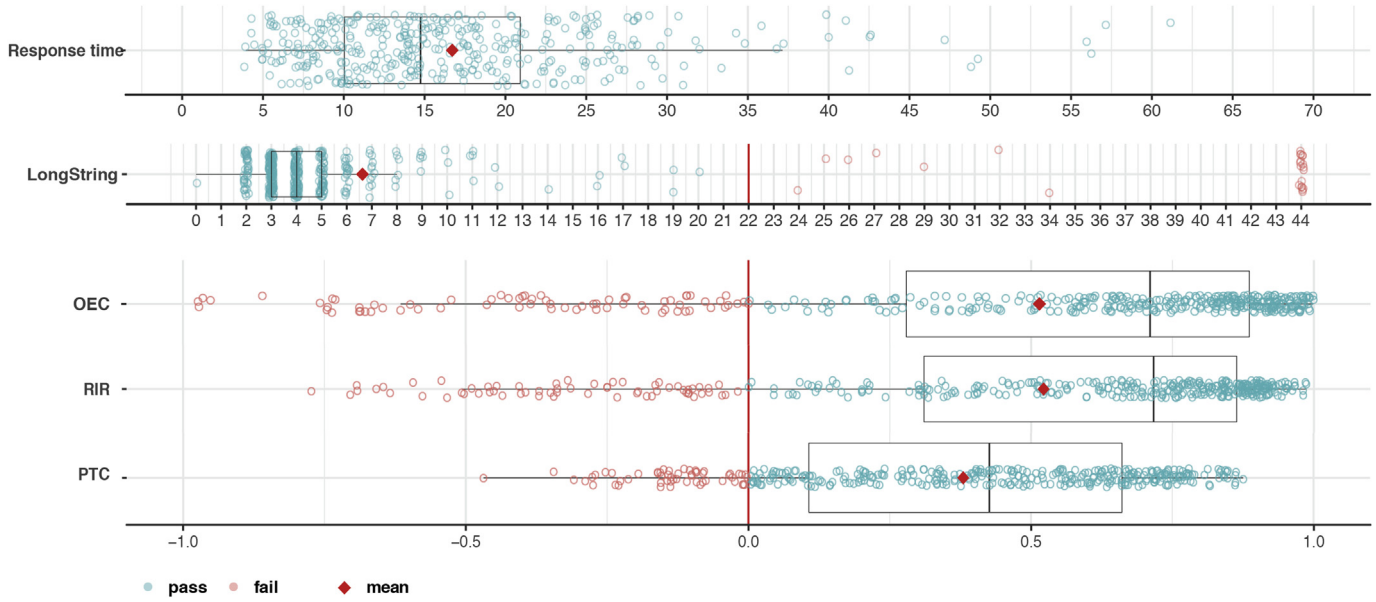

Fig. 3. Boxplots of carelessness detection methods. OEC = Odd-Even Consistency, RIR = Resampled Individual Reliability, PTC = Person-Total Correlation. Response time in minutes for the entire survey.

4.3. Open answer quality

Open answer quality was coded either 0 (= Insufficient), 1 (= High), or 2 (= Excellent). Of the full sample, 100 participants (25.4%) displayed insufficient answer quality in the open question. As we are mainly interested in whether participants failed or succeeded to provide sufficient open answer quality, the high (146, 37.1%) and excellent (148, 37.6%) open answer quality categories were combined for further analysis.

4.4. Correlations between carelessness detection methods

Table 3 depicts how successfully the different methods correlate in their decision to classify participants either as suspicious or not suspicious. All Matthews correlations (Powers, 2011) are moderate to highly positive. Answer quality achieved relatively low correlation with all behavioral and self-report measures of carelessness. The highest correlations of answer quality were observed with the Bogus Item (0.26) and the IRI (0.24). Interestingly, while the IRI and Bogus Item correlated with 0.41, the Bogus Item exhibited a higher correlation with the consistency measures PTC (0.52), RIR (0.51), LongString (0.37), and OEC (0.36). Self-reported data quality correlated substantially with RIR (0.38), the Bogus Item (0.34) and the IRI (0.30). Unsurprisingly, the highest correlation was observed between OEC and RIR (0.68), because RIR is a generalization of OEC. Overall, the correlation pattern demonstrates that among the attention check items the Bogus Item correlated more strongly with several other measures when compared to the IRI. The LongString Index exhibits similar correlations with all behavioral measures, except with IRI. The consistency measures correlate strongly with each other, apart from a relatively weak correlation between OEC and PTC (0.25). However, the relationship of answer quality with other measures is less clear. Based on these correlations, it is difficult to claim that one of the measures is redundant, as all the measures have relatively low overlap.

Table 3
Matthews correlation coefficient (MCC) of each measure pair (N = 394). IRI = Instruced Response Item, OEC = Odd-Even Consistency, RIR = Resampled Individual Reliability, PTC = Person-Total Correlation.

	1.	2.	3.	4.	5.	6.	7.
1. Self-report	–						
2. Bogus Item	.34	–					
3. IRI	.30	.41	–				
4. LongString	.24	.37	.26	–			
5. OEC	.23	.36	.20	.40	–		
6. RIR	.38	.51	.22	.37	.68	–	
7. PTC	.31	.52	.27	.35	.25	.41	–
8. Answer quality	.21	.26	.24	.21	.13	.22	.15

4.5. Classification of respondents based on different methods: Latent Profile Analysis

To identify different classes of carelessness, a Latent Profile Analysis (LPA) was conducted. The LPA is a flexible model-based approach to classification, with less restrictive assumptions than cluster analysis (Muthén, 2002). It aims to find the smallest number of profiles that can describe associations among a set of variables, and a formal set of objective criteria are applied to identify the optimal number of latent profiles in the data. For each participant, LPA provides a probability of membership, which is based on the degree of similarity with each prototypical latent profile. Following the approach by Meade and Craig (2012), we conducted an LPA on the non-self-report indicators of response quality (Open answer quality, response time, IRI, Bogus Item, LongString Index, OEC, RIR, and PTC) using the *mclust* package for R (Scrucca et al., 2016). Self-report indicators were excluded, enabling a comparison of our results with Meade and Craig (2012), and because these indicators might be biased when participants are paid to participate. However, the self-report indicators were subsequently used to describe the different classes found in our data.

Open answer quality, IRI, and Bogus Item were binary variables (pass/fail). Missing data was present because for participants with a LongString Index of 44 (all items with the same answer) no OEC, RIR, and PTC measures could be computed (no variance in the answers). We therefore inputted missing values in these variables with +1 for consistency and reliability and −1 for PTC. Missing values in the response time variable were possible if participants did not respond to the questionnaire in one sitting. These missing values were estimated using an expectation maximization algorithm as implemented in mclust. Based on these variables, multiple models with one to nine classes were fitted. Bayesian information criterion (BIC) and integrated complete-data likelihood (ICL) criterion were used to judge the most appropriate number of

Table 4
Descriptive statistics for each identified class of participants. IRI = Instruced Response Item, OEC = Odd-Even Consistency, RIR = Resampled Individual Reliability, PTC = Person-Total Correlation.

Variable	Class 1	Class 2	Class 3
Class size	181 (45.9%)	129 (32.7%)	84 (21.3%)
Percentages pass			
Answer quality (%)	44.75	100	100
Bogus Item (%)	49.17	100	100
IRI (%)	46.96	100	100
Self-report (%)	59.12	90.70	76.19
SRSI UseMe (%)	90.06	99.22	96.43
Means			
Response time (in Minutes)	14.58	16.94	22.03
OEC	.52	.86	.37
RIR	.44	.83	.43
PTC	.13	.59	.30
LongString	9.61	3.79	4.83
Means (Self-reported)			
Careless responding	2.16	1.36	1.52
Patterned responding	2.28	1.20	1.49
Rushed responding	2.43	1.68	1.88
Skipping instructions	2.53	1.99	2.13

classes. Both indicated that three classes were most appropriate (BIC: −7404.41, ICL: −7414.34). The class sizes were 181 (45.9%) for class 1, 129 (32.7%) for class 2, and 84 (21.3%) for class 3. The frequencies and variable means associated with each class are presented in Table 4.

As shown in Table 4, answers from class 1 were frequently judged as insufficient quality. Moreover, the attention check items were only missed by participants from this class. Further, class 1 participants more frequently self-reported bad quality than those in classes 2 and 3. Classes 1 and 2 responded significantly more quickly than class 3. Concerning OEC, class 3 provided more inconsistent answers than classes 1 and 2. Additionally, class 3 showed slightly stronger agreement to the self-reported responding tendencies than class 2. The defining hallmarks of class 1 were very large LongString Index values and very low PTC. This demonstrates that the consistency within participant answers was relatively high, while these answers were noticeably different from the total sample. Overall, it appears that a large part of class 1, which accounts for 45.9% of the sample, was responding in a careless way. However, class 1 cannot be described by one singular measure of carelessness. Instead, several forms captured by different methods should be included. In contrast, class 2 displayed the best values for all examined measures. Class 3 was slightly more conspicuous in terms of self-reported scales, OEC, RIR, and PTC. This class appeared to answer slightly less consistently than class 2, but still managed to pass all attention checks and to provide sufficient open answer quality.

4.6. *Identifying efficient measures*

It might not always be possible to incorporate all the above-mentioned carelessness detection methods in a study. Therefore, it was of interest to reduce the number of measures but still be able to identify participants of the careless class 1 accurately. Conditional inference trees, as implemented in the *party* package for *R* (Hothorn et al., 2006), were used to identify the most efficient combination of measures to predict class membership for each participant. Conditional inference trees use a recursive algorithm to make an unbiased selection among covariates, and offer several advantages over traditional regression models and random forests (Hothorn et al., 2006; Strobl et al., 2009), such as non-linear relationships and less over-fitting. Nine variables were used to predict class

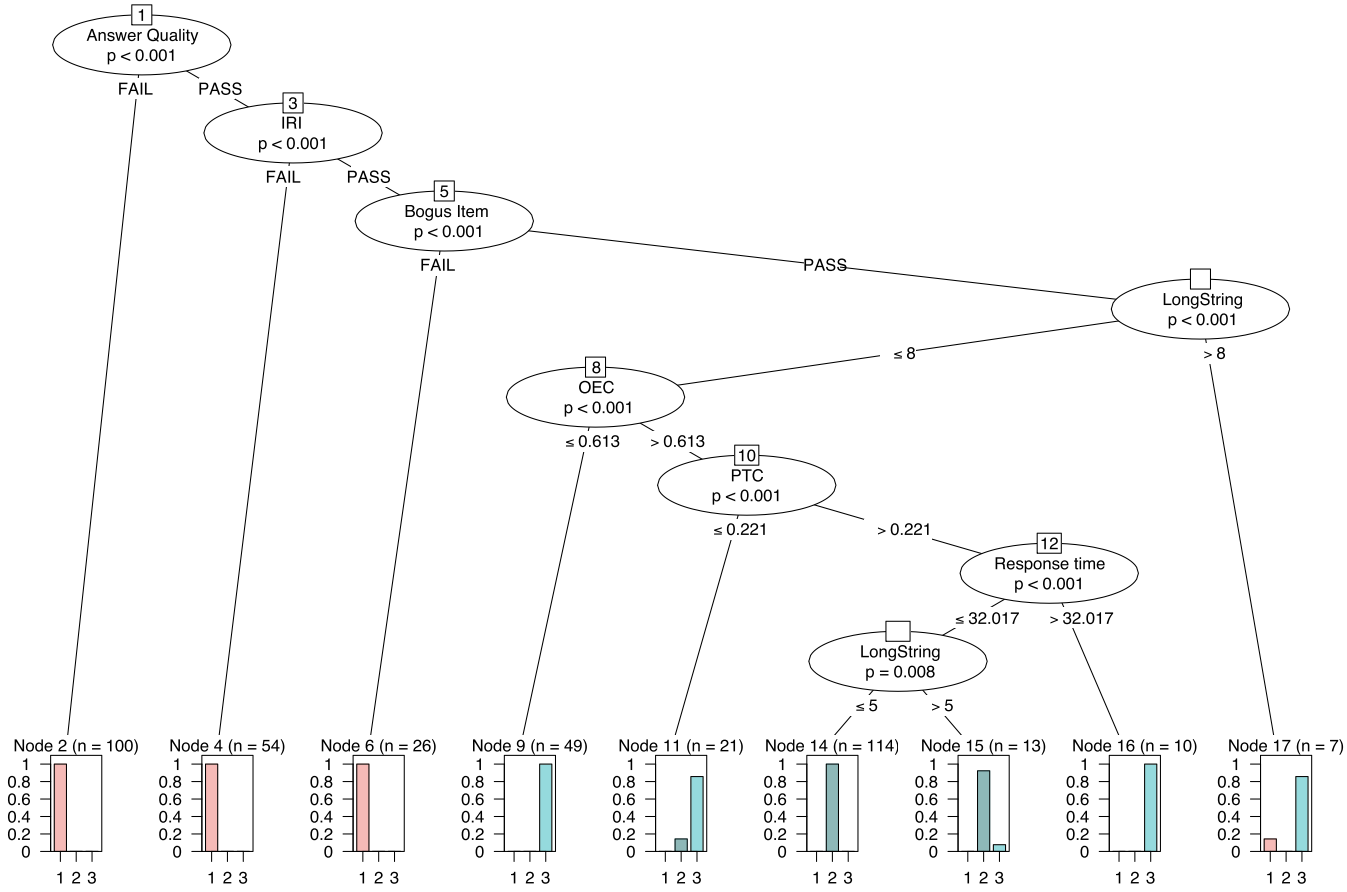

Fig. 4. Conditional inference tree for all carelessness detection methods. For each inner node, the Bonferroni-adjusted p-values are presented, the fraction of participants in each class (1, 2, or 3) is displayed for every terminal node.

Table 5
Performance of the conditional inference tree model in predicting class membership.

	Class 1	Class 2	Class 3
Predicted			
Class 1	180	0	0
Class 2	0	126	1
Class 3	1	3	83

Note. $N = 394$

membership (SRSI UseMe, Bogus Item, IRI, response time, LongString, OEC, RIR, PTC, and open answer quality). The SRSI UseMe variable, Bogus Item, IRI, and the answer quality were included as binary variables (Pass/Fail), whereas the remaining variables were used in their raw form. Results of the analysis depicted in Fig. 4 demonstrate that answer quality, IRI, and Bogus Item are well suited to separate the careless class 1 from classes 2 and 3. Furthermore, taking post hoc detection methods such as LongString analysis, OEC, PTC, and response time into account, the tree successfully separates classes 2 and 3. Table 5 demonstrates that the prediction based on this model is very accurate (Accuracy = 0.987, 95% CI[0.971, 0.996]) in terms of identifying the correct class membership. Only 5 participants out of 394 were assigned to the wrong class based on this model.

4.7. Effects of carelessness on experimental manipulation

The goal of the experiment included in the study was to examine how effect sizes and p-values changed when careless participants were excluded from the analysis. Results of a Welch's *t*-test with the full sample demonstrated that there was a significant difference in perceived trustworthiness of the online shop, $t(381.83) = 5.64$, $p = 3.344e-08$, $d = 0.567$. Participants who saw the low-trust website mock-up rated the company as less trustworthy (M = 4.36, SD = 1.21) than participants in the high-trust condition (M = 4.99, SD = 1). When participants from class 1 (n = 181) were removed, participants in the low-trust condition rated the company slightly less trustworthy (M = 4.34, SD = 1.19), and participants in the high-trust condition rated the website somewhat more trustworthy than the full sample (M = 5.12, SD = 0.95). The standard deviations decreased slightly in both groups, which indicates that some of the noise that could stem from careless participants was reduced. Although the differences between the two conditions was significant in both cases, removing participants from the careless class 1 led to a smaller p-value and increased the effect size, $t(194.32) = 5.83$, $p = 2.277e-08$, $d = 0.803$.

5. Discussion

5.1. Analysis of careless behavior in a crowdsourced sample

Previous work studied carelessness in online samples either with only a few methods (J. J. Chandler and Paolacci, 2017; Hauser and Schwarz, 2016; Kan and Drummey, 2018; Peer et al., 2017), or they assessed carelessness in student samples or mixed online samples (Maniaci and Rogge, 2014; Meade and Craig, 2012). We build on this work with a systematic analysis of carelessness in a crowdsourced sample and by examining also two new methods proposed by Curran (2016): Resampled Individual Reliability (RIR) and Person-Total Correlation (PTC). We applied six measures and corresponding cutoffs, based on recommendations from Meade and Craig (2012), Maniaci and Rogge (2014), and Curran (2016), to identify multiple forms of carelessness in a crowdsourced sample from FigureEight.

Observing the planned detection methods, which require special items or scales, 26.9% of all participants indicated in self-reports to provide careless, patterned, or rushed responses, 24.4% of all participants failed to answer the Instructed Response Item (IRI) correctly, and 23.4% missed the Bogus Item (see Table 2). Weak to moderate correlations between aggregated self-reported carelessness and other detection methods only partially indicate convergent validity for self-report measures (see Table 3). Correlations between attention check items and other detection methods were also weak to moderate, except for the Bogus Item that correlated relatively strongly with RIR and PTC. The 24.4% of participants in our FigureEight sample who failed the IRI surpass the 14% found in the study by Maniaci and Rogge (2014), which examined a mixed sample including MTurk workers, participants from online forums, and psychology students. This indicates that inattentive behavior may be more frequent in samples from crowdsourcing platforms. Taken together, approximately 25% of the sample was flagged as inattentive, based solely on one of the attention check items. It can be expected that the overall number of participants flagged with these items increases with the length of the survey, as one attention check item is recommended for every 50–100 items (Meade and Craig, 2012). In a considerably longer study, applying 4 attention check items, Peer et al. (2017) found 73% of all participants fail at least one attention check item. Post hoc detection methods revealed 6.3% of all participants were flagged by the LongString analysis, which corresponds with findings from Maniaci and Rogge (2014), where 6% were flagged in a mixed sample. However, the Odd-Even Consistency (OEC) and RIR revealed 16% and 15.5%, respectively, as responding too inconsistently. This is more than twice as much as Maniaci and Rogge (2014) identified with the OEC method. Lastly, the PTC flagged 18.8% of all participants as being careless. Hence, the post hoc detection methods of the present study further suggest that careless or inattentive behavior may be more pronounced in a fully crowdsourced sample. It is important to note though, that these comparisons have to be interpreted with caution, as FigureEight may offer a different level of quality control for participants than other crowdsourcing platforms and neither are the aforementioned comparisons based on a controlled setting.

5.2. How are the different planned and post hoc measures of carelessness related? (RQ1)

In general, the observed correlation between the different detection methods were moderate to highly positive. The correlations do not show that one method can be completely replaced by another. One exception might be the OEC and RIR methods, which are conceptually very close. However, these methods achieved a correlation of only 0.68, which was the highest correlation observed. This means that either the examined methods detect different aspects of carelessness, which is conceptually what they should do, or that all these methods have a high measurement error which reduces their precision.

The highest correlation in this study was, unsurprisingly, between RIR and OEC, as one is the generalization of the other. In contrast, PTC correlates to the weakest degree with the OEC, while correlating moderately with other measures. However, for both the RIR and the PTC the Bogus Item showed a relatively high correlation, indicating a stronger relationship between this type of attention check and the new consistency measures than with the IRI. Both measures flagged between 16 and 18.8% of the participants as careless. The strong correlation of the OEC with the RIR (0.68) suggests that these measures have a relatively high overlap. The RIR method showed a substantially stronger correlation with the more objective methods IRI and especially with the Bogus Item. Thus, and as a more general method, the RIR seems to be preferable to the OEC. However, because RIR is a *resampling* method, it has to be noted that it may return different results each time the procedure is run. Out of all post hoc detection methods, the PTC showed the highest number of flagged participants. The PTC correlated even stronger than the RIR with the planned detection methods. The PTC also correlated only moderately with the RIR, suggesting that it may cover a slightly different aspect of inconsistent responding. Therefore, both detection methods are useful especially in situations where planned detection methods are not possible. A disadvantage of the PTC is that it may not be valid when a

large portion of the sample responds carelessly.

We further used the quality assessment of the open-answers to gain an understanding of how careless respondents behave when answering surveys. Out of 394 participants, 100 (25.4%) provided insufficient open answer quality. Significantly fewer participants of this group passed attention checks; they more often self-reported bad data quality and they exhibited significantly higher LongString Index values. Furthermore, participants who failed in providing sufficient answer quality completed the survey in significantly less time, they more often failed to meet the OEC cutoff and the RIR, and they more often failed to meet the PTC cutoff. Hence, these results indicate some convergent validity for open answer quality as a measurement for carelessness. However, correlations between this measure and other planned detection or post hoc methods were rather weak (see Table 3). This brings into question the stability of carelessness, as the open answer question was shown first in our survey and the Likert-type scales and experiment after. It is possible that some participants who carefully respond to the open answer question will become fatigued and stop reading items carefully. This result coincides with findings from Maniaci and Rogge (2014), indicating that inattention or carelessness during specific tasks, such as watching a video or marking pronouns in a text, mostly has a low correlation with other detection methods of carelessness.

5.3. How prevalent is careless responding in samples from crowdsourcing platforms, based on various detection methods for carelessness? (RQ2)

Almost 60% of all participants were flagged by at least one of the methods examined in this study (see Table 2). However, the univariate examination of single measures, and a subsequent cumulative exclusion of participants, might be problematic for various reasons. Firstly, with this strategy, participants are excluded based on methods that do not have a set cutoff or an objective wrong answer, and the researcher has to decide whether one or multiple flags per participant would lead to an exclusion from the sample. Secondly, simply combining the different measures altogether might be too restrictive and lead to many false-positives. Therefore, and in line with Maniaci and Rogge (2014) and Meade and Craig (2012), we base our prevalence estimate of carelessness on the results of the LPA, which takes multiple raw values of various non-self-report methods into account to identify different classes of participants.

The LPA identified three classes in total. Class 1 (the careless participant class), contained 45.9% of all participants. This class cannot be described by one measure, and therefore comprises multiple forms of carelessness, its characteristics can be summarized as follows: Failing in providing sufficient open answer quality and failing attention checks was an exclusive characteristic of this class. This class also self-indicated bad data quality considerably more often than the other two classes. Participants of this class answered more quickly, showed very large LongString Index values, and a very low PTC, indicating excessive consistency within, yet low congruence with the total sample. While the OEC measure also revealed a relatively high inconsistency in the answers of this class, it is important to note that class 3, usually inconspicuous concerning other detection methods of carelessness, provided even more inconsistent answers. This finding suggests using caution with measures of consistency as a means of data cleaning, because they might bear potential for a high false-positive rate.

The LPA from the present study revealed a considerably larger group of careless participants (45.9%) in a crowdsourced sample compared to similar analyses conducted with mixed online samples or student pools in the studies in Maniaci and Rogge (2014) and Meade and Craig (2012). These studies identified approximately 2.2%–11% as being careless. There are multiple plausible reasons why participants were more frequently flagged in the present study: Firstly, specifically crowdworkers were recruited, which sometimes show deceptive behavior when incentives are involved (Gadiraju et al., 2015), rather than a volunteer online sample. Secondly, many different detection methods were introduced, increasing the sensitivity, but possibly reducing the specificity in finding careless respondents. Thirdly, participants on FigureEight were only required to meet level 1 qualification, which is granted after completion of a short exam. Peer, Vosgerau, and Acquisti (2014) has shown that many issues of data quality can be eliminated using the provided filter for reputation on MTurk, thus, increasing qualification requirements on FigureEight might reduce the number of participants flagged as careless.

5.4. What are the consequences of the exclusion of the respondents flagged as careless? (RQ3)

We conducted an experiment to examine the effect of careless responding on the quality of results yielded from crowdworked sources. In this study the experimental groups differed significantly in their ratings of the trustworthiness of a website, even when the data from participants flagged as careless was included. However, we found both an increase in effect size ($\Delta d = 0.236$) from a moderate ($d = 0.567$) to a large effect ($d = 0.803$), as well as a decrease in the standard deviation. This indicates lowered undesirable noise within the data. With this, we can further lend validity to the findings of the LPA which identified this class of participants. Research has further shown that carelessness can not only reduce effects but also disperse known effects (e.g., DeSimone and Harms, 2018; Maniaci and Rogge, 2014). Hence, carelessness may reduce statistical power and increase noise in the data, thus undermining the validity of online experiments.

5.5. Based on our sample, which are efficient detection methods of careless respondents to recommend? (RQ4)

In general, we strongly encourage other researchers to analyze the data quality of crowdsourced surveys. As in Maniaci and Rogge (2014) and Meade and Craig (2012), we refer to the LPA as our reference for careless behavior in our sample. Based on our findings, we recommend a set of measures that are easy to apply, easy to interpret, and at the same time cover the majority of the inattentive class 1.

1. We recommend to include an SRSI UseMe item to assess whether participants indicate that their data should be used for the study. Although this item was not an important predictor of class membership, it acts as a form of revoked consent. Thus, it serves a purpose beyond detecting bad data quality.
2. Attention checks such as an IRI should be included, because these detection methods are easy to create and offer a clear interpretation. We further advise to include a Bogus Item, as the combination of the coding of quality in open-answers, the IRI, and the Bogus Item was successful in classifying 180 of 181 participants correctly in class 1. However, the wording of the Bogus Item should be chosen carefully, because Bogus Items can cause interpretative problems (Meade and Craig, 2012).
3. While we do not recommend including an open-question and coding it for quality in your survey simply for data quality cleaning purposes, the low correlations (see Table 3) between the open-answer coding and other methods of assessing careless respondents indicate that the quality of data can vary strongly between tasks and types of data. Further, open-answers provide face validity as many participant responses can be identified as careless beyond reasonable doubt. We encourage researchers to apply carelessness detection methods according to the given task, while considering the possibility of collecting qualitative data to further their understanding of the quantitative data found.

Taken together, we recommend the following set of carelessness detection methods: an SRSI UseMe item, one or multiple Instructed Response or Bogus Items and quality coding of open-answers, if applicable. These measures either represented important predictors for class 1,

or they provided pragmatic merit. All these methods are relatively straightforward to apply, as they do not need to consider scale dimensions and inverse items. Furthermore, they were clearly associated with class 1 in the LPA, and the prediction for class 1 (based on these detection methods) was very accurate: 180 out of 181 were correctly identified, while none of the participants from classes 2 and 3 were falsely flagged by this combination of methods.

5.6. Limitations and future research

Some limitations must be considered concerning the results and recommendations presented in this paper:

First, the present study was conducted on one particular platform, FigureEight, and this might not readily translate to other platforms or recruitment methods. For instance, Amazon's MTurk offers different methods of community-management, rating possibilities for workers, and other factors such as amount of compensation or the survey interface may also vary, which may cause workers to be more attentive when taking part in a survey. Peer et al. (2017) for example found that participants from FigureEight (Crowdflower) more often failed attention checks in comparison to participants from other crowdsourcing servies. However, carelessness in this case was assessed by merely one type of measure and future research, therefore, should systematically assess data quality differences between various platforms, applying multiple carelessness detection methods. Furthermore, as even a seemingly small number of careless responses can have serious consequences (Huang et al., 2015; Oppenheimer et al., 2009), the investigation of careless responding in a sample should be conducted in any case, even though the prevalence may vary among platforms.

Second, the present study assessed the detection of careless participants which resulted in the exclusion of approximately half of all participants. Excluding this number of participants could have severe methodological and financial implications. Hence, future research should also focus on preventing carelessness, which is presently not well understood. Warnings about monitoring data quality that have been used by Clifford and Jerit (2015) or Meade and Craig (2012) can be effective, but might lead to other biases, such as socially desirable behavior. D. Chandler and Kapelner (2013) have found positive effects of meaning by explaining the purpose of a task on data quality in crowdsourcing tasks. Furthermore, Ward and Pond (2015) found that promising participants the results of the study was effective in increasing data quality. More effort is needed to systematically analyze these measures for preventing carelessness in crowdsourced samples.

Third, the present study merely included one open-ended question to assess the overlap between different behavioral carelessness detection measures and the quality of qualitative data. The low overlap of these methods therefore only provide a first clue regarding the task-dependency of carelessness.

One might further argue, that attention in the different types of tasks was confounded by the order in which these tasks appeared. Although findings from Maniaci and Rogge (2014) suggested a similar approach by applying other forms of different tasks in their survey, future research should aim for a systematic review of a wider variety of different tasks in online surveys. This will facilitate further analysis of the task-depencency of careless behavior. Recent research on microtasks showed that the order (Cai et al., 2016; Newell and Ruths, 2016) as well as the similarity (Aipe and Gadiraju, 2018) of different tasks can affect worker performance, which may also have implications for online surveys.

Finally, as all post hoc detection methods are approximate and uncertain, bad data quality can not clearly and reliably be identified in every case. Our recommendations are based in particular on the prediction of class 1, which was identified using LPA. Only planned detection methods were found to be predictive for this class. However, there are situations where it might not be possible to include attention check items or task-dependent measures of quality, such as voluntary surveys of highly specific populations. Hence, further research is needed to ensure data quality in such situations.

6. Conclusion

The aim of this study was to provide an estimate of the frequency of carelessness in samples from crowdsourcing platforms, based on different identification methods. Results revealed that 45.9% of the participants in our crowdsourced sample display careless behavior. Furthermore, carelessness and inattention appear highly task-dependent, as correlations between open answer quality and other measures are rather low. Finally, based on a predictive model and interpretative problems of several detection methods, we recommend assessing data quality of crowdsourced samples by applying the following: an SRSI UseMe item, attention checks such as the IRI and Bogus Item, and the coding of quality in open-answers or a different task-related measure of quality. A combination of these methods was able to identify 180 out of 181 careless participants, and the subsequent exclusion of this subsample resulted in an increased effect size and smaller p-value in the experiment.

Conflict of interest statement

The authors have no conflict of interest to declare.

Acknowledgments

Special thanks to Sharon T. Steinemann and Sebastian Perrig. This work has been approved by the Institutional Review Board of the Faculty of Psychology, University of Basel under the number D-006-17. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix

Calculating the open answer quality Index

A-priori criteria for the rating of open-ended questions were defined according to the measures used in studies from Holland and Christian (2009); Smyth et al. (2009). The following indicators for calculating an open answer quality index were taken into consideration:

Substantive response

This indicator refers to whether the participant answer thematically corresponds to the open question subject matter. The open answer has been coded with 0 if it merely consisted of meaningless sequences of letters, clearly copy-pasted phrases, or thematically unfit answers which typically emerged from not carefully reading the instructions (such as describing a negative experience in a non-virtual store instead of an online shop). If the open answer corresponded to the subject matter, regardless whether the participant addressed all subquestions, the indicator has been coded with 1.

Number of words

Because it is possible to provide a thematically substantial answer while providing little or zero actual content (such as merely writing one short

sentence about the experience), the number of words has been assessed for each open answer. Given the topic of the open question and the two subsequent subquestions, a minimum of 50 words (±3) was defined as the requirement to answer the questions. Thus, participants were explicitly asked to provide an answer containing at least 50 words. This number is regarded as being a minimum effort to achieve a thematically substantial answer that additionally addresses at least one subquestion. Wordcounts equal or surpassing this number were coded with 1, smaller wordcounts with 0.

Complete sentences

Participants were explicitly asked to provide answers with full sentences. Open answers that mainly or exclusively consisted of unfinished sentences or separate words were coded with 0 in regard to complete sentences. To receive a coding of 1, the majority of all sentences in the open answer needed to be complete and separated with commas or periods.

Number of subquestions answered

If none of the specific subquestions were addressed, the answer was coded with 0 in regard to number of subquestions. This was also the case if the given answer met the requirements for a thematically substantial answer, but failed to answer at least one of the specific subquestions. Accordingly, the answer received a coding of 1 or 2 if one or both subquestions were addressed in the open answer, respectively.

Number of subquestions elaborated

An answer to a subquestion was considered to be elaborate if the according part of the open answer contained at least three complete sentences. If none of the subquestions were elaborated, the answer was coded with 0 in regard to themes elaborated. Accordingly, the answer received a coding of 1 or 2 if one or both subquestions were elaborated in the open answer, respectively.

Calculation of the open answer quality Index

Substantive response and *Number of words* were seen as essential for providing a valuable open answer. Thus, if one or both of these variables were coded with 0, the open answer quality Index was also automatically coded with 0. The other variables, namely *complete sentences*, *number of subquestions answered* and *number of subquestions elaborated*, were seen as being important (but not an absolute necessity) on their own in order to provide a good open answer quality. Thus, for answers that met the minimum requirements, the codings from *complete sentences*, *number of themes*, and *themes elaborated* were counted together. If the sum reached 3 or higher, the overall open answer quality was considered to be adequate, and thus coded with 1.

References

Aipe, A., Gadiraju, U., 2018. Similarhits: revealing the role of task similarity in microtask crowdsourcing. In: Proceedings of the 29th on Hypertext and Social Media. ACM, New York, NY, USA, pp. 115–122. https://doi.org/10.1145/3209542.3209558.

Aust, F., Diedenhofen, B., Ullrich, S., Musch, J., 2013. Seriousness checks are useful to improve data validity in online research. Behav. Res. Methods 45 (2), 527–535. https://doi.org/10.3758/s13428-012-0265-2.

Cai, C.J., Iqbal, S.T., Teevan, J., 2016. Chain reactions: the impact of order on microtask chains. In: Proceedings of the 2016 Chi Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 3143–3154. https://doi.org/10.1145/2858036.2858237.

Casler, K., Bickel, L., Hackett, E., 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. Comput. Hum. Behav. 29 (6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009.

Chandler, D., Kapelner, A., 2013. Breaking monotony with meaning: motivation in crowdsourcing markets. J. Econ. Behav. Organ. 90, 123–133. https://doi.org/10.1016/j.jebo.2013.03.003.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., Ratliff, K.A., 2015. Using nonnaive participants can reduce effect sizes. Psychol. Sci. 26 (7), 1131–1139. https://doi.org/10.1177/0956797615585115.

Chandler, J., Shapiro, D., 2016. Conducting clinical research using crowdsourced convenience samples. Annu. Rev. Clin. Psychol. 12 (1), 53–81. https://doi.org/10.1146/annurev-clinpsy-021815-093623.

Chandler, J., Mueller, P., Paolacci, G., 2014. Nonnaïveté among amazon mechanical turk workers: consequences and solutions for behavioral researchers. Behav. Res. Methods 46 (1), 112–130. https://doi.org/10.3758/s13428-013-0365-7.

Chandler, J.J., Paolacci, G., 2017. Lie for a dime: when most prescreening responses are honest but most study participants are impostors. Soc. Psychol. Pers. Sci. 8 (5), 500–508. https://doi.org/10.1177/1948550617698203.

Cheung, J.H., Burns, D.K., Sinclair, R.R., Sliter, M., 2017. Amazon mechanical turk in organizational psychology: an evaluation and practical recommendations. J. Bus. Psychol. 32 (4), 347–361. https://doi.org/10.1007/s10869-016-9458-5.

Clifford, S., Jerit, J., 2015. Do attempts to improve respondent attention increase social desirability bias? Publ. Opin. Q. 79 (3), 790–802. https://doi.org/10.1093/poq/nfv027.

Comley, P., 2015. Online market research. In: Hamersveld, M.v., Bont, C.d. (Eds.), Market Research Handbook. John Wiley & Sons Ltd, Chichester, England, pp. 401–419. https://doi.org/10.1002/9781119208044.ch21.

Curran, P.G., 2016. Methods for the detection of carelessly invalid responses in survey data. J. Exp. Soc. Psychol. 66, 4–19. https://doi.org/10.1016/j.jesp.2015.07.006.

DeSimone, J.A., Harms, P.D., 2018. Dirty data: the effects of screening respondents who provide low-quality data in survey research. J. Bus. Psychol. 33 (5), 559–577. https://doi.org/10.1007/s10869-017-9514-9.

de Winter, J., Kyriakidis, M., Dodou, D., Happee, R., 2015. Using crowdflower to study the relationship between self-reported violations and traffic accidents. Procedia Manuf. 3, 2518–2525. https://doi.org/10.1016/j.promfg.2015.07.514.

Dogan, V., 2018. A novel method for detecting careless respondents in survey data: floodlight detection of careless respondents. J. Market. Anal. 6 (3), 95–104. https://doi.org/10.1057/s41270-018-0035-9.

Douglas, K.M., McGarty, C., 2001. Identifiability and self-presentation: computer-mediated communication and intergroup interaction. Br. J. Soc. Psychol. 40 (3), 399–416. https://doi.org/10.1348/014466601164894.

Flavián, C., Guinalíu, M., Gurrea, R., 2006. The role played by perceived usability, satisfaction and consumer trust on website loyalty. Inf. Manag. 43 (1), 1–14. https://doi.org/10.1016/j.im.2005.01.002.

Fleischer, A., Mead, A.D., Huang, J., 2015. Inattentive responding in mturk and other online samples. Ind. Organ. Psychol. 8 (2), 196–202. https://doi.org/10.1017/iop.2015.25.

Gadiraju, U., Kawase, R., Dietze, S., Demartini, G., 2015. Understanding malicious behavior in crowdsourcing platforms: the case of online surveys. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 1631–1640. https://doi.org/10.1145/2702123.2702443.

Gadiraju, U., Möller, S., Nöllenburg, M., Saupe, D., Egger-Lampl, S., Archambault, D., Fisher, B., 2017. Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd. In: Archambault, D., Purchase, H., Hoßfeld, T. (Eds.), Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments. Springer International Publishing, Cham, pp. 6–26.

Gosling, S.D., Mason, W., 2015. Internet research in psychology. Annu. Rev. Psychol. 66 (1), 877–902. https://doi.org/10.1146/annurev-psych-010814-015321.

Hassenzahl, M., Burmester, M., Koller, F., 2003. AttrakDiff: ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: Szwillus, G., Ziegler, J. (Eds.), Mensch & Computer 2003: Interaktion in Bewegung. Vieweg+Teubner Verlag, Wiesbaden, pp. 187–196. https://doi.org/10.1007/978-3-322-80058-9_19.

Hauser, D.J., Schwarz, N., 2016, Mar 01. Attentive turkers: mturk participants perform better on online attention checks than do subject pool participants. Behav. Res. Methods 48 (1), 400–407. https://doi.org/10.3758/s13428-015-0578-z.

Holland, J.L., Christian, L.M., 2009. The influence of topic interest and interactive probing on responses to open-ended questions in web surveys. Soc. Sci. Comput. Rev. 27 (2), 196–212. https://doi.org/10.1177/0894439308327481.

Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph Stat. 15 (3), 651–674. https://doi.org/10.1198/106186006X133933.

Huang, J.L., Curran, P.G., Keeney, J., Poposki, E.M., DeShon, R.P., 2012. Detecting and deterring insufficient effort responding to surveys. J. Bus. Psychol. 27 (1), 99–114. https://doi.org/10.1007/s10869-011-9231-8.

Huang, J.L., Liu, M., Bowling, N.A., 2015. Insufficient effort responding: examining an insidious confound in survey data. J. Appl. Psychol. 100 (3), 828. https://doi.org/10.1037/a0038510.

John, O.P., Srivastava, S., 1999. The big five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (Eds.), Handbook of Personality: Theory and Research, vol. 2. Guilford Press, New York, NY, US, pp. 102–138.

Johnson, J.A., 2005. Ascertaining the validity of individual protocols from web-based personality inventories. J. Res. Pers. 39 (1), 103–129. https://doi.org/10.1016/j.jrp.2004.09.009.

Kam, C.C.S., Meyer, J.P., 2015. How careless responding and acquiescence response bias can influence construct dimensionality: the case of job satisfaction. Organ. Res. Methods 18 (3), 512–541. https://doi.org/10.1177/1094428115571894.

Kan, I.P., Drummey, A.B., 2018. Do imposters threaten data quality? an examination of worker misrepresentation and downstream consequences in amazon's mechanical turk workforce. Comput. Hum. Behav. 83, 243–253. https://doi.org/10.1016/j.chb.2018.02.005.

Kazai, G., Kamps, J., Milic-Frayling, N., 2011. Worker types and personality traits in crowdsourcing relevance labels. In: Proceedings of the 20th Acm International Conference on Information and Knowledge Management. ACM, New York, NY, USA, pp. 1941–1944. https://doi.org/10.1145/2063576.2063860.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropr. Med. 15 (2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Lee, H., 2006. Privacy, publicity, and accountability of self-presentation in an on-line discussion group. Socio. Inq. 76 (1), 1–22. https://doi.org/10.1111/j.1475-682X.2006.00142.x.

Maniaci, M.R., Rogge, R.D., 2014. Caring about carelessness: participant inattention and its effects on research. J. Res. Pers. 48, 61–83. https://doi.org/10.1016/j.jrp.2013.09.008.

McKay, A.S., Garcia, D.M., Clapper, J.P., Shultz, K.S., 2018. The attentive and the careless: examining the relationship between benevolent and malevolent personality traits with careless responding in online surveys. Comput. Hum. Behav. 84, 295–303. https://doi.org/10.1016/j.chb.2018.03.007.

Meade, A.W., Craig, S.B., 2012. Identifying careless responses in survey data. Psychol. Methods 17 (3), 437–455. https://doi.org/10.1037/a0028085.

Moshagen, M., Thielsch, M.T., 2010. Facets of visual aesthetics. Int. J. Hum. Comput. Stud. 68 (10), 689–709. https://doi.org/10.1016/j.ijhcs.2010.05.006.

Muthén, B.O., 2002. Beyond sem: general latent variable modeling. Behaviormetrika 29 (1), 81–117. https://doi.org/10.2333/bhmk.29.81.

Newell, E., Ruths, D., 2016. How one microtask affects another. In: Proceedings of the 2016 Chi Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 3155–3166. https://doi.org/10.1145/2858036.2858490.

Nichols, D.S., Greene, R.L., Schmolck, P., 1989. Criteria for assessing inconsistent patterns of item endorsement on the mmpi: rationale, development, and empirical trials. J. Clin. Psychol. 45 (2), 239–250. https://doi.org/10.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1.

Niessen, A.S.M., Meijer, R.R., Tendeiro, J.N., 2016. Detecting careless respondents in web-based questionnaires: which method to use? J. Res. Pers. 63, 1–11. https://doi.org/10.1016/j.jrp.2016.04.010.

Oppenheimer, D.M., Meyvis, T., Davidenko, N., 2009. Instructional manipulation checks: detecting satisficing to increase statistical power. J. Exp. Soc. Psychol. 45 (4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009.

Paolacci, G., Chandler, J., 2014. Inside the turk: understanding mechanical turk as a participant pool. Curr. Dir. Psychol. Sci. 23 (3), 184–188. https://doi.org/10.1177/0963721414531598.

Peer, E., Brandimarte, L., Samat, S., Acquisti, A., 2017. Beyond the turk: alternative platforms for crowdsourcing behavioral research. J. Exp. Soc. Psychol. 70, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006.

Peer, E., Vosgerau, J., Acquisti, A., 2014. Reputation as a sufficient condition for data quality on amazon mechanical turk. Behav. Res. Methods 46 (4), 1023–1031.

Powers, D.M., 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. J. Mach. Learn. Technol. 2 (1), 37–63.

Rieser, D.C., Bernhard, O., 2016. Measuring trust: the simpler the better?. In: Proceedings of the 2016 Chi Conference Extended Abstracts on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 2940–2946. https://doi.org/10.1145/2851581.2892468.

Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E., 2016. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R. J. 8 (1), 205–233.

Seckler, M., Heinz, S., Forde, S., Tuch, A.N., Opwis, K., 2015. Trust and distrust on the web: user experiences and website characteristics. Comput. Hum. Behav. 45, 39—50. https://doi.org/10.1016/j.chb.2014.11.064.

Sheldon, K.M., Elliot, A.J., Kim, Y., Kasser, T., 2001. What is satisfying about satisfying events? Testing 10 candidate psychological needs. J. Pers. Soc. Psychol. 80 (2), 325. https://doi.org/10.1037/0022-3514.80.2.325.

Skitka, L.J., Sargis, E.G., 2006. The internet as psychological laboratory. Annu. Rev. Psychol. 57 (1), 529–555. https://doi.org/10.1146/annurev.psych.57.102904.190048.

Smyth, J.D., Dillman, D.A., Christian, L.M., Mcbride, M., 2009. Open-ended questions in web surveyscan increasing the size of answer boxes and providing extra verbal instructions improve response quality? Publ. Opin. Q. 73 (2), 325–337. https://doi.org/10.1093/poq/nfp029.

Stewart, N., Chandler, J., Paolacci, G., 2017. Crowdsourcing samples in cognitive science. Trends Cognit. Sci. 21 (10), 736–748. https://doi.org/10.1016/j.tics.2017.06.007.

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14 (4), 323–348. https://doi.org/10.1037/a0016973.

Tuch, A.N., Schaik, P.V., Hornbæk, K., 2016. Leisure and work, good and bad: the role of activity domain and valence in modeling user experience. ACM Trans. Comput. Hum. Interact. 23 (6) https://doi.org/10.1145/2994147, 35:1–35:32.

Van Pelt, C., Sorokin, A., 2012. Designing a scalable crowdsourcing platform. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. ACM, New York, NY, USA, pp. 765–766. https://doi.org/10.1145/2213836.2213951.

Ward, M., Pond, S.B., 2015. Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. Comput. Hum. Behav. 48, 554–568 doi: 0.1016/j.chb.2015.01.070.

Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. J. Pers. Soc. Psychol. 54 (6), 1063. https://doi.org/10.1037/0022-3514.54.6.1063.