Survey Research Methods (2019)
Vol. 13, No. 3, pp. 229-248
doi:10.18148/srm/2019.v13i3.7403

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

Too Fast, too Straight, too Weird: Non-Reactive Indicators for Meaningless Data in Internet Surveys

Dominik J. Leiner
Department of Media and Communication
LMU Munich, Germany

Practitioners use various indicators to screen for meaningless, careless, or fraudulent responses in Internet surveys. This study employs an experimental-like design to empirically test the ability of non-reactive indicators to identify records with low data quality. Findings suggest that careless responses are most reliably identified by questionnaire completion time, but the tested indicators do not allow for detecting intended faking. The article introduces various indicators, their benefits and drawbacks, proposes a completion speed index for common application in data cleaning, and discusses whether to remove meaningless records at all.

Keywords: data cleaning, careless responding, meaningless data, paradata, web-based surveys, online surveys.

1 Introduction

Academic researchers and practitioners appreciate respondent-administered Internet surveys for providing an efficient and cost-minimizing method to collect data. The survey mode *Internet* has become increasingly common in survey research, not least because response behavior in web-based surveys was found similar to pen 'n' paper mail surveys (for a summary see Couper & Bosnjak, 2010). The uses and limitations of web-based surveys are broadly discussed in survey methodology textbooks (e.g., Bethlehem & Biffignandi, 2012; Callegaro, Lozar Manfreda, & Vehovar, 2015; Fowler, 2009; Groves et al., 2011; Marsden & Wright, 2010; Sue & Ritter, 2012).

Invalid responses are considered one drawback of web-based surveys. Although they are not specific to web-based surveys, they are more likely to occur on the Internet: A web page "may give respondents a sense of reduced accountability" in comparison to a printed questionnaire (Johnson, 2005, p. 108), and submitting an online questionnaire requires much less effort than submitting a printed questionnaire as letter. The increase in invalid records is accompanied by a decreased chance of incidentally finding them. Every click in the online questionnaire is automatically encoded and written into the data set. There is no human typing the answers from printed questionnaires and recognizing uncommon response behavior, such as zigzag patterns in matrix-style question batteries.

Contact information: Dominik J. Leiner, LMU Munich, Department of Media and Communication, Oettingenstr. 29, 80538 Munich, Germany, email: leiner@ifkw.lmu.de.

Detection of uncommon patterns, however, can be automated and, besides, metadata or paradata (Kreuter, 2013) is easily available in web-based surveys. This includes page/survey completion times, a respondent's IP address, information about the browser, device and screen size, as well as more detailed paradata that can easily be collected with little effort (Cellagaro, 2013; Diedenhofen & Musch, 2017; Olson & Parkhurst, 2013). Paradata has proven helpful to identify multiple submissions by the same respondent (Bowen, Daniel, Williams, & Baird, 2008; Johnson, 2005; Konstan, Rosser, Ross, Stanton, & Edwards, 2005; Selm & Jankowski, 2006) and to supplement the screening for careless responses (Barge & Gehlbach, 2012; Bauermeister et al., 2012; Meade & Craig, 2012).

The paper aims to improve understanding about *bad* or *low-quality* survey data (Schendera, 2007, p. 6). It starts with a summary upon meaningless data and discusses promising indicators of data quality. Subsequently, it presents four studies that evaluate multiple indicators in terms of their ability to identify different kinds of meaningless records in self-administered Internet surveys. Finally, the paper draws practical conclusions on the application of quality indicators in field research.

2 Meaningless Data

As validity and invalidity have several facets, different terms have been used to refer to "bad" survey data. The common characteristic of invalid survey responses is that they do not reflect the true characteristics of the survey respondent, but something else instead—some measurement error. The term *meaningless responses* is more specific in attributing this error to the respondent, that is, the respondent is not *will-*

ing to give a valid response. Meaningless responses shall be distinguished from "pseudo-opinions" (Bishop, Oldendick, Tuchfarber, & Bennett, 1980) and "nonattitudes" (Franzén, 2011; Schuman & Presser, 1980) that are caused by respondents who are *unable* to provide a valid answer, for example, due to insufficient knowledge or understanding. Such responses are typically not meaningless, but often reflect some more general attitudes (Payne, 1950).

Literature describes the phenomenon of meaningless data with attributions to causes and appearance: "Satisficing" (Krosnick, 1991, 1999) and "inattentive or careless response" (Johnson, 2005; Meade & Craig, 2012, p. 438) refer to the respondent's intention to give a qualified answer. "Response sets" (Jandura, Peter, & Küchenhoff, 2012), "response styles" (Van Vaerenbergh & Thomas, 2012) and "content nonresponsivity" (Meade & Craig, 2012, p. 437; Nichols, Greene, & Schmolck, 1989) refer to the observation that an answer is more or less independent from what was asked. Sometimes the term "random responding" is used to express that answer options are selected *arbitrarily*, but this is somewhat misleading because meaningless answers rather follow effortless patterns (e.g., always selecting the first option, Meade & Craig, 2012) instead of being statistically random.

The respondent, of course, is only one possible source of invalid data. A fit between the research questions and the employed measures, the wording of questions (Converse & Presser, 2003; Payne, 1980) are necessary prerequisites for useful survey data; and depending on the research design, sampling, coverage, and non-response (Dillman, 2013) may threaten data quality more severely than a few meaningless records (Weisberg, 2009). Yet, given a systematically and rigorously designed questionnaire, those few records have the potential to cause serious errors, such as faux-significant effects in an experiment.

2.1 Systematic Result Biases

The hazard posed by meaningless data depends on how the measurement error affects data structures. In the best case only the accuracy (Wang & Strong, 1996) of the data set is affected, causing type II errors (not rejecting wrong null-hypotheses; for details see Meade & Craig, 2012). Given that respondents will usually not give statistically random responses, the best case is unlikely to occur. More often, we have to assume that meaningless data is systematically different from valid data regarding response distributions. For example, when a respondent always selects the first response option in scales with unbalanced items (for a summary on response styles see Van Vaerenbergh & Thomas, 2012). Using such data exposes the researcher to the risk of drawing wrong conclusions (type I errors) and possibly to make detrimental recommendations (Bauermeister et al., 2012; Woods, 2006).

2.2 Detection of Meaningless Records

Meade and Craig (2012) distinguish two routes for detecting meaningless data. If researchers anticipate meaningless data to be a serious issue *a-priori*, additional questions may be included in the questionnaire to identify meaningless data (for an overview see DeSimone, Harms, & DeSimone, 2015). This first route is comprised of self-reports (direct questions whether to use the answers for analysis, or scales for response behavior; also see Aust, Diedenhofen, Ullrich, & Musch, 2012) and covered measures (scales designed to measure language understanding or consistent responding, bogus items, or instructed response items). With a focus on faking behavior, Burns and Christiansen (2011) present a systematic framework and summary of such methods (also see Allen, 1966; Azfar & Murrell, 2009; Lim & Butcher, 1996; Pine, 1995). Their summary also covers the second route for detecting meaningless data: *post-hoc* analysis of the data collected in the survey. While the first route is reactive (Lavrakas, 2008), the second route identifies anomalies in the responses' means, variance, and correlation structure. Paradata collected during the survey allows for a third path that is often available, even if the researcher did not consider meaningless data a problem a-priori.

The academic community is just beginning to establish standards on how to identify and handle potentially problematic records (Osborne, 2013). In spite of the significant threats that meaningless records pose to scholarly research, their identification in web-based surveys has mostly been subject to practitioners (Bhaskaran & LeClaire, 2010; Rogers & Richarme, 2009). But when data cleaning is based on untested assumptions, removing data may render new biases (Bauermeister et al., 2012; Harzing, Brown, Köster, & Zhao, 2012). A researcher may even face the accusation of data manipulation, if data cleaning is not argued on systematic research and the cleaned data fits the model better than the original data. With a clear focus on the non-reactive indicators (paradata and post-hoc analyses of data), this paper asks (RQ 1):

Which non-reactive data quality indicators are the most efficient ones in identifying records of meaningless data in an Internet survey?

Nichols et al. (1989) suggest to differentiate between careless responding and faking. To give consideration to different kinds of meaningless data, a secondary research question (RQ 2) is:

Which are the most efficient quality indicators to identify specific types of meaningless data?

Literature provides two studies that empirically test non-reactive indicators to identify careless cases in Internet surveys. Both studies' questionnaires include one or more extensive personality inventories with 300 and 400 items, respectively. Johnson (2005) focuses on the distributions of four quality indicators in a large sample (N = 23076). An el-

bow criterion identifies clear cut-points in the distributions of a straightlining index and the number of missing responses. Two further indices for inter-item correlation do not show such clear cut points. Notably, Johnson (2005, p. 119) found the different consistency measures to identify mostly independent sets of cases as being meaningless.

Meade and Craig (2012) include bogus items and self-reports on response quality in the questionnaire. The article compares indices from seventeen quality indicators and finds them to correlate only low to moderately (N = 438). Based on the results from latent cluster analysis, Meade and Craig (2012) argue that correlation measures that detect inconsistent answers, bogus items, and a diligence scale are most efficient in identifying careless respondents.

Johnson (2005), Meade and Craig (2012) provide valuable insights into the distribution of data quality indicators and the relation of different indicators. Yet, both studies do not employ an external criterion for careless responses. The conclusion that a response is careless is based on data structure and response anomalies – assuming that the indicators actually predict careless responding. To test this assumption, and therefore, to determine the predictive validity of the indicators, this study employs an experimental-like design: Some respondents are asked for careless and fraudulent responses, and indicators compete to identify their records.

3 Non-Reactive Data Quality Indicators

This paper distinguishes five classes of non-reactive data quality indicators. (1) The *percentage of missing data* is important for data cleaning in general (Barge & Gehlbach, 2012; Börkan, 2010; Kwak & Radler, 2002; Shin, Johnson, & Rao, 2012). Unanswered questions are a significant limitation for nearly any kind of data analysis and can render a record unusable. Internet surveys can automatically probe or reject missing answers to ensure complete data sets (Franzén, 2011; Krosnick & Fabrigar, 2003; K. C. Schneider, 1985; Schuman & Presser, 1980). Yet, such filtering may obfuscate cases in which a person just leafs through the questionnaire. Regarding the quality of answers, missing data may be of ambivalent informative value. Unmotivated respondents likely skip questions (Barge & Gehlbach, 2012), but highly motivated respondents could as well express an "I do not feel qualified to answer this question" by omitting the answer.

In printed questionnaires, (2) *patterns in matrix-style questions*, such as a Likert question battery ("scale"), are the most obvious indicator for suspicious data. Annoyed respondents typically paint straight vertical lines (the same response option is chosen for each item of a scale, also known as *straightlining*, (Schonlau & Toepoel, 2015), diagonal lines, and a combination of both (figure 1). Such patterns do not necessarily render the response invalid, but it seems likely that the respondents had the pattern in mind rather than the battery items.

The (3) *distance from the sample means* is a straightforward measure to identify respondents giving atypical answers. There is a high face validity of removing outliers, if the sample shows a clean normal distribution and single cases or small groups cause "peaks" or lie far outside the limits of three or four standard deviations. Respondents who click the first option for every item of a scale, for example, can cause such an outlier group (figure 2). Outliers may indicate meaningless data, but Bhaskaran and LeClaire (2010) argue that outliers may as well be valid answers from atypical respondents. Removing outliers will directly affect a sample's variance and means.

The (4) *correlation structure* within the answers (consistency) is a chimera. On the one hand, answers about the same construct shall be consistent and therefore highly correlated. The same is expected for measures on related or dependent constructs. This renders inconsistent answers suspicious of being invalid answers. On the other hand, differentiation between similar but non-identical items might rather indicate the respondent's cognitive effort (Krosnick & Alwin, 1988). Vice versa, straightlining results in very high consistency if scales do not contain reversed items. Kurtz and Parrish (2001) argue that valid responses also may seem inconsistent and Sniderman and Bullock (2004) argue that inconsistent answers may simply indicate that the respondent is not familiar with the issue under research. In such a case, inconsistent response behavior may just indicate loaded question wording and weak attitudes (Klirs & Revelle, 1986), if not even the attitude itself is inconsistent (Ajzen, 1988; Katz, 1968; Kuhn, 1991). Last but not least, data cleaning based on correlations may interfere with hypothesis testing. If only those respondents are selected for analysis who show a correlation one seeks to test, this is clearly a violation of prudence.

When using computer-assisted survey modes, (5) *completion time* is routinely available, for example, measured per question (CATI) or per questionnaire page (web-based survey). Survey completion time is predicted by the personality trait reliability (Furnham, Hyde, & Trickey, 2013) and correlates to (less) measurement artifacts (Malhotra, 2008) and increased attention (Revilla & Ochoa, 2014). Completion time is no quality indicator per se: Many reasons can cause a respondent to complete a 15-minute questionnaire in 5 minutes. Filter questions may have hidden substantial parts of the questionnaire, or the respondent may be an expert who can respond very quickly. If the interview was face-to-face or via telephone, the interviewer will have an idea of the reasons, but the response codes from a self-administered questionnaire provide little clues.

If no legitimate explanation can be found for increased completion speed (*rushing*), then we must assume that the respondent did not answer carefully or did even not read the questions. Notably, research on interviewer-administered surveys originally focused on *response latencies* as a mea-

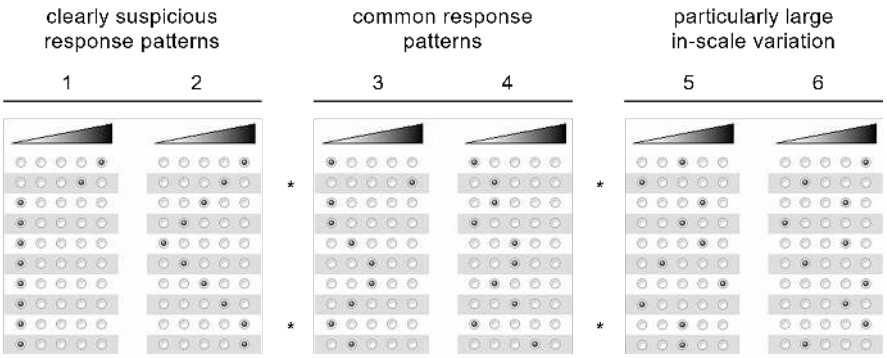

Figure 1. Response patterns observed in a Likert-like scale on elaboration. Items 2 and 9 were reversed. Overall scale consistency was $\alpha = 0.85$, $N = 11201$.

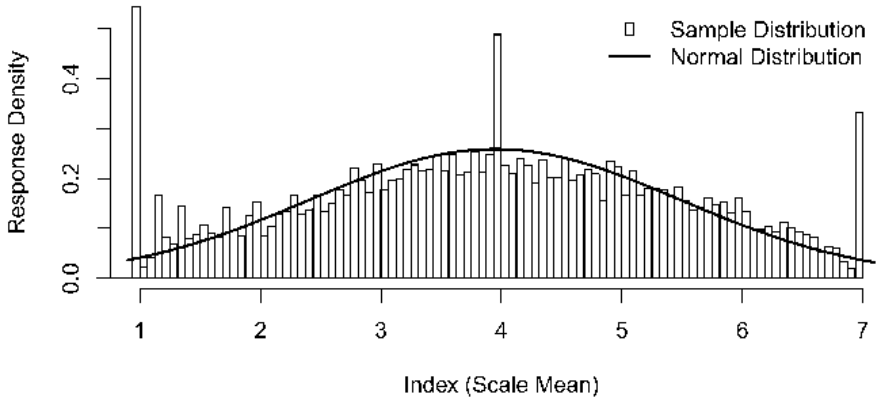

Figure 2. Anomalies in the distribution of attitudes measured in a bi-polar scale. Means from a 7-point scale with 16 items ($N = 11032$, see studies 1 and 2 for sample details). The peaks at the scale's middle and extremes are likely caused by straightliners.

sure for attitude availability, and found that a longer latency indicates low-quality data (Draisma & Dijkstra, 2004). A general downside of completion times is their large interpersonal variation (Fazio, 1990; Meade & Craig, 2012, p. 447; Yan & Tourangeau, 2008), which may outweigh variation in the respondents' effort. Leiner and Doedens (2010), for example, point out that completion time does not predict test-retest reliability, except for extreme cases. The duration between starting and finishing a questionnaire comprises time to read, think, and respond, but also time for technical processing (Internet transmission, server processing, also see Stieger & Reips, 2010), and break times that are not actively spent on the questionnaire, but with leaving the room, checking e-mails, social media, and so on. Technical delays typically become negligible when pages ask for more than three or four responses, but breaks may introduce substantial artifacts.

Practical research will often employ a sixth class of indicators: *meaning and plausibility*. If the questionnaire asks for 23 online services, and a respondent reports to use each with the highest possible frequency, the record will appear in the outlier analysis and it most likely contains meaningless data. If the respondent reports an age of 118 years, this is probably not correct, but may be a typo. An age of 99 years may be a statement agains disclosing personal data. Open-ended text questions often provide rich information to better understand the respondent, and they may support the decision of whether to remove a record from the data set or not. Nonetheless, *meaning and plausibility* can hardly be generalized and standardized throughout different surveys, which is the reason for excluding content-specific indicators from this paper.

4 Method

To test the indicators' efficiency in identifying cases of meaningless data, four studies were conducted. Their design follows an experimental logic so far as respondents were assigned into (two) groups that received different treat-

ments. Similar to a design employed by Kemper and Menold (2014) to research interview data fabrication, respondents in the "low-quality" group were explicitly asked to produce meaningless data. The success of the studies depends on whether the participants follow this instruction. If the respondents self-reported that they had not followed it (manipulation check), their records were removed from analyses. The "high-quality" group completed the questionnaire normally. The experimental outcome variables, technically, are the data quality indicators as introduced above. But to better fit practical application, the analysis logic is reversed: Analyses do not test whether there was an effect of the treatment on the indicators, but if those indicators allow the researcher to tell the high- and low-quality groups apart.

4.1 Treatment

The aim was to elicit *meaningful* answers from the high-quality group (HQ) and *meaningless* data from the low-quality group (LQ). The treatment started with the invitation: The HQ group was invited to a survey on "public opinion" (studies 1 and 2), "stress" (study 3), or "communication with managers" (study 4). The LQ group was invited to a study on "poor survey data" (studies 1 and 2) or an "uncommon study" (studies 3 and 4). The first page of the questionnaire then gave away more information. The questionnaire instructions in the HQ group asked the respondents to give their personal opinions but included no appeal to answer the questions particularly carefully, except for a subsample in study 3.

In the LQ group, the instructions explained that unlike in other surveys, this questionnaire expects them *not* to give qualified answers. Then eye-catching red, large friendly letters (studies 1 and 2, with repetition on the next page) or a yellow box (studies 3 and 4, only once) gave the exact instruction on how to complete the questionnaire. We assume that all respondents had previous experience with not being motivated to do something, so the introduction for the LQ group asked them to put themselves in such a situation. The LQ group's instructions in studies 3 and 4 said: "Please imagine, you had no interest in the subsequent questions, but your only interest is to attend the lottery." Studies 1 and 2 were more exploratory: One our of three instructions was chosen randomly to provoke different kinds of meaningless data (rushing, careless responding, intended faking). Studies 3 and 4 then focused on careless responding.

The instructions in studies 3 to 4 also announced a question at the questionnaire's end—the manipulation check—that, unlike the other questions, would require an honest and careful answer. The manipulation check was then preceded by another eye-catching instruction to answer this question honestly and carefully.

4.2 Participants

To collect the data in a realistic situation, and to avoid wasting the respondents' valuable efforts, all studies were coordinated with other research projects. While the respondents in the HQ group actually participated in an academic survey, additional participants in the LQ group did the same questionnaire under the "meaningless data" condition.

Participants for the studies were randomly drawn from the *SoSci Panel* (for sample sizes see table 1, below), an academic access panel whose participants receive no compensation for completing questionnaires (Leiner, 2016a). The topic of study 4 required to invite only pool respondents who identified themselves as employees (age median 42 years), whereas the demographics of the studies 1 to 3 resemble the pool's demographics (40-50% employees, 35-39% students, 80-88% matriculation standard and above, age median 28-36 years, 52-62% female). In the pilot studies 1 and 2 there was a time lag between the original survey (HQ group) and a second sample forming the LQ group, which imposes restrictions on group comparability—although these are of minor relevance, compared to traditional experiments. In studies 3 and 4, all respondents were invited at the same time, and assigned randomly to the experimental conditions.

4.3 Questionnaires

The questionnaires were designed by the respective researchers in distinct research projects (Beckert, Koch, & Jakubowitz, 2018; Leiner, 2016b; E. E. Schneider, Schönfelder, Wolf, & Wessa, 2017), and amended by few questions to allow for analyses on meaningless data. The questionnaires' contents are very different and there is some variation in question formats (see below), which avoids overestimating specific indicators in the subsequent analyses. Studying the quality of answers is only possible if the questionnaires meet methodological standards. The *SoSci Panel*'s terms of use support this objective as they demand sufficient pretesting, and questionnaires undergo a peer review prior to the survey.

4.4 Manipulation Check

Records were used for analyses regardless of whether optional questions were answered (see table 1 for details about forced-choice/optional questions), but only if the respondents clicked through all pages of the respective questionnaire. In the pilot studies 1 and 2 only respondents in the LQ group were asked to complete a manipulation check, in studies 3 and 4 all respondents were asked for an honest rating of *how* they completed the questionnaire. The manipulation check contained two bipolar items per LQ group sub-condition in studies 1 and 2 (6-point response format), and two items in studies 3 and 4 ("answered superficially/thoroughly", "hardly/completely read the questions", 4-point response format). According to the manipulation

check, several respondents in the LQ group were too careful for this kind of a study. In studies 1 and 2 records were removed from the LQ group if respondents reported that they had rudely violated the experimental instructions (table 1). For example, if they had given particularly correct and careful answers (average ≥ 5 of 6), while they should have responded carelessly, or they had answered slowly and thought about questions when they should have rushed. In studies 3 and 4, records were removed from both, the HQ and LQ groups, if the average rating given in the two self-report items was below the scale's middle or above, respectively.

4.5 Measures

The goal of this study was to find those quality indicators that could separate "good" from "bad" records. Five categories of non-reactive data quality indicators were introduced above (missing data, response patterns, distance from the sample means, correlation structure, completion time), and we computed a series of indicators for each category.

To quantify the (A) *amount of missing data*, we must distinguish compulsory from optional questions, and questions offering a "don't know" (DK) option from those that do not. Open-ended text inputs where a response is not necessarily expected (e.g., inputs for "other" and collection of arguments or word associations) need special attention. One strategy is to exclude such variables when computing indices for missing data, another strategy is to weight each "miss" with the probability that the variable is answered by the overall sample (the quotient of cases where the variable was answered and those where it was not). Both strategies were employed, whereas any answer—meaningful or not—counted as an answer. Multiple-choice checkbox questions that allow none, one, or multiple options to be checked, were excluded from the indices: In the same way as forced-choice questions, they cannot *not* be answered, unless there is a minimum of options to check. From a practical point of view, it does not matter whether such variables are excluded from the index or not, because their inclusion does not add variance to the index, only changes the absolute percentages of missing responses. The percentage of DK responses was then used as third indicator for data quality, as choosing the DK option could indicate a lack of motivation or understanding (Shoemaker, Eichholz, & Skewes, 2002, p. 195).

Matrix-style item batteries (scales) are analyzed to find (B) *visual response patterns*. Reversed items that had been re-coded during data collection were re-reversed so that the response code of every item resembled the column in the matrix (1=first response option from left, 2=second from left, etc.). A series of indicators for visual patterns was tried, and an obvious one was skipped: The number of straightlined (short-)scales, i.e., scales where each item received the same response. This indicator strongly depends on the number of item batteries and their length, and provides little distinction. More differentiation is provided by the "longest string" (Johnson, 2005, p. 109), which is the length of the longest sequence of the same answer within an item battery. To account for other patterns, mathematical functions and algorithms were employed to compute indices (Baumgartner & Steenkamp, 2001; Jong, Steenkamp, Fox, & Baumgartner, 2008; Van Vaerenbergh & Thomas, 2012). For each indicator, the index was computed for every matrix-style question battery separately, and then these partial indices were averaged.

First, the standard deviation (SD) was chosen to indicate straightlining with minor deviations (Barge & Gehlbach, 2012). If the respondent checks options in a nearly straight line, the SD is close to zero, regardless if the response format is a 5, 6, or 7-point scale. Second, an algorithm was created, giving one point if two subsequent items receive the same answer (detecting straightlining), one point if the change between subsequent items is the same like the recent change (detecting diagonal lines), and half a point if the change is the same as the next-to-recent change (detecting left-right clicking). No more than one point is given per item, and the point sum is divided by the number of items (k) minus one, resulting in a value between 0 and 1. Third, pretests with manufactured patterns show that the absolute second derivation (d) of response values (r_i) is sensitive to straight, diagonal, and zigzag lines:

$$\begin{aligned} d &= \text{mean}(\text{abs}(\text{diff}(\text{diff}(r)))) \\ &= \frac{\sum_{i=1}^{k-2} |r_i''|}{k-2} \\ &= \frac{\sum_{i=1}^{k-2} |r_{i+2} - 2r_{i+1} + r_i|}{k-2} \end{aligned}$$

Two indicators are based upon the (C) *distance from the sample means*, effectively identifying atypical records (outliers). One indicator is the absolute z-scored response per item, averaged over all scale items. Inter-case z-standardization levels the items' different standard deviations and toughens the index against missing responses. The other indicator is the Mahalanobis distance (Johnson, 2005; Mahalanobis, 1936), which is a multivariate measure. Missing responses pose a significant challenge for the Mahalanobis distance, therefore, variables with more than 20% item non-response are excluded, and the covariance matrix is computed pairwise.

The (D) *correlation structure* puts the focus on records that influence correlations between variables in an atypical way. The even-odd consistency (Johnson, 2005; Meade & Craig, 2012), as the first indicator, requires the scale batteries being half-split into even and odd items. An index value (mean) is computed for each half set of items after recoding reversed items. This procedure results in two series (even and odd) of k index values per respondent, where k is the number of scale questions. The within-subject correlation

Table 1
Description of the studies, sample sizes, and deletion of records after manipulation check

	Study 1 (Pilot Study)		Study 2 (Heterogeneous Data)		Study 3 (Replication I)		Study 4 (Replication II)	
Questionnaire topic	political opinions and attitude characteristics (constant attitude issue)		political opinions and attitude characteristics (randomized issue)		stress perception, physical condition, and stress-related behaviors		interpersonal communication in organizations	
Underlying academic study/publication	Leiner (2016b) (studies 1 and 2 are based on the same survey)				E. E. Schneider, Schönfelder, Wolf, and Wessa (2017)		Beckert, Koch, and Jakubowitz (2018)	
Questionnaire characteristics[a]	10 scales (2–10 items, ∑ 65 items) plus 42 closed-ended and 1 open-ended questions, 2 questions asking for arguments (also open-ended), length 18.5 Min., 16 pages (wave 1 questionnaire, only)				5 scales (5–30 items, ∑ 85 items) plus 12 closed and 1 open-ended questions, 12.5 Min., 17 pages		12 scales (4–10 items, ∑ 85 items) plus 9 closed and 2 open-ended questions, 11.6 Min., 18 pages	
Forced-choice items	first 21 of 102 closed-ended single-choice items				all		none	
Sample specifics[b]	-		-		-		limited to employees	
Condition	LQ group	HQ group	LQ group	HQ group	LQ group	HQ group	LQ group	HQ group
Data collection (survey adm. period)	06/2013	07/2011–10/2012	11/2012	07/2011–10/2012	01–02/2016 (random. assignment)		08/2016 (random. assignment)	
Survey response rate[c]	26%	25%	27%	25%	31%	21%	18%	12%
Complete records	625	621	427	10580	418	854	411	727
Records removed[d]	134	-[e]	97	-[e]	50	6	90	4
Percentage	22%		23%		12%	1%	22%	<1%
Analysis sample	475	621	321	10580	368	848	321	723

The high-quality group (HQ) is the original sample of the respective *underlying academic study*. The low-quality group (LQ) completed the same questionnaire, but carelessly or giving fake responses.

[a] The length is the HQ group's median completion time. [b] All samples were recruited from the SoSci Panel (Leiner, 2016a) and are very similar in their characteristics, except for study 4. [c] Minimum response rate (response rate 1, AAPOR). [d] Records of respondents were removed when they reported in the manipulation check that they did not follow the instructions to complete the questionnaire either carefully (HQ) or carelessly (LQ). [e] There was no manipulation check for the HQ groups in studies 1 and 2.

coefficient between these series is a combined measure of consistent responding within the scales and differentiating between the scales. Another measure for intra-scale consistency is inspired by the idea of using regressions (Burns & Christiansen, 2011; Jandura et al., 2012): Within each scale battery, linear regression models predict every scale item's response based on the responses received for the other scale items. The absolute residuals for each item are averaged per scale. To create an index of scale inconsistency, the average scale residuals are again averaged. A large index of residuals then indicates arbitrary responding. Note, that only intra-scale consistency is used as an indicator. Residuals from an all-dataset-model (inter-measure consistency) were not tested, as such an indicator is prone to increase type I errors in hypotheses testing when used for cleaning data.

The (E) *completion time* (also known as *response time*) was server-side recorded for each page in the questionnaire, with the pages usually containing several questions and/or items. In studies 1 and 2 some pages (the instructions and the manipulation check) showed different contents, depending on the experimental condition; these were removed from analyses. The same would be necessary if filter questions were to vary the content substantially. The first indicator then is the absolute time spent to complete all (relevant) pages. The second indicator is the same, but after replacing outlier times by the typical completion time for the respective page. As the distribution of completion times is heavily skewed (e.g., skewness = 12, kurtosis = 151 for the overall completion time in study 1), the per-page medians serve as *typical completion time*, and an outlier is defined as taking 3/1.34 times the interquartile range (IQR) longer than the median completion time (this would be 3 *SD*, if the distribution was normally distributed). The third indicator is an index of relative completion speed: For each page, the sample's median page completion time is divided by the individual completion time, resulting in a speed factor. A factor of 2 means that the respondent has completed a page twice as fast as the typical respondent. An average speed factor per respondent is computed after the page factors are clipped to a maximum value of 3. This avoids disqualifying respondents who incidentally skip a single page. The limit of 3 is based on trials with the data from studies 1 and 2. Therefore this measure's efficiency is possibly overestimated for those two studies

Although this paper is about non-reactive indicators, studies 3 and 4 also include few (F) *reactive indicators* to put the non-reactive indicators' performance into perspective. Study 3 employs an instructional manipulation check (IMC, Oppenheimer, Meyvis, & Davidenko, 2009, p. 868) and a variant of the IMC that is passed more easily (figure 3). Both variants aim to indicate whether the instructions have been read and understood. Study 4 includes either the easier IMC variant, or bogus items like "I am currently filling out a questionnaire" (Hargittai, 2009; Meade & Craig, 2012), or instructed response items like "please select «fully disagree» in this line" (DeSimone et al., 2015). In the latter both conditions, three such items were placed in different scale batteries.

An indicator's effectiveness is quantified by its capability to correctly identify the records from the LQ group (*true positive*). Given the indicator's distribution for the overall sample, a threshold/cut-off value (percentile) is calculated for every indicator that identifies (*predicted positive*) as many records as there are records in the LQ group. The primary performance criterion then is the percentage of LQ records that have been correctly identified by the chosen cut-off value. This metric is also known as *sensitivity, true positive rate*, or *hit rate*. An ideal indicator has a sensitivity of 100%; it identifies all LQ records and none of the HQ records. The distribution of some indicators lacks differentiation (many records have the same indicator value) and does therefore not allow for a cut-off value that identifies the exact number of records. In that case, a larger number (*over-identification*) of records is identified as *predicted positive*, and the sensitivity is linearly corrected for that over-identification. While the *sensitivity* gives an impression for practical application, it depends on the relation of the LQ and HQ group sizes, and is specific for the chosen cut-off value. Therefore, the *area under the curve* (*AUC*, Fawcett, 2006; Hanley & McNeil, 1982) is calculated as a secondary criterion, describing the indicators' accuracy, taking chances and varying cut-off values into account. The fact that there is no manipulation check for the HQ groups in studies 1 and 2 causes us to underestimate the sensitivity and *AUC* but does not change the indicators' relative ranking.

5 Pilot (Study 1)

Studies 1 and 2 were conducted with the same questionnaire about "public opinion". This questionnaire starts with polling opinions on public issues (allowing "don't know" but no missing data), and then asks detailed questions on one of these issues (attitudes, relation to values, ambivalence, elaboration, uncertainty) mostly by means of five short multi-item scales, presented in a matrix layout (see Leiner, 2016b for the questionnaire). A significant amount of formally missing data was generated by open-ended questions asking for arguments pro and contra the issue (Cappella, Price, & Nir, 2002). In Study 1 the detail questions (the second part of the questionnaire) were about the same political issue for all respondents.

The LQ group was subdivided to cover three possible origins of meaningless survey data: (1) rushing, only in study 1, (2) careless responding, and (3) intended faking (Nichols et al., 1989), with two possible treatment instructions for each sub-condition (Appendix A.2). These six instructions were randomly assigned to the respondents.

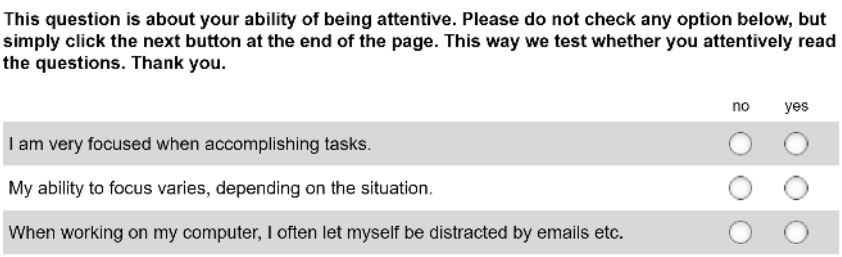

Figure 3. Simplified version of the instructional manipulation check (S-IMC). Instructions in a survey are often repetitive. Just screening such instructions seems a sufficient strategy, and does not necessarily reduce data quality. The simplified IMC therefore gives the instruction away in the first line. The questionnaire allows to uncheck the radio buttons, in case that the instruction is read only after answering has been started. The figure is a translation from the German version employed in the questionnaire.

5.1 Results

After removing records from the LQ group that have failed the manipulation check, 475 records with meaningless data had to be distinguished from 621 mostly meaningful records (table 1). The resulting random chance to correctly identify a record from the LQ group is 0.43, which is also the baseline for the indicators' *sensitivity* as listed in table 2.

In the overall sample, not separated by sub-condition, the indicators based on completion times are most successful in identifying records from the LQ group. These indicators identify about 66% of the LQ records, which, for comparison, corresponds to Nagelkerke's pseudo R^2 of .26 when understood as binomial regression (completion time with outliers replaced). The scales' even-odd consistency and the weighted non-response are considerably less efficient. For the latter, indicators based on item non-response, a good part of the drop can be attributed to a lack of differentiation: Near the cut-off value, many records share the same non-response rate, so far too many records exceed the cut-off value (over-identification). The other indicators barely exceed random chance or indicate the LQ records even worse than chance, such as the number of DK responses and the simple distance from the sample mean $AUC < 0.5$).

When it comes to different kinds of meaningless data (sub-conditions), the LQ groups are smaller while the HQ group remains the same. Therefore, random chance decreases. As shown in table 2 (right columns), the *careless responding* sub-condition reflects the indicators' performance observed for the overall sample. In the *rushing* sub-condition, completion time is the *only* relevant indicator. This suggests that the manipulation failed: Doing a questionnaire as fast as possible (as the rushing instruction asked for) does not necessarily provoke meaningless responses. Rushing is more likely *one* possible outcome of careless responding than its cause. Consequently, the rushing sub-condition is not used for the subsequent studies. In the *faking* sub-condition, completion time performs much worse than in the other two sub-conditions. The weighted non-response is the only indicator to identify faked records. Yet, its fair performance may be an artifact: The participants in the faking sub-condition were instructed not to disclose any true information about themselves (Appendix A.2), which might be understood as not answering at all. Looking at this pragmatically, intended faking is virtually invisible with the indicators applied in study 1.

6 Heterogeneous Data (Study 2)

Study 2 used the same questionnaire as study 1, but respondents were randomly assigned to one of 17 different political issues in the second questionnaire part. The samples in studies 1 and 2 already have substantial variance in age and location, but not in education. The variation triggered in study 2 increases heterogeneity in response behavior like we would expect, for example, in a representative sample.

6.1 Results

The HQ group in study 2 is much larger ($n_{\text{HQ}} = 10,580$) than in study 1. As there is some probability to have meaningless data in the HQ group as well, we must understand the efficiency presented in table 3 as a conservative estimate. Not a single indicator can identify a substantial part of the *faked* records in the heterogeneous data from study 2, including the non-response rate that had shown a fair performance in study 1. We also find differences regarding *careless responding*: While completion time is, again, the most efficient indicator, (in)consistent responding cannot identify LQ records in study 2. On the other hand, effortless response patterns can play their strengths in study 2, nearly closing up to the sensitivity of completion time. The fact that this respectable sensitivity is not accompanied by a similarly convincing *AUC*

Table 2
Indicator efficiency in Study 1

	Overall sample (not separated by sub-condition)				Sensitivity by sub-condition		
Data Quality Indicator	sensitivity	not computed	over-identification	AUC	careless rushing	intended responding	faking
Item non-response (irrel. variables excluded)	0.489	0	+59.2	0.611	0.310	0.321	0.246
Item non-response (weighted)	0.529^{+}	0	+51.4	0.729	0.308	0.500^{+}	0.503^{+}
DK responses	0.392^{-}	0	−44.2	0.451	0.183^{-}	0.218	0.125^{-}
Straightlining (longest string)	0.477	0	+4.6	0.540	0.260	0.331	0.270
Straightlining (avg. within scale SD)	0.501	8	-	0.586	0.277	0.458	0.353
Patterns (algorithmic)	0.494	0	+1.1	0.566	0.287	0.340	0.287
Patterns (second derivation)	0.480	22	-	0.575	0.287	0.389	0.331
Average Item Distance from Sample Mean	0.326^{-}	8	-	0.355	0.205^{-}	0.063^{-}	0.074^{-}
Mahalanobis Distance from Sample Mean	0.491	0	-	0.557	0.313	0.417	0.279
Even-odd consistency (split-half scales)	0.556	29	-	0.640	0.282	0.403^{+}	0.294
Intra-scale residuals (inconsistency)	0.463	22	-	0.537	0.287	0.375	0.243
Absolute completion time	0.665^{+}	0	-	0.766	0.518^{+}	0.604^{++}	0.353
Absolute completion time (outliers replaced)	0.659^{+}	0	+0.6	0.760	0.487^{+}	0.597^{++}	0.360
Relative completion speed (speed index)	0.636^{+}	0	-	0.740	0.451^{+}	0.625^{++}	0.324
Random chance to identify a LQ record	0.433	-	-	0.500	0.239	0.188	0.180
(LQ group nLQ : HQ group nHQ)	(475:621)				(195:621)	(144:621)	(136:621)

The sensitivity indicates the indicator's efficiency to identify meaningless data (see chapter 4 for details on sensitivity, over-identification and AUC). The column Not computed indicates the number of records for which the indicator could not be computed, such records were treated as "not identified as meaningless."
$^{-}$ Below random chance $^{+}$ AUC ≥ 0.7 $^{++}$ AUC ≥ 0.8 $^{+++}$ AUC ≥ 0.9

Table 3
Indicator sensitivity in Study 2 per sub-condition

Data Quality Indicator	Careless responding	Intended faking
Item non-response (irrelev. variables excluded)	0.033	0.008^{-}
Item non-response (weighted)	0.031	0.008^{-}
DK responses[a]	0.101	0.069
Straightlining (longest string)	0.050	0.040
Straightlining (avg. within scale SD)	0.262^{+}	0.073
Patterns (algorithmic)	0.252	0.106
Patterns (second derivation)	0.246	0.073
Average Item Distance from Sample Mean	0.041	0.024
Mahalanobis Distance from Sample Mean	0.139	0.114
Even-odd consistency (split-half scales)	0.000^{-}	0.016
Intra-scale residuals (inconsistency)	0.082	0.114
Absolute completion time	0.266^{++}	0.033
Absolute completion time (outliers replaced)	0.254^{++}	0.024
Relative completion speed (speed index)	0.246^{++}	0.041
Random chance	0.011	0.011
($n_{\text{LQ}} : n_{\text{HQ}}$)	(122:10580)	(123:10580)

For intended faking (right column), the AUC does not exceed 0.624. Due to the large sample, over-identification was not an issue
[a] Except for DK responses (22% / 16%).
$^{-}$ Below random chance, $^{+}$ AUC ≥ 0.7, $^{++}$ AUC ≥ 0.8, $^{+++}$ AUC ≥ 0.9.

suggests that the strength of effortless patterns lies in identifying a rather specific part of careless responding.

7 Replication I (Study 3)

Study 3 employed a questionnaire on stress perception, physical condition, and stress-related behaviors (E. E. Schneider et al., 2017). The questionnaire includes scales with considerably more items than studies 1 and 2. The longest scale consisted of 30 items that were presented on two pages but analyzed as one scale to obtain the pattern and consistency indicators (when treated as two scales, the indicators' performance would improve slightly). DK options are not offered by this questionnaire. To allow for comparison of non-reactive and reactive indicators, two instructional manipulation checks (the original IMC, and a simplified version) were included in the questionnaire.

The LQ group (n_{LQ} = 368) was asked to imagine that they had no interest in the questionnaire and only to fill it out to enter a lottery for a 25 € voucher. In response to concerns that respondents may answer "too careful" in the LQ group, an announcement was included that one would afterward have the option to complete the questionnaire carefully. These records did *not* become part of the HQ group. Respondents in the HQ group could also enter a lottery. In study 3, the respondents from the HQ group (n_{HQ} = 851) were assigned to two different conditions: They were either instructed to attentively read the questions and complete the questionnaire very carefully (n_{H1} = 413) or were not given such an instruction (n_{H2} = 438). The instructions in the LQ and HQ groups, if given, were labeled "important advice" and highlighted visually.

7.1 Results

The instructional manipulation check (IMC) can correctly identify 365 out of 368 LQ records (99%). This seems a compelling rate, but the IMC also identifies 347 of 852 HQ records (41%) as meaningless data. According to other studies with respondents from the same access panel, it is very unlikely that more than 5% of the HQ records actually contain meaningless data. Such an over-sensitive indication is not untypical for the IMC (Revilla & Ochoa, 2014). Table 4 accounts for the over-identification and therefore reports a sensitivity of about 0.6 for the IMC, instead of 0.99. This is, of course, only a theoretical value, since the IMC does not provide any differentiation that would allow for not losing substantial parts of the meaningful records.

Over-identification was better for the simplified IMC, but even this version misidentified 30% HQ records, while 87% of LQ records were identified correctly. The completion time that showed above-average performance in studies 1 and 2, and also performs best in study 3, achieves an identification rate similar to the IMC (87% for the relative completion speed) while losing much fewer records from the LQ group (6%). This suggest following Aust et al. (2012, no. pg) who

Table 4
Indicator efficiency in Study 3

	LQ v. HQ no instruction		LQ v. attentive instruction	
Data Quality Indicator	Sensitivity	*AUC*	Sensitivity	*AUC*
Item non-response (exclusion)	0.375^{-}	0.231	0.385^{-}	0.205
Item non-response (weighted)	0.163^{-}	0.239	0.160^{-}	0.209
Straightlining (longest string)	0.484	0.614	0.522	0.640
Straightlining (within-scale SD)	0.625^{+}	0.712	0.630	0.691
Patterns (algorithmic)	0.633	0.695	0.639^{+}	0.711
Patterns (second derivation)	0.641^{+}	0.712	0.663^{+}	0.715
Avg. Item Distance	0.318^{-}	0.306	0.361^{-}	0.339
Mahalanobis Distance	0.622	0.664	0.628	0.669
Even-odd consistency	0.394^{-}	0.410	0.399	0.390
Intra-scale residuals	0.571	0.611	0.579	0.615
Absolute completion time	0.883^{+++}	0.944	0.883^{+++}	0.951
Abs. completion time (outliers)	0.894^{+++}	0.959	0.897^{+++}	0.964
Relative completion speed	0.897^{+++}	0.961	0.905^{+++}	0.966
IMC (solved perfectly)	0.574	0.689	0.609^{+}	0.715
IMC (clicked title)	0.578	0.697	0.612^{+}	0.725
Simplified IMC	0.690^{+}	0.772	0.743^{++}	0.802
Random chance	0.457	0.500	0.471	0.500
($n_{\text{LQ}} : n_{\text{HQ}}$)	(368:438)		(368:413)	

The LQ group ($n_{LQ} = 368$) was compared either to an HQ group that received no instruction how to complete the questionnaire or to an HQ group that received an *attentive instruction*.
$^{-}$ Sensitivity below random chance, $^{+}$ $AUC \geq 0.7$, $^{++}$ $AUC \geq 0.8$, $^{+++}$ $AUC \geq 0.9$.

state that the applicability of the IMC is "limited to studies in which data quality is dependent on the careful reading of instructions."

Results, when controlled for random chance, do not show systematic differences between the HQ group's sub-conditions employed in study 3. Records from the LQ group are no easier to identify if the HQ group was instructed to complete the questionnaire carefully, than if they were not.

8 Replication II (Study 4)

Study 4 employed yet another questionnaire, researching interpersonal communication in organizations (Beckert et al., 2018; Breitsohl & Steidelmüller, 2018). Only employees having a supervisor were allowed for this study. The instruction for LQ group respondents was to imagine that they had no interest in the questionnaire, but only wanted to attend the lottery. They were explicitly asked *not* to complete the questionnaire carefully. Like in study 3 an option was announced to complete the questionnaire carefully, afterwards, and again, these respondents did not become part of the HQ group.

Both, the HQ and LQ groups were split into three conditions that employed different reactive quality indicators. The questionnaire either contained a simplified IMC, or three bogus items (items with only one possible response option), or three instructed response items (items saying which response option to check) spread throughout three of six scales (5 to 24 items). A fourth sub-condition without any reactive quality indicators was excluded from the analyses for its different overall number of items.

8.1 Results

Although a different issue was presented to a different sample in study 4, the overall indicator performance is similar to the previous studies (table 5). Again, completion time is the most effective non-reactive indicator. The IMC, a reactive indicator, again is disproportionately strict—but not the newly included reactive indicators: Instructed response items perform similarly good as completion time, and the bogus items can outperform any other indicator. Only three bogus items are sufficient to correctly identify 92% of the meaningless records.

9 Discussion

If participants take the time to plausibly falsify a questionnaire (faking), we are virtually unable to recognize this from the non-reactive indicators applied in studies 1 and 2. This finding was replicated with data from another online survey (Bergwinkl et al., 2018) that is not presented in this paper,

Table 5
Indicator efficiency in Study 4

Data Quality Indicator	Sensitivity	Over-identification	*AUC*
Item non-response (weighted)[a]	0.431	+24.9%	0.628
Straightlining (longest string)	0.399	+24.0%	0.571
Straightlining (within-scale SD)	0.567^{+}	-	0.711
Patterns (algorithmic)	0.486	-	0.633
Patterns (second derivation)	0.559^{+}	+0.3%	0.759
Avg. Item Distance	0.134^{-}	-	0.227
Mahalanobis Distance	0.321	-	0.437
Even-odd consistency	0.271^{-}	-	0.542
Intra-scale residuals	0.349	-	0.475
Absolute completion time	0.850^{+++}	-	0.947
Abs. completion time (outliers replaced)	0.857^{+++}	-	0.950
Relative completion speed	0.857^{+++}	-	0.950
Simple Instructional Manipulation Check[b]	0.384	+111.8%	0.631
Instructed Response[c]	0.851^{+++}	+3.1%	0.919
Bogus Items[d]	0.923^{+++}	−1.1%	0.951
Random chance	$0.307^{b,c,d}$		0.500
$(n_{LQ} : n_{HQ})$	(321:723)		

[a] The performance of weighted and non-weighted item non-response was nearly identical. Statistics on reactive indicators (end of the table) are based on subsets with $N =$ [b]343, [c]330, [d]323 (random chance [b] 0.297, [c] 0.335, [d] 0.297).

[-] Sensitivity below random chance, [+] $AUC \geq 0.7$, [++] $AUC \geq 0.8$, [+++] $AUC \geq 0.9$.

researching the perception of transgender persons. The questionnaire had been answered by a substantial number of respondents who had publicly stated that they would try and disturb the survey (*survey trolls*). As the malicious respondents could reliably be identified by other means (timestamps and the HTTP referer), the non-reactive indicators could be tested. The distances from the sample means achieved the best *AUC* with 0.73, which confirmed that the survey trolls would have been successful in biasing the means, but also that intended faking is barely identifiable by non-reactive indicators. In this respect, interviews sophistically falsified by respondents substantially differ from interviews falsified by lazy interviewers (Menold & Kemper, 2014). The present studies cannot tell if reactive indicators had performed better, but there are reasonable doubts: Consciously intended, motivated faking seems to involve a certain amount of attention to the questionnaire. A respondent making up coherent responses will likely pass attention checks.

Much more promising are the results on careless responding: The response to the first research question (RQ1) is that a substantial share of meaningless records can be identified by completion time. This indicator class consistently shows the best performance among the non-reactive indicators. A lack of response variation (near-straightlining) also achieves a respectable identification rate for careless responses in studies 1 and 2. Yet, the replication of this good result fails in studies 3 and 4 where different questionnaires are employed. The other classes of non-reactive indicators (item non-response, untypical responses, and inconsistent answering within scales) are of little help in identifying meaningless data.

Studies 3 and 4 also employ reactive indicators to allow direct comparison to the non-reactive indicators. Instructed response items identify careless records similarly good as completion time. Bogus items outperform completion time, although this paper cannot address the question, whether this outstanding performance can be generalized: Bogus items were employed only in a sub-sample of only one study.

10 Implications and Recommendations

Two recommendations can be derived from these results. The first is, to make use of completion times to identify meaningless data in web-based surveys. At least when the questionnaire does not require respondents to look up information or is otherwise more quickly to complete for experts. The integration of multiple indicators is beyond the scope of this paper but could improve the identification of meaningless data. The second recommendation is to reduce reliance on *post-hoc* analyses by, for example, sprinkling a few bogus items throughout different scale batteries of the

questionnaire. As an orientation: Study 4 used three items "I'm currently filling out a questionnaire", "I have never ever used a computer", and "My supervisor was born on February 30th." (note, that the questionnaire in study 4 was about the respondent's supervisor). Negative effects of such items have been discussed (Curran, 2016; Goldsmith, 1989), yet Breitsohl and Steidelmüller (2018) present empirical evidence that these bogus items have little effect on response behavior.

The identification of meaningless data is part of a larger data cleaning process (Appendix B), and an important decision is what to do with records that have been identified as meaningless? The answer depends on the survey. Losing records to deletion has much worse implications in a representative sample than in a convenience sample, and yet it could be the preferable choice. And even when the data is from a convenience sample, there is no data cleaning without side effects: Completion times are independent of most constructs measured in the questionnaires, but still correlate to some of them. Data not presented here shows a moderate correlation between completion time and political interest in studies 1 and 2, for example. On the contrary, what do we gain by removing meaningless records? This paper did not discuss to what degree meaningless data distorts the results. Greszki, Meyer, and Schoen (2015) as well as Moran and Cutler (1997) argue that those responses may not affect the results at all. Preliminary results for the above studies suggest that it largely depends on the research question and questionnaire design. Meaningless responses may only increase statistical noise or may cause substantial biases and type I errors.

Even the best available non-reactive indicators identify only a fraction of the problematic records, and the present findings on bogus items need replication. Therefore, it is still an uncertain diagnosis, whether a questionnaire was completed carefully or not. If there are arguments to remove *probably meaningless* records, this means to also remove some valid questionnaires. The amount of valid data that is lost depends on the percentage of problematic records in the data set.

Regarding the question for a definite cut-off value (DeSimone et al., 2015, p. 179), this paper can only give a rough direction. In the above studies, records were designed to be either meaningful or meaningless. In general, we face a broad continuum between completely meaningless answers and painstaking accuracy. The results for the IMC (studies 3 and 4) and the lack of fully inconsistent ("random") response behavior suggest that some average attention is much more typical for online survey respondents than utmost diligence or completely meaningless responding. Facing this continuum, it might be a good strategy to focus on the most careless records, assuming that these threaten data quality the worst.

The average *relative completion speed* is a good candidate to screen for meaningless data, as it allows comparison between different questionnaires. Its cut-off to distinguish the LQ from the HQ group in studies 1, 3, and 4 is about 1.3 (study 2 is excluded here, because even a small percentage of careless records in the exceedingly large HQ group biases the cut-off value). This cut-off was calculated to estimate the potential of the indicator. A pragmatic recommendation is to use a much more lenient cut-off of 2.0 to identify *particularly suspicious* records. For those, we must assume that the respondents have not read the questions at all. Depending on the questionnaire and sample, one may assess additional indicators beyond the scope of this paper, such as the presence of meaningful open-ended responses.

One restriction, of course, is crucial before removing records with a *relative speed index* above 2.0: If the questionnaire asks for facts or knowledge instead of opinions, then completion time is not a valid indicator for meaningless data. Experts are obviously faster in giving facts than non-experts who must select the information from the filing cabinet. This does not mean that the experts' information would be meaningless. The same applies to Internet newcomers and frequent users: the latter will do a web-based questionnaire faster, but do not necessarily answer less carefully.

11 Limitations

The manipulation employed in studies 1 to 4 can provoke meaningless data only to a limited degree: Respondents from the LQ groups often "failed" to respond carelessly. At the same time, there may be some careless responding in the HQ groups as well. This causes predominantly conservative estimates of indicator efficiency. Also, some liberties were taken regarding group assignment in studies 1 and 2: The HQ group (the reference group in both studies) and the LQ groups were invited separately at different points of time. This limitation has been addressed in studies 3 and 4. Finally, this study includes neither self-reports nor scales designed to measure inconsistent or faked responding. Studies 3 and 4 give some impression on how effective reactive measures for data quality could be, but do not cover the full range of such measures.

12 Conclusion

Meaningless data is a complex issue. Different motivations and behaviors of the respondents cause disperse outcomes: Some respondents give inconsistent answers, others skip items, reduce differentiation or follow further response styles (Van Vaerenbergh & Thomas, 2012). Yet, none of these patterns is characteristic of the majority. The closest thing to a common characteristic of careless responding is the motivation to save time: Completion times were found to be a useful indicator of meaningless data throughout four studies. But in summary, the accuracy of isolated non-reactive indicators is limited. A direction might be to address the

complexity of different response behaviors by using multiple non-reactive indicators at the same time. First attempts suggest a non-linear combination so that each indicator only identifies particularly abnormal cases in its own respect.

This paper leaves several questions unanswered: In which settings are reactive indicators, such as bogus items, a better replacement for non-reactive indicators? To what degree does the removal of meaningless records affect the results' quality? Shall we accept the possible removal of carefully completed records? And what are the ethical implications of evaluating metadata that respondents don't know about?

Collecting survey data through the Internet is more important for the social sciences than ever before, but this interview mode is still facing substantial challenges. Not only in view of what became known as the *replication crisis* in the social sciences, it is probably good advice to seek a better understanding of problematic data and its impact on research results.

References

Ajzen, I. (1988). *Attitudes, personality and behavior. Mapping social psychology series*. Chicago: Dorsey.

Allen, I. L. (1966). Detecting respondents who fake and confuse information about question areas on surveys. *Journal of Applied Psychology*, *50*(6), 523–528.

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2012). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-012-0265-2

Azfar, O. & Murrell, P. (2009). Identifying reticent respondents: Assessing the quality of survey data on corruption and values. *Economic Development and Cultural Change*, *57*(2), 387–411. doi:10.1086/592840

Barge, S. & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education*, *53*(2), 182–200. doi:10.1007/s11162-011-9251-2

Bauermeister, J. A., Pingel, E., Zimmerman, M., Couper, M., Carballo-Dieguez, A., & Strecher, V. J. (2012). Data quality in HIV/AIDS web-based surveys: Handling invalid and suspicious data. *Field Methods*, *24*(3), 272–291. doi:10.1177/1525822X12443097

Baumgartner, H. & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, *38*(2), 143–156. doi:10.1509/jmkr.38.2.143.18840

Beckert, J., Koch, T., & Jakubowitz, M. (2018). *Not what you say, but how you say it. Effects of managerial communication on employee-organization relationships*. Poster presented at the 7th European Communication Conference (ECC/ECREA).

Bergwinkl, S., Hille, L., Marinova, K., Nothvogel, R., Schwarzer, F., & Xu, A. (2018). *Weniger Vorurteile gegenüber Transgendern durch Kontakt über Youtube-Videos?* Unpublished manuscript.

Bethlehem, J. & Biffignandi, S. (2012). *Handbook of web surveys*. Hoboken, NJ: Wiley.

Bhaskaran, V. & LeClaire, J. (2010). *Online surveys for dummies*. Hoboken, N.J: Wiley.

Birnbaum, M. H. (2003). Methodological and ethical issues in conducting social psychology research via the internet. In C. Sansone, C. C. Morf, & A. T. Panter (Eds.), *The Sage handbook of methods in social psychology* (pp. 359–382). Thousand Oaks, CA: Sage.

Bishop, G. F., Oldendick, R. W., Tuchfarber, A. J., & Bennett, S. E. (1980). Pseudo-opinions on public affairs. *Public Opinion Quarterly*, *44*(2), 198–209.

Börkan, B. (2010). The mode effect in mixed-mode surveys: Mail and web surveys. *Social Science Computer Review*, *28*(3), 371–380. doi:10.1177/0894439309350698

Bowen, A. M., Daniel, C. M., Williams, M. L., & Baird, G. L. (2008). Identifying multiple submissions in internet research: Preserving data integrity. *AIDS and Behavior*, *12*(6), 964–973. doi:10.1007/s10461-007-9352-2

Breitsohl, H. & Steidelmüller, C. (2018). The impact of insufficient effort responding detection methods on substantive responses: Results from an experiment testing parameter invariance. *Applied Psychology*, *67*(2), 284–308. doi:10.1111/apps.12121

Burns, G. N. & Christiansen, N. D. (2011). Methods of measuring faking behavior. *Human Performance*, *24*(4), 358–372. doi:10.1080/08959285.2011.597473

Callegaro, M., Lozar Manfreda, K. L., & Vehovar, V. (2015). *Web survey methodology*. Los Angeles, CA: Sage.

Cappella, J. N., Price, V., & Nir, L. (2002). Argument repertoire as a reliable and valid measure of opinion quality: Electronic dialogue during campaign 2000. *Political Communication*, *19*(1), 73–93.

Cellagaro, M. (2013). Paradata in web surveys. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 261–279). Hoboken, NJ: John Wiley & Sons.

Converse, J. M. & Presser, S. (2003). Survey questions: Handcrafting the standardized questionnaire (20th ed.) In *Sage University Papers, Quantitative Applications in the Social Sciences* (Vol. 63). Newbury Park, CA: Sage.

Couper, M. P. & Bosnjak, M. (2010). Internet Surveys. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 527–550). Bingley: Emerald.

Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, *66*, 4–19.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2015). Best practice recommendations for data screening. *Journal of Organizational Behavior*, *36*(2), 171–181. doi:10.1002/job.1962

Diedenhofen, B. & Musch, J. (2017). PageFocus: Using paradata to detect and prevent cheating on online achievement tests. *Behavior Research Methods*, *49*(4), 1444–1459. doi:10.3758/s13428-016-0800-7

Dillman, D. A. (2013). *Mail and internet surveys*. Hoboken, N.J.: Hoboken, N.J.

Draisma, S. & Dijkstra, W. (2004). Response latency and (para)linguistic expressions as indicators of response error. In S. Presser, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, & E. Singer (Eds.), *Wiley Sreries in Survey Methodology. Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: John Wiley & Sons.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. doi:10.1016/j.patrec.2005.10.010

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research. In C. Hendrick & M. S. Clark (Eds.), *Review of Personality and Social Psychology: Vol. 11. Research methods in personality and social psychology* (pp. 74–97). Newbury Park: Sage.

Fowler, F. J. (2009). Survey research methods (4th ed.) In *Applied Social Research Methods Series* (Vol. 1). Los Angeles: Sage.

Franzén, M. (2011). *Nonattitudes/pseudo-opinions: Definitional problems, critical variables, cognitive components and solutions*. (C/D Extended Essay No. 14). Retrieved from http://www.diva-portal.org/smash/get/diva2:1032161/FULLTEXT01.pdf

Furnham, A., Hyde, G., & Trickey, G. (2013). On-line questionnaire completion time and personality test scores. *Personality and Individual Differences*, *54*(6), 716–720. doi:10.1016/j.paid.2012.11.030

Goldsmith, R. E. (1989). Reducing spurious response in a field survey. *The Journal of Social Psychology*, *129*(2), 201–212. doi:10.1080/00224545.1989.9711721

Göritz, A. S. (2004). The impact of material incentives on response quantity, response quality, sample composition, survey outcome, and cost in online access panels. *International Journal of Market Research*, *46*(3), 327–345.

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the effects of removing "too fast" responses and respondents from web surveys. *Public Opinion Quarterly*, *79*(2). doi:10.1093/poq/nfu058

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2011). *Survey Methodology* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Hanley, J. A. & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29–36.

Hardeman, S. (2013). Organization level research in scientometrics: A plea for an explicit pragmatic approach. *Scientometrics*, *94*(3), 1175–1194. doi:10.1007/s11192-012-0806-6

Hargittai, E. (2009). An update on survey measures of web-oriented digital literacy. *Social Science Computer Review*, *27*(1), 130–137. doi:10.1177/0894439308318213

Harzing, A.-W., Brown, M., Köster, K., & Zhao, S. (2012). Response style differences in cross-national research. *Management International Review*, *52*(3), 341–363.

Jandura, O., Peter, C., & Küchenhoff, H. (2012). *Die Guten ins Töpfchen, doch wer sind die Schlechten? Ein Vergleich verschiedener Strategien der Datenbereinigung [Picking the good ones, but which are the bad ones? Comparing different strategies of cleaning data]*. 14th annual conference of the DGPuK methods group, Sept. 27.-29., Zürich, Swiss.

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from web-based personality inventories. *Journal of Research in Personality*, *39*(1), 103–129. doi:10.1016/j.jrp.2004.09.009

Jong, M. G. d., Steenkamp, J.-B. E. M., Fox, J.-P., & Baumgartner, H. (2008). Using item response theory to measure extreme response style in marketing research: A global investigation. *Journal of Marketing Research*, *45*(1), 104–115. doi:10.1509/jmkr.45.1.104

Katz, D. (1968). Consistency for What? The Functional Approach. In R. P. Abelson, E. E. Aronson, W. J. McGuire, T. M. Newcomb, M. J. Rosenberg, & P. H. Tannenbaum (Eds.), *Theories of cognitive consistency* (pp. 179–191). Chicago: McNally.

Kemper, C. J. & Menold, N. (2014). Nuisance or remedy? The utility of stylistic responding as an indicator of data fabrication in surveys. *Methodology*, *10*(3), 92–99. doi:10.1027/1614-2241/a000078

Klirs, E. G. & Revelle, W. (1986). Predicting variability from perceived situational similarity. *Journal of Research in Personality*, *20*(1), 34–50.

Konstan, J. A., Rosser, B. R. S., Ross, M. W., Stanton, J., & Edwards, W. M. (2005). The story of subject naught: A cautionary but optimistic tale of Internet survey research. *Journal of Computer-Mediated Communication*, *10*(2). doi:10.1111/j.1083-6101.2005.tb00248.x

Kreuter, F. (2013). Improving surveys with paradata: Introduction. In F. Kreuter (Ed.), *Improving surveys with*

paradata: Analytic uses of process information (pp. 1–9). Hoboken, NJ: John Wiley & Sons.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236. doi:10.1002/acp.2350050305

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*(1), 537–567. doi:10.1146/annurev.psych.50.1.537

Krosnick, J. A. & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, *52*(4), 526–538. doi:10.2307/2749259

Krosnick, J. A. & Fabrigar, L. R. (2003). *"Don't Know" and "No Opinion" responses: What they mean, Why they occur, and how to discourage them*. Paper presented to the Workshop on Item Non-response and Data Quality in Large Social Surveys, Basil, Switzerland, 10/2003.

Kuhn, D. (1991). *The skills of argument*. Cambridge: Cambridge Univ. Press.

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. doi:10.1207/S15327752JPA7602_12

Kwak, N. & Radler, B. (2002). A comparison between mail and web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics*, *18*(2). Retrieved from http://www.barold.com/www/JOS%20article.pdf

Lavrakas, P. (2008). *Encyclopedia of survey research methods*. Thousand Oaks, CA: Sage.

Leiner, D. J. (2016a). Our research's breadth lives on convenience samples: A case study of the online respondent pool "SoSci Panel". *Studies in Communication | Media (SCM)*, *5*(4), 367–396.

Leiner, D. J. (2016b). *Stabilität öffentlicher Meinung. Wie der Charakter einer Streitfrage den Einfluss der Medien begrenzt*. Dissertation. Wiesbaden, DE: Springer VS.

Leiner, D. J. & Doedens, S. (2010). Test-Retest-Reliabilität in der Forschungspraxis der Online-Befragung. In N. Jackob, T. Zerback, O. Jandura, & M. Maurer (Eds.), *Methoden und Forschungslogik der Kommunikationswissenschaft: Vol. 6. Das Internet als Forschungsinstrument und -gegenstand in der Kommunikationswissenschaft* (pp. 316–331). Köln: Halem.

Lim, J. & Butcher, J. N. (1996). Detection of faking on the MMPI–2: Differentiation among faking-bad, denial, and claiming extreme virtue. *Journal of Personality Assessment*, *67*(1), 1–25. Retrieved from 10.1207/s15327752jpa6701_1

Mahalanobis, P. C. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India* (Vol. 2, *1*, pp. 49–55).

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public Opinion Quarterly*, *72*(5), 914–934.

Marsden, P. V. & Wright, J. D. (Eds.). (2010). *Handbook of survey research* (2nd ed.). Bingley: Emerald.

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

Menold, N. & Kemper, C. J. (2014). How do real and falsified data differ? Psychology of survey response as a source of falsification indicators in face-to-face surveys. *International Journal of Public Opinion Research*, *26*(1), 41–65. doi:10.1093/ijpor/edt017

Moran, G. & Cutler, B. L. (1997). Bogus publicity items and the contingency between awareness and media-induced pretrial prejudice. *Law and Human Behavior*, *21*(3), 339–344. doi:10.1023/A:1024846917038

Musch, J. & Reips, U.-D. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 61–87). San Diego, CA: Academic Press.

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. *Journal of Clinical Psychology*, *45*(2), 239–250. doi:0.1002/1097-4679(198903)45:2<239::AID-JCLP2270450210>3.0.CO;2-1

Olson, K. & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata: Analytic uses of process information* (pp. 43–72). Hoboken, NJ: John Wiley & Sons.

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. doi:10.1016/j.jesp.2009.03.009

Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Los Angeles, CA: Sage.

Payne, S. L. (1950). Thoughts about meaningless questions. *Public Opinion Quarterly*, *14*(4), 687–696.

Payne, S. L. (1980). *The art of asking questions*. Princeton, N.J.: Princeton Univ. Press.

Pine, D. E. (1995). Assessing the validity of job ratings: An empirical study of false reporting in task inventories. *Public Personnel Management*, *24*(4), 451–460.

Revilla, M. & Ochoa, C. (2014). What are the links in a Web survey among response time, quality, and auto-evaluation of the efforts done? *Social Science*

Computer Review, *33*(1), 97–114. doi:10 . 1177 / 0894439314531214

Rogers, F. & Richarme, M. (2009). *The honesty of online survey respondents: Lessons learned and prescriptive remedies*. Retrieved from https : / / www . decisionanalyst.com/whitepapers/onlinerespondents/

Schendera, C. F. G. (2007). *Datenqualität mit SPSS*. München: Oldenbourg.

Schneider, E. E., Schönfelder, S., Wolf, M., & Wessa, M. (2017). *Krank und gestresst? Subjektiv erlebter Stress in gesunden und klinischen Populationen: Validierung, Eigenschaften und Populationsunterschiede einer Deutschen Version der Perceived Stress Scale [Sick and stressed out? Validation, psychometric properties and group differences of a German Version of the Perceived Stress Scale in both healthy and clinical samples]*. Poster presented at the 10th Conference of the German Psychological Society (DGPs), Chemnitz, Germany.

Schneider, K. C. (1985). Uninformed response rates in survey research: New evidence. *Journal of Business Research*, *13*(2), 153–162.

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

Schuman, H. & Presser, S. (1980). Public opinion and public ignorance: The fine line between attitudes and nonattitudes. *American Journal of Sociology*, *85*(5), 1214–1225.

Selm, M. v. & Jankowski, N. W. (2006). Conducting online surveys. *Quality and Quantity*, *40*(3), 435–456. doi:10.1007/s11135-005-8081-8

Shin, E., Johnson, T. P., & Rao, K. (2012). Survey mode effects on data quality: Comparison of web and mail modes in a US national panel survey. *Social Science Computer Review*, *30*(2), 212–228. doi:10.1177/0894439311404508

Shoemaker, P. J., Eichholz, M., & Skewes, E. A. (2002). Item nonresponse: Distinguishing between don't know and refuse. *International Journal of Public Opinion Research*, *14*(2), 193–201.

Slomczynski, K. M., Powalko, P., & Krauze, T. (2017). Nonunique Records in International Survey Projects: The Need for Extending Data Quality Control. *Survey Research Methods*, *11*(1), 1–16. doi:10.18148/srm/2017.v11i1.6557

Sniderman, P. M. & Bullock, J. (2004). A consistency theory of public opinion and political choice: The hypothesis of menu dependance. In W. E. Saris & P. M. Sniderman (Eds.), *Studies in public opinion: Attitudes, nonattitudes, measurement error and change* (pp. 337–357). Princeton: Princeton Univ. Press.

Stieger, S. & Reips, U.-D. (2010). What are participants doing while filling in an online questionnaire: A paradata collection tool and an empirical study. *Computers in Human Behavior*, *26*(6), 1488–1495. doi:10.1016/j.chb.2010.05.013

Sue, V. M. & Ritter, L. A. (2012). *Conducting online surveys* (2nd ed.). Los Angeles: Sage.

Van Vaerenbergh, Y. & Thomas, T. D. (2012). Response styles in survey research: A literature review of antecedents, consequences, and remedies. *International Journal of Public Opinion Research*, *25*(2), 195–217. doi:10.1093/ijpor/eds021

Wang, R. Y. & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33.

Weisberg, H. F. (2009). *The total survey error approach: A guide to the new science of survey research*. Chicago, IL: University of Chicago Press.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*(3), 186–191. doi:0.1007/s10862-005-9004-7

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

Appendix A

Table A1
Cross-correlation between different data quality indicators (study 2)

	Indicator	1.a	1.b	2.a	2.b	3.a	3.b	4.a	4.b	5.a	5.b
1.a	Missing Data (absolute)		.75	.06	.07	−.04	−.08	.04	−.01	.10	.22
1.b	Missing Data (weighted)	.75		.11	.13	−.02	−.11	.01	−.04	.18	.28
2.a	Straightlining (within scale SD)	.06	.11		.78	−.37	−.67	−.11	−.71	.12	.17
2.b	Patterns (second derivation)	.07	.13	.78		−.22	−.58	−.15	−.68	.13	.19
3.a	Avg. Item Dst. f. Sample Mean	−.04	−.02	−.37	−.22		.44	−.37	.31	−.05	−.07
3.b	Mahalanobis Dst. f. Spl. Mean	−.08	−.11	−.67	−.58	.44		.07	.86	−.11	−.17
4.a	Even-odd consist: (split-half)	.04	.01	−.11	−.15	−.37	.07		.13	−.01	−.02
4.b	Intra-scale residuals (incons.)	−.01	−.04	−.71	−.68	.31	.86	.13		−.09	−.13
5.a	Fast Completion (absolute time)	.10	.18	.12	.13	−.05	−.11	−.01	−.09		.80
5.b	Fast Completion (index)	.22	.28	.17	.19	−.07	−.17	−.02	−.13	.80	

N = 10901. The table gives rank correlations (Spearman), as the absolute value of an indicator and its distribution is irrelevant for removal by cut-off (threshold).

Table A2
Instructions in studies 1 and 2

Sub-condition	Instruction
Rushing (1)[a]	Please complete this questionnaire as fast as possible.
Rushing (2)[a]	Please try and reach the questionnaire's end as quickly as possible.
Careless responding (1)	Please take as little care as possible in doing this questionnaire. Do this questionnaire deliberately carelessly.
Careless responding (2)	Please imagine that you're not interested in the questions, but your only interest is to attend the lottery.
Intended faking (1)	Please imagine that you're not interested in the questions, but your only interest is to attend the lottery – yet make your answers look authentic.
Intended faking (2)	Please disclose as little as possible about you and your opinion.

The instructions, like the questionnaire, were in German. [a] The "rushing" instructions were only employed in study 1, not in study 2.

Appendix B
Practical Application – Data Cleaning

The identification of meaningless data and the choice of how to handle such records is part of a larger data cleaning process. Step 1 of this process usually removes ineligible cases where respondents are not part of the population under research. In step 2 records may be removed for which important questions have not been answered: dropouts and records with substantial item non-response. The analytic value of incomplete cases is usually limited to estimating self-selection biases and the identification of problematic questions. Step 3 is the removal of multiple submissions by the same respondents (Bauermeister et al., 2012; Bowen et al., 2008; Konstan et al., 2005) or submissions for the same analysis unit from different respondents (data doublets). Multiple submission is often considered a minor problem, because doing the same survey twice is very unattractive in non-/low-incentivised, lengthy Internet surveys (Birnbaum, 2003, p. 372; Göritz, 2004). Should a study provoke multiple submissions, various techniques help with their identification (Musch & Reips, 2000). Data doublets are rarely an issue when the research units are respondents (Slomczynski, Powalko, & Krauze, 2017), but studies researching organizations (Hardeman, 2013) or households typically need deduplication and merging strategies to cope with heterogeneous reports on the same unit. Step 4 might identify and handle cases with meaningless data, which is the focus of this paper. Depending on the applied statistic methods, step 5 is to remove extreme outliers that would disproportionally skew statistical analyses and/or remove outlier responses from otherwise valid cases (for a discussion "to remove or not to remove" records in representative samples see Osborne, 2013, p. 165). The steps' order may vary throughout studies, and data cleaning may include fewer or further steps.