Detecting computer-generated random responding in questionnaire-based data: A comparison of seven indices

Marc Dupuis¹ · Emanuele Meier¹ · Félix Cuneo¹

Published online: 8 August 2018
© Psychonomic Society, Inc. 2018

Abstract

With the development of online data collection and instruments such as Amazon's Mechanical Turk (MTurk), the appearance of malicious software that generates responses to surveys in order to earn money represents a major issue, for both economic and scientific reasons. Indeed, even if paying one respondent to complete one questionnaire represents a very small cost, the multiplication of botnets providing invalid response sets may ultimately reduce study validity while increasing research costs. Several techniques have been proposed thus far to detect problematic human response sets, but little research has been undertaken to test the extent to which they actually detect nonhuman response sets. Thus, we proposed to conduct an empirical comparison of these indices. Assuming that most botnet programs are based on random uniform distributions of responses, we present and compare seven indices in this study to detect nonhuman response sets. A sample of 1,967 human respondents was mixed with different percentages (i.e., from 5% to 50%) of simulated random response sets. Three of the seven indices (i.e., response coherence, Mahalanobis distance, and person–total correlation) appear to be the best estimators for detecting nonhuman response sets. Given that two of those indices—Mahalanobis distance and person–total correlation—are calculated easily, every researcher working with online questionnaires could use them to screen for the presence of such invalid data.

Keywords Botnet · Functional method · Mahalanobis distance · Mechanical Turk · Person–total correlation · Random responding · Response coherence

For the last two decades, the Internet has become crucially important in every domain of the academic work, including data collection methods. With the development of online surveys, paid participation for online survey responding has become a current practice, as highlighted by the ubiquitous use of Amazon's Mechanical Turk (MTurk) and other crowd-sourcing instruments in some psychology and social science domains (Litman, Robinson, & Abberbock, 2017; Necka, Cacioppo, Norman, & Cacioppo, 2016). However, the use of such instruments remains questioned (Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013; Gleibs, 2017; Goodman, Cryder, & Cheema, 2013). Currently, online questionnaire-based research is vulnerable to problematic

respondent behaviors (Clifford & Jerit, 2014). Despite Briones and Benham's (2017) findings that very few differences exist between crowd-sourced and traditionally recruited or convenience samples, researchers have provided evidence that distant participant samples are more likely to complete surveys with less care than are local participant samples, irrespective of possible financial compensations (Litman, Robinson, & Rosenzweig, 2015). In addition, crowd-sourced samples are likely to consist of people who participate regularly in surveys, making them overrepresented (Buchanan & Scofield, 2018; Chandler, Mueller, & Paolacci, 2014). In worse-case scenarios, the presence of fraudulent respondents may be suspected (Chandler & Paolacci, 2017).

Malicious programs, generating purely invalid data in order to earn money (e.g., botnets, automated form-fillers, and survey bots) represent new and important economic and scientific threats to the completion of online questionnaires. Interestingly, few measures have been taken thus far to deal with nonhuman response sets.

Currently, the prevalence of invalid data sets completed by malicious programs can hardly be estimated. However,

Marc Dupuis and Emanuele Meier contributed equally to this work.

✉ Marc Dupuis
Marc.Dupuis@unil.ch

¹ Institute of Psychology, University of Lausanne, Bâtiment Geopolis 4122, CH-1015 Lausanne, Switzerland

arguments supporting the idea that such programs represent a major threat merit mention. Chandler and Paolacci (2017) reported that 14% to 18% of imposters could be found in online study samples with rare inclusion conditions. Sharpe Wessling, Huber, and Netzer (2017) highlighted the existence of MTurk user forums and websites listing strategies that bypass screening criteria, supporting the idea that the proportion of impostors in some online studies is not negligible. Users may freely download several pieces of software, such as survey bots or automated form-fillers, from the Internet. Such automated form-fillers can be easily programmed (e.g., Buchanan & Scofield, 2018). Liu and Wronski (2018) found that Captcha-type questions represent a promising trap technique to capture inattentive respondents; nonetheless, impostors are likely to bypass such steps by responding to the Captcha and then running automated form-fillers.

Malicious programs are likely to provide random response sets—namely, uniform deviates (Buchanan & Scofield, 2018; Holden, Wheeler, & Marjanovic, 2012; Meade & Craig, 2012). The main effects of random responding cover the mean level and variance in total scores (Osborne & Blanchard, 2011) and correlations (DeSimone, DeSimone, Harms, & Wood, 2018; Holtzman & Donnellan, 2017; Huang, Liu, & Bowling, 2015), including the factor structure of a psychological test or its reliability (DeSimone et al., 2018), which is at high risk of affecting both a study's representativeness and validity. Even a low proportion of random response sets can largely bias statistical analyses and distort their results (Credé, 2010). For instance, in psychological practice, random responding may result in misleading conclusions for clinical trials, due to the possible increase in both Type I and Type II error rates (Marjanovic, Struthers, Cribbie, & Greenglass, 2014; Osborne & Blanchard, 2011). Although the effects of random responding have been well-studied, random responding is still deemed a minor threat by many researchers. This is surprising, because not only extreme incidences of random responding could lead to bias (Credé, 2010; Holden et al., 2012; McGrath, Mitchell, Kim, & Hough, 2010; Niessen, Meijer, & Tendeiro, 2016).

In the scientific literature, random responding is most frequently considered a synonym for careless responding (Meade & Craig, 2012) or insufficient effort responding (Bowling et al., 2016). The underlying assumptions for this absence of differentiation are, first, that respondents who complete questionnaires carelessly provide responses following some random distributions; second, that there is a continuum between adequate responding (i.e., resulting in usable information) and random responding, with the latter considered as the most extreme form of careless responding. From this point of view, searching for botnets or automated form-fillers that provide completely randomly generated responses is relevant, considering that partially random responses are already problematic. These theoretical assumptions are bold and maintain

unclear notions that merit clarification. However, their problematic definitions are not within the scope of this article. In a more modest perspective, herein, *random responding* will refer to a narrow definition of providing truly randomly distributed response sets.

Different indices have been introduced to detect problematic human response sets under the assumption of their random distribution, which makes these indices eligible to detect non-human response sets. Niessen, Meijer, and Tendeiro (2016) regrouped such indices into different categories, including consistency indices, outlier statistics, and external measures.

- **Consistency indices** Consistency indices rely on the assumption that “sufficient effort” responding differs from random responding in terms of stability of response patterns and coherence regarding items related to the same dimensions. Five consistency indices exist based on *psychometric antonyms*, *psychometric synonyms*, *odd–even consistency* (Curran, 2016; Meade & Craig, 2012; Niessen et al., 2016; Ward & Meade, 2018), and a multi-dimensional index of *response reliability* resulting from Gendreau's functional method (Dupuis, Meier, Capel, & Gendreau, 2015). The fifth index, *response coherence*, is another that Gendreau developed to measure the part of variance in individual responses explained by the factor structure of the questionnaire. Despite being derived from different approaches, these five indices all measure stability in response patterns, a condition that random response sets hardly fulfill. The calculation procedure for these five indices is detailed below.
- **Outlier statistics** As their name suggests, outlier statistics are used to detect response sets that are particularly uncommon in comparison to the rest of a sample. They include two measures: Mahalanobis distance (Curran, 2016; Niessen et al., 2016; Ward & Meade, 2018) and person–total correlation (Curran, 2016). The Mahalanobis distance is used to detect multivariate outliers using a chi-square test to determine significantly different variable sets, and assumes they are outlying. The person–total correlation is based on a very different technique that consists of correlating an individual's responses on each item and the average value on each item for the entire group of respondents.
- **Long-string analysis** Long-string analysis consists of calculating the number of consecutive identical responses throughout a questionnaire. This index is intuitive and frequently used to detect insufficient-effort responding patterns (Curran, 2016; Johnson, 2005; Meade & Craig, 2012; Ward & Meade, 2018). However, it is not useful for identifying strict random responding.
- **External indices** External indices are additional measures collected with the questionnaire that are not a part of the questionnaire. These indices include response time (Buchanan & Scofield, 2018; Curran, 2016; Meade &

Craig, 2012; Niessen et al., 2016), bogus items (Curran, 2016; Meade & Craig, 2012; Niessen et al., 2016), explicit instructed response items (Niessen et al., 2016), and click count, when data are collected electronically (Buchanan & Scofield, 2018).

An ever-growing body of research covers the question of random responding (Bowling et al., 2016; Caldwell-Andrews, Baer, & Berry, 2000; Credé, 2010; Curran, 2016; Fronczyk, 2014; Holden et al., 2012; Holtzman & Donnellan, 2017; Huang et al., 2015; Johnson, 2005; McGonagle, Huang, & Walsh, 2016; Meade & Craig, 2012; Niessen et al., 2016; Osborne & Blanchard, 2011; Ward & Pond, 2015). However, most of these authors have investigated forms of careless responding rather than random responding, itself. The absence of a clear distinction between these notions is questionable when dealing with computer-generated response sets. First, Johnson has reminded us that careless responding includes leaving multiple blank items, which differs from random responding. Second, though the prevalence of human insufficient-effort responding can be established, the prevalence of malicious programs completing online surveys in order to earn money is unpredictable and more difficult to estimate. In our opinion, computer-generated response sets, “botnets,” and other automated form-filling programs represent new threats to be addressed in themselves, and few researchers have discussed these threats thus far (e.g., Buchanan & Scofield, 2018; Meade & Craig, 2012). These deserve specific attention due to their threat to data validity.

Several authors have addressed the detection of invalid human data, in general, but few studies have been undertaken to propose techniques for identifying entirely computer-generated data. Because bots represent the worst-case scenario of random responding patterns, wherein all items are answered completely at random, it is important to test the effectiveness of existing and new techniques to identify these invalid data before ascertaining the quality of real human data. In addition, Buchanan and Scofield (2018) showed that the best indices for detecting nonhuman data are not the indices that best detect low-quality human responses.

In this article, we aim to provide preliminary insight on the detection of computer-generated response sets in online questionnaires under controlled conditions. We compared seven indices in terms of the detection of nonhuman response sets.

Method

Data sources

Not every index has proven its validity in detecting nonhuman random response sets. Thus, the purpose of this study was to measure whether each index was able to detect such response

sets under controlled conditions. The present study was based on data from a group of 1,981 respondents who completed a personality questionnaire within a larger study (Dupuis et al., 2016). This sample was used for the present study because it is highly representative of high-quality surveys and was subject to frequent issues inherent in such studies, such as participants whose motivation waned in the course of responding, or those whose native language was not that of the questionnaires. Because the data were collected using paper-and-pencil self-administered questionnaires, none of the external indices of response validity listed so far (i.e., response time, bogus items, explicit instructed response items, and click count) were applicable.

The paper-and-pencil nature of administration here avoided the complication of collecting computer-generated invalid data. The questionnaire of interest, the NEO Five Factor Inventory (NEO-FFI; Costa & McCrae, 1992), was the sixth of the six-questionnaire battery. The sample consisted of adults who participated in a population study on cardiovascular health. Some respondents demonstrated careless responding (Dupuis et al., 2016), which proved suitable for testing whether they would have been identified as human, had this questionnaire been completed online.

Then, a second sample of simulated data was created for comparison. For theoretical and practical reasons, we first generated a sample of equivalent size. Both the risks and the effects of biased results are maximized when the proportion of valid respondents is about 50% of the total sample (Holtzman & Donnellan, 2017). A large sample was also needed in order to maximize statistical power. Thus, we used the statistical software R to generate a group of 2,000 random response sets to compare to the human sample. To test the method, the simulated data were mixed with the human sample. Second, because the 50%-proportion standard of computer-generated response sets was not likely to occur frequently, we tested the indices with variable proportions of simulated data (i.e., from 5% to 50% of the total sample), to assess the indices' effectiveness in identifying lower proportions of nonhuman data. Statistical power analyses (i.e., sensitivity analyses) were conducted prior to every other analysis, to determine which differences would lead to statistically significant results at a 5% level with a power of 99%.

For this study, we assumed that nonhuman responses were uniformly distributed (e.g., Buchanan & Scofield, 2018; Holden et al., 2012; Meade & Craig, 2012). In other words, the values were randomly distributed among the different item modalities with equivalent probabilities.

Measures

Both the human and nonhuman response sets represented data from the 60-item version of the NEO-FFI (Costa & McCrae, 1992). The NEO-FFI investigates the “Big Five” personality

factors: neuroticism, extraversion, openness, agreeableness, and conscientiousness. Each factor comprises 12 specific analytic items, including reverse-coded items. Item values on a 5-point Likert-type scale range from 0 (i.e., *it does not correspond to me at all*) to 4 (i.e., *it totally corresponds to me*). A total of 25 of the 60 items are reverse-coded, which is mainly used to detect inattention in responding. Due to index computational requirements, missing values were replaced by the neutral response option (i.e., 2).

Indices of the validity of response sets

Seven indices were calculated to assess the validity of each response set. Four were indices of internal consistency within one set of responses (i.e., response reliability, psychometric synonyms, psychometric antonyms, even–odd consistency), one was meant to measure the part of variance of a response set that was explained by the factor structure of a questionnaire (i.e., response coherence), and two were designed to detect outlying data (i.e., the Mahalanobis distance and person–total correlation).

Five well-known indices were assessed by Meade and Craig (2012) as indicators of the presence of careless responding, whereas the remaining two indices (i.e., *response coherence* and *response reliability*) result from the application of Gendreau's functional method, which was recently presented in the literature. Dupuis et al. (2015) provided detailed comparisons between classic test theory, item response theory, and Gendreau's method. Interestingly, the functional method was specifically created for multidimensional attitude questionnaires, unlike classic test theory and item response theory, which were not initially intended for use with such instruments.

The functional method relies on the creation of a hyperspheric and orthonormal measurement space that makes the classical theorems and axioms from vector geometry applicable. Such a metrical space results from repeated iterations of principal component analyses (PCAs). A first PCA is performed as usual—that is, on the answers to the items. The next step entails conducting another PCA on the resulting loading matrix. The only aim of this second PCA is to transform the loading matrix so that its dimensions are orthonormal; thus, the number of extracted components is constrained to the same number of components extracted from the first PCA. The resulting factor scores are related to the items instead of to the individual respondents. The final step is a reiteration of PCAs on the extracted factor scores. After each PCA on the extracted factor scores, each matrix line is divided by its norm; according to the generalized Pythagorean theorem, when factor scores are squared and summed by row, the norm corresponds to the square root of the sum of the squares. PCAs tend to reduce the sphericity of the measurement space

slightly, and dividing matrix lines by their norm tends to reduce the orthonormality of the components; so, the successive PCA iterations are conducted to warrant that both conditions are eventually met at the same time. Thus, the last-step PCAs were iterated 30 times, which is generally sufficient to obtain a stable matrix that meets both criteria and also corresponds to the default number of iterations of an R package currently under development. The number of components retained was constrained to five, to match with the factor structure of the test.

As a result of the procedure, a loading matrix with specific qualities is calculated: the matrix of item characteristics. In this matrix, lines are the expression of the items as radius vectors in the same hyperspheric and orthonormal space. Then, functional factor scores can be calculated by establishing the scalar products of items responses on the original Likert-type scale and each column of the matrix. Given the specificities of the measurement space, the scalar products range from -1 to 1 and can be considered equivalent to correlations: They represent the coordinates of a vector of an individual's response strategy that takes place in the aforementioned measurement hyperspheric space, wherein individuals, items, and factors can be represented together.

Response coherence Response coherence represents a multiple correlation index that indicates how predictable a responses set is. Mathematically, response coherence is equal to the square root of the sum of squared coordinates of the vector of an individual's response strategy, and varies from 0 to 1 . Psychologically speaking, response coherence indicates whether one individual has provided clear and interpretable responses to a given questionnaire.

Response reliability Response reliability is based on the application of the bisection method on the matrix of item characteristics and individual responses. First, pairs of closest items are identified within the matrix, using minimal Euclidean distance to split the data into two parallel versions of the questionnaire. Second, the scalar product of both strategy vectors calculated for both parallel versions is used as an individual split-half reliability index covering the reliability and the stability of one's response pattern. The Spearman–Brown correction formula is eventually applied, and values are bounded between -1 and 1 .

Mahalanobis distance The Mahalanobis (1960) distance is a multivariate outlier detection index used to establish the distance between one response set and the other responses. Mahalanobis distance is used to detect outliers, on the basis of the statistical significance of the difference. For this study, the distance itself was used as an estimate of the likelihood of considering respondents as outliers.

Person–total correlation The person–total correlation measures the extent to which a set of responses fits general tendencies. It relies on the very simple calculation of the correlation between one person's scores and the mean scores for different items (Curran, 2016). This index is also used to detect a specific form of careless responding, in which the respondent inverts his or her responses due to misreading the questionnaire instructions (Dupuis et al., 2015).

Psychometric antonyms and psychometric synonyms Indices of response consistency based on psychometric antonyms or synonyms are popular techniques for detecting invalid response sets (Huang, Curran, Keeney, Poposki, & DeShon, 2011; Meade & Craig, 2012). Like the calculation of response reliability, the calculation of such indices of response consistency is based on the creation of items pairs—either antonyms or synonyms, based on the largest negative or positive correlations between items, respectively. Once those pairs of items are identified, the responses are split into two scales. Then the correlation of the items on both scales is taken as the index of response consistency. These correlations are then corrected using the Spearman–Brown formula and bound between -1 and 1 . For this study, eight pairs of items with the highest or lowest correlations were used to calculate the two indices. Specifically, the correlations between the pairs of antonyms ranged from $-.27$ to $-.22$, and the correlations between the pairs of synonyms ranged from $.27$ to $.30$, depending on the simulated scenario.

Odd–even consistency Odd–even consistency is often considered the simplest measure of individual consistency. To calculate an odd–even consistency index, each scale of a test is divided into odd and even items, and scores are computed for each half. The correlation between the resulting scores across multiple scales is used as a unidimensional index of response consistency (Curran, 2016; Johnson, 2005; Meade & Craig, 2012; Ward & Meade, 2018). Consistent with the other correlation indices, the Spearman–Brown correction formula is used and the values are bound between -1 and 1 .

Longest string The longest string of identical responses was measured (Curran, 2016; Johnson, 2005; Meade & Craig, 2012). According to questionnaire length, this index could vary between 0 to 59 consecutive responses. Given that this index was not designed to detect random responding, it was only used to screen for some specific cases.

Statistical analyses

First, means and standard deviations were calculated for each variable in both the human and the nonhuman data. Cronbach's α was calculated to measure internal consistency

of the Big Five dimensions (Cronbach, 1951). In addition, the correlations between indices were also calculated.

Concerning inferential statistics, normality assumptions can be considered violated by design, but simulation studies have shown that conditions of nonnormality bias the results of mean comparisons only slightly, and the t test is powerful under both normal and uniform distributions (Poncet, Courvoisier, Combescure, & Perneger, 2016). Thus, the mean differences between groups were assessed using the Satterthwaite–Welch adjusted t test (Satterthwaite, 1946; Welch, 1947). Cohen's (1992) d was used to measure the standardized effect size. In addition, the large sample tended to prevent risks of the occurrence of Type II error and from variations of Type I error rates resulting from the nonnormality of the variables.

Receiver operator characteristic (ROC) curves were then used to measure how well the different indices predicted the presence of nonhuman responses. For each ROC curve, the area under the curve (AUC) was calculated in order to quantify the percentages of both nonhuman (i.e., true positive) and human (i.e., true negative) responses correctly detected by using each index. Differences in the predictions were compared using DeLong's test (DeLong, DeLong, & Clarke-Pearson, 1988). Finally, because several index computations require nonzero interitem standard deviations, we initially computed the longest-string index (Curran, 2016; Johnson, 2005; Meade & Craig, 2012), indicating the maximum number of consecutive identical response options. Therefore, we screened for extreme longest-string patterns that could potentially make several indices incalculable, by using a very restrictive cutoff score wherein only individuals with 20-item (one third of the questionnaire) or longer strings of consecutive same responses were excluded from the sample. This conservative approach led to the exclusion of the worst cases (Curran, 2016) among the human respondents.

Additional samples consisting of 200, 400, 800, 1,200, or 1,600 computer-generated response sets were created to correspond to fractions of 5% to 40% of the simulated data within the entire sample. ROC curves were calculated for each condition in order to measure whether each index would show the same accuracy with different fractions of simulated botnet responses. Power analyses were conducted a priori. For the sample sizes with 5% botnet responses ($n_1 = 1,981$, and $n_2 = 200$), AUC values above 58.9% were considered significant at the 5% level with a statistical power of 99%. Thus, the risk of Type II errors was minimal by design, and even lower in the scenarios with more random data included.

The analyses were two-tailed, with $\alpha = 5\%$, and were performed using the R package pROC (Robin et al., 2011). The Bonferroni correction was used to correct for significance resulting from multiple testing iterations.

Results

Main sample's characteristics

Prior to the statistical analyses, 14 respondents from the 1,981 human participants were excluded because they entered the same response to more than 20 items consecutively. Statistics for the remaining 1,967 human respondents and the 2,000 simulated random data sets are presented in Table 1. Regarding personality traits, internal consistency coefficients ranged from .66 to .84 in the human respondents; by contrast, Cronbach's α coefficients ranged from $-.05$ to $.04$ in the simulated data, whereas the mean scores for each factor were about 24, confirming the uniform distribution of the response sets and the expected absence of associations resulting from random responding. Concerning mean levels, the human and simulated responses differed significantly on each variable of interest. Every difference was significant at the 0.1% level, even after Bonferroni correction. Effect sizes above 1.42 were found for every index. In particular, standardized mean differences of 4.32 and 3.99 were measured for response coherence and Mahalanobis distance, respectively, highlighting how much the two groups differed on these indices. A smaller, but still very large, effect size of 3.21 was found concerning the person–total correlation.

Index correlations

Large correlations, ranging from $-.87$ to $.87$, were measured between different indices, underscoring the large associations between some of the indices. These correlations are presented in density plots in Fig. 1. Nonetheless, several associations are

linear—specifically, the associations between consistency indices and both response coherence and Mahalanobis distance, as illustrated in the scatterplots.

ROC curve analyses

The ROC curve analyses resulted in high proportions of correct detections of simulated response sets. In the condition with 50% simulated data, we compared the 1,967 human respondents to the 2,000 simulated data sets. In this case, AUCs ranging from 86.92% to 99.12% were measured (see Fig. 2). Specifically, an AUC of 99.12% was found for response coherence, an AUC of 96.37% for response reliability, an AUC of 98.66% for Mahalanobis distance, an AUC of 97.25% for person–total correlation, an AUC of 86.92% for psychometric antonyms, an AUC of 94.64% for psychometric synonyms, and an AUC of 89.67% for odd–even consistency. Each proportion was significant ($p < .001$) after Bonferroni correction. Furthermore, the DeLong test showed that almost every difference between AUC values was significant at the 0.1% level. The only exceptions concerned the differences between response coherence and Mahalanobis distance and between response reliability and person–total correlation. These two differences were not significant after the Bonferroni correction.

To generalize these results, the same analyses were performed by mixing the human sample with various percentages of simulated data. The results were consistent across conditions (see Fig. 2). In particular, response coherence showed an AUC of about 99.10% in every condition. High AUC values were also measured for the Mahalanobis distance, person–total correlation, and response reliability.

Table 1 Descriptive statistics and mean comparisons

Variable	Human Respondents ($n = 1,967$)			Random Response Sets ($n = 2,000$)			t	df	p	95% Confidence Interval		Cohen's d
	α	M	SD	α	M	SD				Lower	Upper	
Neuroticism	.84	18.14	7.77	$-.05$	23.93	4.81	-28.2	3,269	$<.001$	-6.19	-5.38	-0.90
Extraversion	.77	28.00	6.24	.02	23.95	4.95	22.1	3,744	$<.001$	3.61	4.31	0.70
Openness	.72	29.84	6.01	.04	24.02	5.00	33.0	3,815	$<.001$	5.46	6.15	1.05
Agreeableness	.66	33.39	5.23	.01	24.14	4.90	57.4	3,940	$<.001$	8.93	9.56	1.82
Conscientiousness	.79	34.94	5.75	$-.05$	24.14	4.78	64.2	3,816	$<.001$	10.47	11.13	2.04
Response coherence	—	.72	.11	—	.28	.09	135.8	3,695	$<.001$	.43	.45	4.32
Response reliability	—	.89	.16	—	$-.11$	.68	64.0	2,231	$<.001$	.97	1.03	2.02
Mahalanobis distance	—	33.86	15.66	—	85.88	9.83	-125.1	3,298	$<.001$	-52.84	-51.20	-3.99
Person–total correlation	—	.55	.20	—	.01	.13	100.7	3,295	$<.001$	.54	.56	3.21
Psychometric antonyms	—	.66	.39	—	$-.08$	.63	44.7	3,368	$<.001$	.71	.78	1.42
Psychometric synonyms	—	.82	.27	—	$-.12$	.62	61.9	2,734	$<.001$	.91	.97	1.96
Odd–even consistency	—	.81	.33	—	$-.09$	.73	50.2	2,780	$<.001$	.86	.93	1.59

α = Cronbach's alpha; M = mean; SD = standard deviation; t = Welch's t test; df = degrees of freedom; p = p value.

Fig. 1 Correlation and density plots for each index ($n = 3,967$)

Discussion

In this article, we aimed to compare how well different indices can detect computer-generated response sets in large datasets. This discussion reviews each index across three features: *validity* (i.e., the extent to which simulated data were identified correctly), *feasibility* (i.e., how easily indices can be calculated), and potential specificities.

Response coherence showed AUC values of 99.10%, which was the highest proportion of correct detections. The index is especially designed for multidimensional psychometric questionnaires, which potentially makes it more robust than other indices when applied to such instruments. However, this index is not easily calculable with user-friendly statistical programs. An R package is in preparation to make this index feasible to use.

Response reliability showed AUC values of 96.40%. Like response coherence, response reliability is not easily calculable with existing software. This index is likely to be more accurate in multidimensional questionnaires.

Mahalanobis distance showed AUC values of 98.70%. This index can be calculated using widely available statistical software (e.g., in SPSS, by using the linear regression function). In addition, since this index is a distance, it is applicable to nonnormal data.

Person–total correlation showed AUC values of 97.30%. This index is easily computable in every spreadsheet program with correlation functions. It is not impacted by the number of factors in a questionnaire, but is more likely to be sensitive to the number of items.

Psychometric antonyms and *psychometric synonyms* showed consistent AUC values of 86.90% and 94.80%, respectively. However, these indices are difficult to calculate,

Fig. 2 Receiver operator characteristic (ROC) curve analysis for the detection of random responding, with 5% to 50% simulated responses. AUC: area under the curve

because user-friendly programs do not exist for these computations.

Odd-even consistency showed AUC values of 89.7%. However, this index is again not easily computable.

As a result of these comparisons, response coherence, Mahalanobis distance, and person-total correlation appear to be far better predictors of botnet or other program random responding than are other indices. Response coherence resulted in the highest number of correct detections of random response sets, but the difference from Mahalanobis distance was small (about 0.45% in each condition) and not significant. The difference from the other indices was, however, meaningful.

This study focused specifically on computer-generated sets of responses that were generally based on a uniform distribution of item-level responses. According to several researchers, techniques exist to detect careless responding.

Several authors have stated that practitioners and researchers are often reluctant to use techniques to check for invalid data because they consider that the adequate indices are too complex and difficult to calculate (Borsboom, 2006; Liu, Bowling, Huang, & Kent, 2013). This study has been a straightforward attempt to compare seven indices, some of which can be easily calculated and applied to different

questionnaires, even beyond the psychology disciplines. A crucial finding is that two of the best indices to detect random responding can be calculated easily using basic statistical programs (e.g., SPSS for the Mahalanobis distance, or spreadsheet software for the person-total correlation; <https://youtu.be/kTBM0uC2d8w>; <https://youtu.be/vSnq9npL4J0>). Thus, data could and should be screened systematically for the presence of invalid response sets generated by malicious programs.

The present study represents a modest but useful step for research based on online questionnaires. However, there are some limitations. The fact that the present study focused on the 60-item Five Factor Inventory is one. Because the indices might vary in accuracy depending on questionnaire length, this question could not be addressed in this study. Moreover, univariate and multivariate indices were presented and compared, but only some of them are likely to be more accurate when applied to multiple-factor tests. Specifically, because response coherence and response reliability rely on every relation between the factors and items, we can hypothesize that these indices are more accurate than psychometric antonyms and synonyms or odd-even consistency, for instance. By contrast, response coherence and response reliability are specially designed to assess response quality in multiple-factor

questionnaires (Dupuis et al., 2015), using them on a one-factor instrument would be of little interest; therefore, using classical indices might be more relevant in those cases. Finally, the quality of responses provided by the participants might also depend on the proportion of reverse-coded items present in the questionnaire, which merits consideration and should be controlled for in further studies.

The present findings could be generalized through future research that controls for different aspects that were unaccounted for in this study (e.g., the type of random distribution in nonhuman responses, the length and structure of the questionnaire, etc.).

In this study we aimed to focus on the detection of nonhuman data in online research. The main characteristic of such data is the equal probability of each response to occur, which distorts mean and standard deviation estimates and nullifies test sphericity. Researchers have addressed the question of random responding in humans, by comparing indices in order to screen data for invalid response sets. By contrast, there has been a gap concerning the detection of nonhuman data generated by bots, which represents an additional threat to study validity. This study was an attempt to fill this specific gap. By design, the question under study was not whether the indices could detect invalid data, but rather how well each index could distinguish nonhuman from human data. This implies that the accurate indices did not consider invalid human data to be simulated. Thus, the most accurate indices could be used only to exclude data created by malicious software, either as the very first step of a wider data-cleaning procedure or as a specific data-cleaning procedure designed to exclude only nonhuman data.

We recommend further research on low-quality data, including external indices of computer-generated data, such as those proposed by Buchanan and Scofield (2018), to compare theirs to the indices tested in this study. In addition, some specific external indices could result in decisive arguments concerning the presence of nonhuman datasets in online survey data, which represents a further important gap to fill.

Buchanan and Scofield (2018) also showed that differences between computer-generated data and human invalid data exist, making some indices specifically suited to detect either the former or the latter. However, we assume that screening for nonhuman data should take place within a larger data-screening process, and not before or after a data-cleaning procedure covering human invalid data.

The present findings also provide insight regarding human invalid data, although this was not the main aim of the study. Furthermore, more complex simulations are recommended to increase scientific knowledge on partially random forms of careless responding that are more similar to low-quality human response patterns.

Author note The present study was supported in its entirety by the University of Lausanne. However, the data used to represent the human respondents were collected within the “CoLausPsyCoLaus” cohort study and previously were used in other studies supported by research grants from GlaxoSmithKline, the Faculty of Biology and Medicine at the University of Lausanne, and the Swiss National Science Foundation (Grants 3200B0-105993, 3200B0-118308, 33CSO-122661, 33CS30-139468, and 33CS30-148401). The authors have no conflict of interest to disclose. The authors express their sincere appreciation to Sarah Stauffer Brown, for copyediting the manuscript, and Jessica Gale, for dubbing the tutorial videos.

References

- Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, 71, 425–440. <https://doi.org/10.1007/s11336-006-1447-6>
- Bowling, N. A., Huang, J. L., Bragg, C. B., Khazon, S., Liu, M., & Blackmore, C. E. (2016). Who cares and who is careless? Insufficient effort responding as a reflection of respondent personality. *Journal of Personality and Social Psychology*, 111, 218–229. <https://doi.org/10.1037/pspp0000085>
- Briones, E. M., & Benham, G. (2017). An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples. *Behavior Research Methods*, 49, 320–334. <https://doi.org/10.3758/s13428-016-0710-8>
- Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*. Advance online publication. <https://doi.org/10.3758/s13428-018-1035-6>
- Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon’s Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6, 3–5. <https://doi.org/10.1177/1745691610393980>
- Caldwell-Andrews, A., Baer, R. A., & Berry, D. T. R. (2000). Effects of response sets on NEO-PI-R scores and their relations to external criteria. *Journal of Personality Assessment*, 74, 472–488. https://doi.org/10.1207/S15327752jpa7403_10
- 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. *Computers in Human Behavior*, 29, 2156–2160. <https://doi.org/10.1016/j.chb.2013.05.009>
- Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaivete among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, 46, 112–130. <https://doi.org/10.3758/s13428-013-0365-7>
- Chandler, J., & Paolacci, G. (2017). Lie for a Dime. *Social Psychological and Personality Science*, 8, 500–508. <https://doi.org/10.1177/1948550617698203>
- Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, 1, 120–131. <https://doi.org/10.1017/xps.2014.5>
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
- Costa, P. T., & McCrae, R. R. (1992). Revised Neo Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI). Lutz, FL: Psychological Assessment Resources.

- Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, 70, 596–612. <https://doi.org/10.1177/0013164410366686>
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Curran, P. G. (2016). Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology*, 66, 4–19. <https://doi.org/10.1016/j.jesp.2015.07.006>
- DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44, 837–845. <https://doi.org/10.2307/2531595>
- DeSimone, J. A., DeSimone, A. J., Harms, P. D., & Wood, D. (2018). The differential impacts of two forms of insufficient effort responding. *Applied Psychology*, 67, 309–338. <https://doi.org/10.1111/apps.12117>
- Dupuis, M., Capel, R., Meier, E., Rudaz, D., Strippoli, M.-P. F., Castela, E., . . . Vandeleur, C. L. (2016). Do bipolar subjects' responses to personality questionnaires lack reliability? Evidence from the PsyCoLau study. *Psychiatry Research*, 238, 299–303. <https://doi.org/10.1016/j.psychres.2016.02.050>
- Dupuis, M., Meier, E., Capel, R., & Gendre, F. (2015). Measuring individuals' response quality in self-administered psychological tests: An introduction to Gendre's functional method. *Frontiers in Psychology*, 6, 629:1–12. <https://doi.org/10.3389/fpsyg.2015.00629>
- Fronczyk, K. (2014). The identification of random or careless responding in questionnaires: The example of the NEO-FFI. *Rzeczni Psychologiczne*, 17, 457–473.
- Gleibs, I. H. (2017). Are all “research fields” equal? Rethinking practice for the use of data from crowdsourcing market places. *Behavior Research Methods*, 49, 1333–1342. <https://doi.org/10.3758/s13428-016-0789-y>
- Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26, 213–224. <https://doi.org/10.1002/bdm.1753>
- Holden, R. R., Wheeler, S., & Marjanovic, Z. (2012). When does random responding distort self-report personality assessment? An example with the NEO PI-R. *Personality and Individual Differences*, 52, 15–20. <https://doi.org/10.1016/j.paid.2011.08.021>
- Holtzman, N. S., & Donnellan, M. B. (2017). A simulator of the degree to which random responding leads to biases in the correlations between two individual differences. *Personality and Individual Differences*, 114, 187–192. <https://doi.org/10.1016/j.paid.2017.04.013>
- Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2011). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, 27, 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. *Journal of Applied Psychology*, 100, 828–845. <https://doi.org/10.1037/a0038510>
- Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality*, 39, 103–129. <https://doi.org/10.1016/j.jrp.2004.09.009>
- Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 49, 433–442. <https://doi.org/10.3758/s13428-016-0727-z>
- Litman, L., Robinson, J., & Rosenzweig, C. (2015). The relationship between motivation, monetary compensation, and data quality among US- and India-based workers on Mechanical Turk. *Behavior Research Methods*, 47, 519–528. <https://doi.org/10.3758/s13428-014-0483-x>
- Liu, M., Bowling, N., Huang, J., & Kent, T. (2013). Insufficient effort responding to surveys as a threat to validity: The perceptions and practices of SIOP members. *Industrial–Organizational Psychologist*, 51, 32–38.
- Liu, M., & Wronski, L. (2018). Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research*, 60, 32–49. <https://doi.org/10.1177/1470785317744856>
- Mahalanobis, P. C. (1960). A method of fractile graphical analysis. *Econometrica*, 28, 325–351. <https://doi.org/10.2307/1907724>
- Marjanovic, Z., Struthers, C. W., Cribbie, R., & Greenglass, E. R. (2014). The Conscientious Responders Scale: A new tool for discriminating between conscientious and random responders. *SAGE Open*, 4, 215824401454596. <https://doi.org/10.1177/2158244014545964>
- McGonagle, A. K., Huang, J. L., & Walsh, B. M. (2016). Insufficient effort survey responding: An under-appreciated problem in work and organisational health psychology research. *Applied Psychology*, 65, 287–321. <https://doi.org/10.1111/apps.12058>
- McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, 136. <https://doi.org/10.1037/a0019216>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17, 437–455. <https://doi.org/10.1037/a0028085>
- Necka, E. A., Cacioppo, S., Norman, G. J., & Cacioppo, J. T. (2016). Measuring the prevalence of problematic respondent behaviors among MTurk, campus, and community participants. *PLoS ONE*, 11, e0157732. <https://doi.org/10.1371/journal.pone.0157732>
- Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, 63, 1–11. <https://doi.org/10.1016/j.jrp.2016.04.010>
- Osborne, J. W., & Blanchard, M. R. (2011). Random responding from participants is a threat to the validity of social science research results. *Frontiers in Psychology*, 2, 220. <https://doi.org/10.3389/fpsyg.2010.00220>
- Poncet, A., Courvoisier, D. S., Combescure, C., & Perneger, T. V. (2016). Normality and sample size do not matter for the selection of an appropriate statistical test for two-group comparisons. *Methodology*, 12, 61–71. <https://doi.org/10.1027/1614-2241/a000110>
- Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12, 77. <https://doi.org/10.1186/1471-2105-12-77>
- Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2, 110–114. <https://doi.org/10.2307/3002019>
- Sharpe Wessling, K., Huber, J., & Netzer, O. (2017). MTurk character misrepresentation: Assessment and solutions. *Journal of Consumer Research*, 44, 211–230. <https://doi.org/10.1093/jcr/ucx053>
- Ward, M. K., & Meade, A. W. (2018). Applying social psychology to prevent careless responding during online surveys. *Applied Psychology*, 67, 231–263. <https://doi.org/10.1111/apps.12118>
- Ward, M. K., & Pond, S. B. (2015). Using virtual presence and survey instructions to minimize careless responding on Internet-based surveys. *Computers in Human Behavior*, 48, 554–568. <https://doi.org/10.1016/j.chb.2015.01.070>
- Welch, B. L. (1947). The generalization of students problem when several different population variances are involved. *Biometrika*, 34, 28–35. <https://doi.org/10.2307/2332510>