Article

Detecting Careless Responding in Survey Data Using Stochastic Gradient Boosting

Ulrich Schroeders[1] , Christoph Schmidt[2], and Timo Gnambs[3]

Abstract

Careless responding is a bias in survey responses that disregards the actual item content, constituting a threat to the factor structure, reliability, and validity of psychological measurements. Different approaches have been proposed to detect aberrant responses such as probing questions that directly assess test-taking behavior (e.g., bogus items), auxiliary or paradata (e.g., response times), or data-driven statistical techniques (e.g., Mahalanobis distance). In the present study, gradient boosted trees, a state-of-the-art machine learning technique, are introduced to identify careless respondents. The performance of the approach was compared with established techniques previously described in the literature (e.g., statistical outlier methods, consistency analyses, and response pattern functions) using simulated data and empirical data from a web-based study, in which diligent versus careless response behavior was experimentally induced. In the simulation study, gradient boosting machines outperformed traditional detection mechanisms in flagging aberrant responses. However, this advantage did not transfer to the empirical study. In terms of precision, the results of both traditional and the novel detection mechanisms were unsatisfactory, although the latter incorporated response times as additional information. The comparison between the results of the simulation and the online study showed that responses in real-world settings seem to be much more erratic than can be expected from the simulation studies. We critically discuss the generalizability of currently available detection methods and provide an outlook on future research on the detection of aberrant response patterns in survey research.

[1]University of Kassel, Kassel, Germany
[2]eoda GmbH, Kassel, Germany
[3]Leibniz Institute for Educational Trajectories, Bamberg, Germany

Corresponding Author:
Ulrich Schroeders, Psychological Assessment, Institute of Psychology, University of Kassel, Hollaendische Str. 36-38, Kassel 34127, Germany.
Email: schroeders@psychologie.uni-kassel.de

Deliberate misreporting is a serious threat to the reliability and validity of social science research. Increasing evidence suggests that some participants engage in dishonest behavior and cheat in unproctored achievement tests (Steger, Schroeders, & Gnambs, 2020), distort their responses in line with prevalent social norms (Gnambs & Kaspar, 2015), or simply do not invest the necessary effort to properly process a survey question and indicate an elaborated response (Huang et al., 2015; Litman & Robinson, 2020). Rather, a substantial proportion of survey respondents adopt cognitive shortcuts, for example, by using specific response patterns such as selecting the same response category for multiple items instead of evaluating the actual item content (Johnson, 2005). Particularly, in web-based studies, careless responses have become a major hindrance to data quality (Bowling & Huang, 2018; Weiner & Dalessio, 2006), which can substantially affect study results (Arias et al., 2020; Huang et al., 2015). Accordingly, various data screening methods have been proposed to identify careless respondents (Meade & Craig, 2012; Niessen et al., 2016), such as probing items that directly assess test-taking behavior (e.g., bogus items), auxiliary or paradata (e.g., response times), or data-driven techniques (e.g., Mahalanobis distance). In contrast, prediction methods that are commonly used in the machine learning literature, but (as of yet) are rarely used in sociological and psychological research, have been somewhat neglected. Thus, we introduce stochastic gradient boosting (Friedman, 2001) as a means to identify careless respondents in self-report surveys. The novel approach is compared with established techniques using simulated data that exhibit a clear allocation of cases to regular and careless respondents. In addition, empirical data from a web-based experiment in which participants were instructed to display different types of test-taking behavior (regular, inattentive) probe the usefulness of the machine learning algorithm as compared with traditional techniques for the detection of careless respondents.

Types and Sources of Careless Responding

The phenomenon of *careless responding* has been explained with different nuances and labels: ''content nonresponsivity'' (Nichols et al., 1989, p. 239), ''careless inattentiveness'' (Johnson, 2005, p. 104), ''insufficient effort responding'' (Huang et al., 2012, p. 100), or ''careless/insufficient effort'' (Arias et al., 2020). Most commonly, it is considered a bias in survey responses disregarding the actual item content. Sometimes, this has been described as random responding (e.g., Berry et al., 1992). However, this term is misleading because these responses can also follow some nonrandom pattern (e.g., a recurring sequence of 1-2-3-4-5) and, more important, humans

are incapable of producing true random sequences (Figurska et al., 2008). But the distinction between careless and random responses may point to the difference between intentional and nonintentional response behavior. As Niessen et al. (2016) reflect on a reviewer's comment, careless responding is perceived as ''the more benign form of aberrant responding . . . , whereas random responding is more blatant and intentional'' (p. 1). Accordingly, in our understanding, careless responding covers both intentionally and unintentionally answering items (Meade et al., 2017, p. 417) without the intention of creating a specific image of oneself. Thus, careless responding does not include, for example, malingering—the fabrication, feigning, or exaggeration of physical or psychological symptoms to achieve some beneficial outcome (e.g., early retirement, mitigating punishment, claim for compensation)—which is a common concern in clinical and forensic psychology (Mittenberg et al., 2002). This distinction between careless responding and faking or cheating may be subtle, but it is important because the latter hinges on participants' ability to fake (Geiger et al., 2018) and is presumably much harder to detect (e.g., Karabatsos, 2003).

Prevalence estimates on the extent of careless responding vary considerably, from modest 3.5% (Johnson, 2005), over 3% to 9% (Maniaci & Rogge, 2014), 10.6% (Kurtz & Parrish, 2001) and 10% to 12% (Meade & Craig, 2012) up to 35% to 46% (Oppenheimer et al., 2009). The large heterogeneity in estimates is presumably due to the different operationalizations of *true* careless responding. Moreover, it is reasonable to assume that studies differ in their extent of careless responding depending on the study's intention, the recruitment procedure, and incentives given. Regarding the consequences, careless responding introduces systematic bias, whereas random responses add noise to the measurement. Ultimately, such error variances reduce reliability estimates, measurement precision, and also attenuate or inflate correlations and, therefore, pose a threat to validity (Hong et al., 2020; Huang et al., 2015; McGrath et al., 2010). In simulation studies, Credé (2010) convincingly showed that base rates of careless responding as low as 5% can affect correlations in a similar manner as, for example, range restriction or score unreliability. Accordingly, variance–covariance-based methods are also affected by careless responding. For instance, in case of scales with partially reversely coded items, careless responding might contribute to the emergence of an additional method factor capturing the variance of some negatively worded items of the same scale. Already 10% to 20% of careless respondents in the sample are sufficient to render a single-factor confirmatory model incorrect (Woods, 2006; see also Huang et al., 2012). In the same vein, Gnambs and Schroeders (2020) compared different measurement models for the factor structure of the *Rosenberg Self-Esteem Scale* across several continuous moderators and concluded that the scale becomes more unidimensional with increasing reading competence. Thus, superficial reading or comprehension problems are an obvious source of careless responding, but there are several, potential reasons (e.g., test-taking motivation).

Traditional Techniques for the Detection of Careless Respondents

Given the risk of biased results, a number of different methods have been proposed to detect careless respondents. A rough distinction can be made between (a) special items or scales assessing participants' care, (b) the evaluation of auxiliary data from computerized testing, and (c) statistical, data-driven methods. The first group includes *bogus items* with an obvious correct answer (e.g., ''I am paid biweekly by leprechauns''), or *instructed response items* to which participants have to react correctly (e.g., ''Please click strongly agree''; Gummer et al., 2018, or the so-called blue dot task; Oppenheimer et al., 2009), or self-report measures often placed at the end of a survey that directly ask for participants' diligence and engagement (e.g., ''I put forth my best effort in responding to this survey''; Meade & Craig, 2012). However, the usefulness of such items is still debated (Curran & Hauser, 2019), because their inclusion can result in negative spillover effects by irritating participants or introducing reactance.

The second group of indices relies on so-called paradata to gain further insights into participants' test-taking behavior (Couper, 2005). Paradata are additionally recorded in computerized testing and include, among others, log data (Kroehne & Goldhammer, 2018), response latencies (Leiner, 2019), keystrokes, and mouse clicks (Kieslich & Henninger, 2017; Olson & Parkhurst, 2013). Major advantages of paradata are that they can unobtrusively be assessed and are hard to fake. For example, click counts and page timing have been used to flag bot-generated answers in a web-based survey (Buchanan & Scofield, 2018). In the same vein, Steger, Schroeders, and Wilhelm (2020) demonstrated in an experimental setting, comparing a proctored with an unproctored knowledge assessment, the superiority of paradata such as response times and focus shift of browser tabs in comparison to classical self-report scales (e.g., honesty scales) in detecting cheaters. And recently, both responses and auxiliary response time information was used in a mixture hierarchical modeling approach to detect aberrant responses (Wang et al., 2018; Wang & Xu, 2015).

Finally, several data-driven techniques, such as statistical outlier methods, consistency analysis, or response pattern functions (see Table 1), are conducted after the assessment to identify careless respondents. The detection and removal of *statistical outliers* (i.e., data points that differs significantly from other observations) is one of the most common methods of data cleaning. Multivariate approaches such as the Mahalanobis distance evaluate the entire response pattern in a series of items (e.g., a scale) to identify respondents with aberrant response behavior. Previous research indicated high sensitivity and specificity of outlier detection for complete random responses, but not for the midpoint response style or more sophisticated cheating strategies (Maniaci & Rogge, 2014; Meade et al., 2017). *Consistency analysis* comprises a heterogeneous set of techniques (Curran, 2016; Maniaci & Rogge, 2014; Meade & Craig, 2012), all relying on the basic notion that careless respondents produce responses that are internally inconsistent. For example, the odd–even consistency is calculated by breaking each individual's responses into even–odd items sets for each unidimensional subscales. On the item level, the within-person correlations

Table I. Overview of Data-Driven Mechanisms to Detect Careless Respondents.

Index (Abbr.)	Description	Strengths	Weaknesses	Key references
Statistical outlier functions				
Mahalanobis distance (Maha.)	Multivariate distance between a respondent's response vector and the vector of sample means	• Easy to calculate and understand • Widespread usage	• Effective only for truly random responses	Mahalanobis (1936)*, Meade & Craig (2012)
Consistency analysis				
Synonym/antonym score (Ant.)	Within-person correlation between highly correlated item pairs (e.g., $|r| > .60$)	• Good, sensitive detection of careless respondents	• For semantic synonym/ antonym score similar/ contrasting worded items • Varying and arbitrary cutoff for correlation	Meade & Craig (2012), Maniaci & Rogge (2014), Curran (2016)
Even–odd consistency (EvenOdd)	Within-person correlation across unidimensional subscales formed by even–odd split halves	• High scale dependency	• Relies on unidimensional scales • Sufficient amount of items	Meade & Craig (2012), Curran (2016)
Intraindividual response variability (IRV)	Intraindividual standard deviation across a set of consecutive item responses	• Very easy to calculate and understand	• Should be calculated across multiple constructs and reversely coded items • Both low and high variability might indicate careless responding	Marjanovic et al. (2015)*, Dunn et al. (2018)
Response pattern functions				
Longstring (Long.)	Maximum (or average) of consecutive items answered with the same response option	• Easy to calculate and understand	• Uniform response scale • Needs larger sets of items covering different constructs • Arbitrary cutoff	Johnson (2005)*
Number of Guttman errors	Number of item pairs that behave contrary to expectations regarding the solution probabilities	• Nonparametric version	• Relies on large sample sizes • Rather complicated to calculate	Meijer et al. (1994)*, Niessen et al. (2016)

33

Table 1. (continued)

Index (Abbr.)	Description	Strengths	Weaknesses	Key references
Polytomous U_3 person-fit statistics	Extent to which a person's nonparametric polytomous IRT estimate matches the probability of correctly solving items	• Nonparametric version	• Relies on large sample sizes • Rather complicated to calculate • Presupposes unidimensional scales	Emons (2008)*
Z_h statistics (Z_h)	Extent to which a person's polytomous IRT estimate corresponds to the probability of correctly solving items	• Uses information on the structure of the measure	• Relies on large sample sizes • Rather complicated to calculate	Dragsow et al. (1985)*

Note. References marked with an asterisk (*) are key references. The intraindividual response variability is also known as inter-item standard deviation and can also be classified as a response pattern function. The abbreviation in parentheses is used in the Results section. IRT = item response theory.

34

between highly correlated item pairs (irrespective of the content) are called psychometric synonym/antonym scores. Although such consistency indices yield good, sensitive detection of careless respondents (Curran, 2016), they are accompanied by a variety of difficulties, for example, the necessity of relatively homogeneous, redundant scales with similarly (or oppositely) formulated items. Last, *response pattern functions* are used to detect uncommon patterns in parts or the complete response vector of an individual in comparison to others.[1] Some of them are easy to calculate such as the longstring index (i.e., the number of consecutive items answered with the same response alternative; Johnson, 2005) or the inter-item standard deviation, which is also known as intraindividual response variability (IRV; i.e., standard deviation across a response vector; Marjanovic et al., 2015). They are useful if the items can be assigned to different constructs and some of them are negatively worded. More complex procedures include the number of Guttman errors (Curran, 2016) or person-fit statistics (Meijer, 1996), which have been devised for achievement data with a dichotomous answer format, whereas there are only few procedures for polytomous items (Emons, 2008; Tendeiro et al., 2016), especially in a multidimensional context (Drasgow et al., 1985; Glas & Dagohoy, 2007). More specifically, person-fit statistics have been used to identify participants with spuriously low or high test scores by comparing participants' actual responses with the expected responses (Karabatsos, 2003) and have been reported to detect both deliberate cheating (e.g., answer copying; Sotaridona & Meijer, 2002) and random responding (Niessen et al., 2016).

Stochastic Gradient Boosted Trees

Although machine learning methods have existed in computer science for decades, they have only recently found wider application in psychology for the analysis of large and complex data structures (Fan et al., 2014), including the prediction of personality from online social networking behavior (Kosinski et al., 2013) or the prediction of physiological health (e.g., diabetes, hypertension) through personality nuances (i.e., items; Seeboth & Mõttus, 2018). In psychological assessment, Zopluoglu (2019) used an extreme gradient boosting algorithm, which is an efficient implementation of gradient boosting (Friedman, 2002), to identify test-takers with preknowledge more precisely compared with more traditional item exposure measures (area under the receiver operating characteristic curve [AUROC] = .930). Thus, the algorithm might help in classification problems such as separating careless from diligent respondents in case the class labels are known. Besides the good predictive performance in similar contexts (Zopluoglu, 2019), also the properties of the algorithm favor this method over others (see Friedman, 2002).

In general, decision trees are a versatile and increasingly popular method for the analysis of complex data structures with many variables and nonlinear relationships that do not require the a priori specification of a functional form between an outcome and its predictors (James et al., 2017). They sequentially partition the sample into mutually exclusive classes and, thus, build a tree-like structure such as in Figure 1.
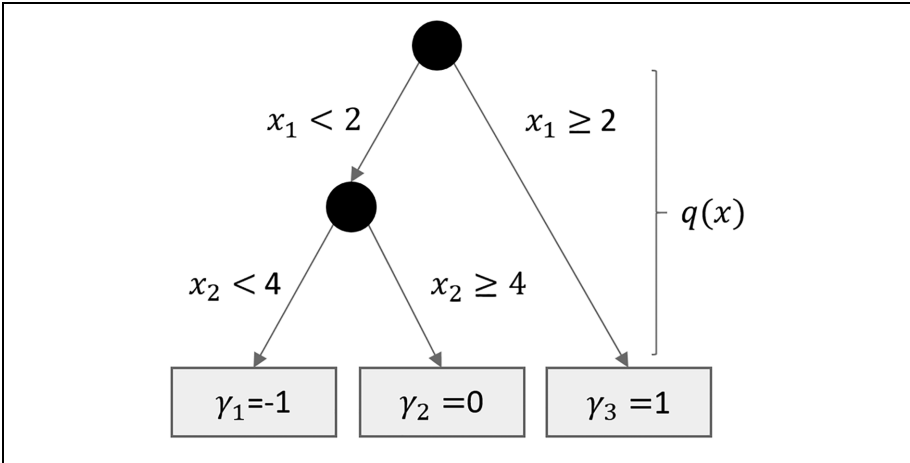
Figure 1. Example of a decision tree with two nodes and three leaves.

Let y_i denote the outcome for respondent $i \in \{1, ..., I\}$ and $x_i = \{x_{i1}, x_{i2}, \ldots, x_{iq}\}$ a Q-dimensional vector of predictors. Then, a set of decision rules and associated parameter values is identified that splits the data at each branch of the tree into two smaller subsets according to a variable in x, thus, resulting in a tree that consists of J end points (i.e., leaves). Given a tree structure T with its decision rules $q(x)$ that maps each observation i to one of the k leaves and a vector of scores on each leaf $\gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_k\}$ the regression relationship between y_i and x_i can be formally expressed as

$$\hat{y}_i = f(x_i; T), \tag{1}$$

with f as a function describing the tree. The unknown tree parameters T are obtained by minimizing a differentiable loss function L that measures the goodness of the prediction for each data point:

$$\hat{T}_m = \arg\min \sum_{i=1}^{N} L(y_i, \hat{y}_i). \tag{2}$$

As the present study uses a binary outcome that represents the regular versus careless respondents group, a logistic loss function L is used

$$L(y_i, \hat{y}_i) = y_i \ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \ln\left(1 + e^{\hat{y}_i}\right). \tag{3}$$

Single trees usually have a rather poor predictive performance, but can be combined as ensembles for higher accuracies (Breiman, 2001). The gradient boosting machine (GBM; Friedman, 2001) sequentially estimates M decision trees by choosing

the optimal split points for x_i. At the same time, the maximum number of leaves and splits is limited to guarantee that each tree acts as a weak learner and does not dominate the prediction. The M decision trees are added one at a time using a gradient descent procedure. The loss function L is approximated by calculating its gradient with respect to the predicted value of the ensemble and adding a new tree to the ensemble, if the tree moves the loss in the direction of the prediction values. Following this logic, the prediction $\hat{y}$ at iteration m can be expressed as the prediction at the previous iteration $\hat{y}_i^{(m-1)}$ and the prediction by the current tree $f_m(x_i; T_m)$, thus rewriting the loss function at iteration m as

$$\hat{T}_m = \arg\min \sum_{i=1}^{N} L\left(y_i, \hat{y}_i^{(m-1)} + f_m(x_i; T_m)\right). \tag{4}$$

The numerical optimization algorithm suggested by Friedman (2001) reformulates Equation (4) using the negative gradient g_{im} for the ith observation at iteration m and proceeds in three steps. First, g_{im} is estimated for each observation i as

$$g_{im} = -\left[\frac{\partial L(y_i; f_{(m-1)}\left(x_i; T_{(m-1)}\right))}{\partial f\left(x_i; T_{(m-1)}\right)}\right]. \tag{5}$$

These pseudo-residuals reflect the potential reduction in loss and, in the second step, are used as outcomes for estimating the parameters for the decision tree at iteration m using Equation (5). Finally, the predictions are updated to derive the predicted values at iteration m as

$$\hat{y}_i^m = \hat{y}_i^{(m-1)} + \lambda \cdot f_m\left(x_i; \hat{T}_m\right), \tag{6}$$

with λ as the gradient descent step size (i.e., learning or shrinkage parameter) that controls the rate at which the algorithm updates the predictions. Friedman (2002) showed that incorporating a boosting approach in the algorithm substantially improves the predictive performance. Thus, stochastic gradient boosted trees do not use the entire sample in Step 2 but are limited to a random subsample (without replacements) of observations.

In stochastic gradient boosting, various tuning parameters need to be specified that affect the performance of the algorithm. Because there is no convergence criterion, the number of iterations M determines the amount of residual error. The maximum interaction depth determines the number of splits allowed in the tree, whereas the leaf size indicates the minimum number of observations in a given leaf. Larger interaction depths and smaller leaves sizes allow for more complex models, but also lead to less stability and higher computational demands (Berk, 2017). In addition, parameters regarding the minimum impurity decreasement for a split to be considered worthy, the learning rate λ, and the number of trees are commonly optimized in grid searches. Small values for the learning rate λ not only slow down the learning, requiring more iterations, but also improve the stability of the predictions.

The Present Study

The main objective of the study is to evaluate different statistical techniques to identify participants who likely engaged in careless responding. Therefore, we implemented different post hoc statistics that have previously been described in the literature and that serve as a benchmark (e.g., statistical outlier methods, consistency analyses, and response pattern functions, see also Table 1). To identify careless respondents, we also used a GBM (Friedman, 2002). In comparison to other machine learning algorithms, the GBM has several advantages: First, it is a black box approach that does not require the a priori specification of a functional relationship between the predictors and group membership (regular vs. careless respondents), which is advantageous because the relationship is basically unknown. Such a model evaluation favors prediction over explanation (Shmueli, 2010). Second, the GBM can capture complex relationships (James et al., 2017) between the predictors (i.e., interactions) and between predictors and outcome (i.e., nonlinear relations). Third, the GBM is eligible to search for patterns in complex data structures with many heterogeneous variables, including paradata such as response times. Paradata is considered more difficult to counterfeit because the recording is done incidentally and clear assumptions of the test-takers about the relation between paradata and careless responding are missing. The evaluation of binary classifiers (e.g., specificity, sensitivity) is done for traditional detection mechanisms and gradient boosted trees with simulated data and with empirical data of an experimental study.

Method

Study 1: Simulation Study

Data Generation. In comparison to real data where the prevalence of careless responding is usually unknown (if not experimentally induced or otherwise controlled for), Monte Carlo simulations offer the advantage that different response patterns can be modeled reliably, so that the classification quality of the detection mechanisms can be evaluated without measurement uncertainty. Although they allow nuanced conclusions about the specific procedures, they can be criticized for their idealized distributions of both regular and irregular response patterns. In general, Monte Carlo simulations require a large number of specifications (see also previous simulations by Credé, 2010; Meade & Craig, 2012; Niessen et al., 2016); we describe the most important ones below and refer for detailed specifications to an Open Science Framework (OSF) repository (Soderberg, 2018) in which we provide all data and syntax files to foster transparency and reproducibility: https://osf.io/mct37.

We simulated nine data conditions in a 3×3 design manipulating the amount of careless respondents in the overall sample (5%, 10%, 15%) and three types of response styles (random, midpoint, and fixed pattern respondents; see also Meade & Craig, 2012). We aligned the data generation of the simulation study as closely as possible to the empirical data (which will be described as Study 2) to enable a proper comparison between both studies. More precisely, we simulated valid responses to

60 ordinal variables with five response categories based on the empirical correlation matrix and the response distribution from Study 2 (see OSF repository for the syntax) using the R package genOrd (Version 1.4.0; Ferrari & Barbiero, 2012). With regard to the sample size, we also closely mirrored the conditions of the empirical study allowing for a direct comparison of the detection rates with simulated and empirical data (for more information see Appendix A in the OSF repository).

In line with previous simulation studies,[2] we differentiate three types of careless respondents: (a) random respondents, (b) midpoint respondents, and (c) fixed pattern respondents. In more detail, the answers of the *random respondents* were randomly drawn from a uniform distribution across all response categories (i.e., 20%, 20%, 20%, 20%, 20%). Accordingly, the responses can be considered as completely random, which is not necessarily a realistic condition. The *midpoint response styles* mirrors the response behavior of ''persons that may stick primarily to the scale midpoints over concerns of being identified via outlier analysis'' (Meade & Craig, 2012, p. 448). In the simulation, the midpoint response style camouflaging inattentive responses is characterized by successively higher probabilities for the middle categories (i.e., 5%, 20%, 50%, 20%, 5% for the response categories). For simulating responses of *fixed pattern respondents*, we generated a repeating sequence of random numbers for each respondent (e.g., 1-2-3-1-2-3 . . . if the random numbers were 1 and 3).[3] Modeling fixed response patterns with varying responses instead of repeating the same response across several consecutive items seems a more realistic setting than constantly choosing the same response option and it should disfavor detection via longstring or IRV. Sample sizes of simulated valid and careless respondents match those of the empirical study.

Study 2: Empirical Study

Design and Participants. The web-based survey was administered through the survey tool *SoSci-Survey* (Leiner, 2020) between February 2020 and April 2020. Participants were recruited through social media, mailing lists, and a German online panel (Göritz et al., 2019). Participants were randomly assigned to one of two conditions in a between-subject design: regular responding or inattentive responding.[4] The only difference between the two conditions was the instruction (see Appendix B in the OSF repository). The first instruction for regular response behavior was a standard instruction as given in many measures of typical behavior emphasizing that participants should select the answer that fits best after carefully considering each statement and taking the time to think about the answer. The second instruction tried to emulate the mind-set and response behavior of inattentive respondents (see also Niessen et al., 2016) by focusing on a speedy and superficial processing. Thus, the respondents were asked to quickly finish the survey without carefully reading the items. The instruction also emphasized the fact that participation in the lottery would not be affected by the actual responses. In addition to a data protection agreement, participants had to answer a control question (''What were your instructions?'') after reading the instruction and prior to the actual questionnaire to ensure that the instruction was read and

understood. After the questionnaire, participants were asked whether they followed the instruction (yes/no) and had to answer the control question again.

Participants were included in the analyses if they had completed the questionnaire and indicated that they had followed and correctly remembered the instruction (for a flowchart with the dropout on the different stages see Figure S1 in the online repository). In total, 605 respondents completed the test, either under the regular condition ($n = 361$) or the careless responding condition ($n = 244$). Two thirds of the participants were female (66.9%, 32.4% male, and 0.7% diverse); the average age was 43.1 years ($SD = 17.8$). The sample of the regular and the careless condition was made up as follows: 28.1% pupils/students, 2.8% manual workers, 39.3% employees, 5.0% self-employed, 16.5% retired, and 8.3% others (information was available for $n = 601$). Almost all participants ($>98\%$) stated that they understood German *well* or *very well* on a 4-point scale.

Selection of Analysis Sample. The empirical sample consisted of 605 respondents with 361 in the regular condition and 244 in the careless condition. For the statistical analyses, we split this sample into a test sample (30% or 180 respondents) and a training sample (70% or 425 respondents) with the following boundary condition: Based on the literature, we constrained the ratio of regular respondents to careless respondents in the test sample to 9:1 (or 162:18 in absolute numbers). The remaining respondents constituted the training sample with an almost balanced ratio (i.e., 199 regular and 226 careless respondents). A balanced ratio in the learning phase (i.e., training) and a realistic ration in the evaluation phase (i.e., test) provide optimal conditions, both from a learning and an evaluation perspective. The constrained random sampling procedure was repeated 1,000 times to quantify the variability of the results. The analysis sample derived from the empirical study corresponded to one of the conditions of the simulation study because both have a testing sample of 180 respondents, including 10% careless respondents (for more information see Appendix A in the OSF repository).

Measures. In addition to basic demographic data (i.e., age, gender, profession, command of the German language), participants worked on the German version of the HEXACO-60 (Ashton & Lee, 2009) measuring six personality traits: Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). Participants had to indicate their agreement to the different statements (e.g., ''I would be quite bored by a visit to an art gallery'') on 5-point scales (1 = *strongly disagree*, 2 = *disagree*, 3 = *neutral*, 4 = *agree*, 5 = *strongly agree*). Skipping items was not possible to prevent test-takers from rushing through the questionnaire. We collected response times on the item level, which were winsorized for extremely long response times (i.e., replacing outliers with the 95th percentile) and aggregated to parcels of 10 (in accordance with the page layout) for subsequent analyses.

Statistical Analyses for Studies 1 and 2

All traditional data-driven detection mechanisms (see also Table 1) were calculated with the R package careless (Version 1.1.3; Yentes & Wilhelm, 2018), except for the Z_h statistic that was calculated using mirt (Version 1.32.1; Chalmers, 2012). In more detail, the Mahalanobis distance was calculated (with a confidence level of 95%) on data sets in which reversely formulated items were recoded. Psychometric antonyms depict the within-person correlation of all negatively correlated item pairs, irrespective of the scale they belong to. A critical correlation between two items of $r = -.20$ was considered sufficient given the typical low mean item-intercorrelation of the HEXACO items of .05 $[-.45; .63]$. In contrast to semantic antonyms that take the item content into account, psychometric antonyms are purely data-driven (see also Curran, 2016). Even–odd consistency was computed as the within-person correlation between the even and the odd items of a subscale; that is, information on the six-dimensional structure of the HEXACO questionnaire was taken into account in the calculation of the coefficient. At least six identical responses in a row (before reverse coding negatively phrased items) were considered conspicuous for the longstring index (see also Niessen et al., 2016). IRV can manifest on both sides of the distribution, either as low IRV scores reflecting straight lined responses (Dunn et al., 2018) or as high IRV scores indicating highly random responses (Marjanovic et al., 2015). Accordingly, IRV scores were calculated twice and the higher accuracy rates were taken, which constitutes an overestimation, because in realistic settings such informed adjustments cannot be made. The Z_h person-fit statistic was based on a polytomous multidimensional item response theory (IRT) model with the HEXACO traits as dimensions (Drasgow et al., 1985).

We used the GBM algorithm of the R package gbm (Version 2.1.5; Greenwell et al., 2019) as supervised machine learning algorithm and caret (Kuhn, 2008, 2020) as interface for parallel modeling and prediction. To assess the correctness of the model classification, we split the sample into a holdout (also called test) sample (30%) to test model performance in an independent test sample and a training-validation sample (70%) to train and tune the model. To minimize overfitting (Yarkoni & Westfall, 2017), we used 10-fold cross-validation. The optimal settings for the tuning parameter for each of 1,000 random training-validation samples were identified via grid search with the following parameter settings: interaction depth of 2, 3, or 4, the minimum leaf size between 4 and 10, the shrinkage as a sequence between .001 and .03 in small steps of .002, and trees between 250 and 800 in steps of 50.

In the empirical study, the samples of the regular and careless respondents in the training sample were almost equal in size, providing optimal learning conditions for the algorithm. In the simulation study, however, we used the same prevalence of careless respondents in the data generation process for both the training and the test sample (5%, 10%, or 15% depending on the condition, see Appendix A for detailed information on the sample size). It is a well-known fact that machine learning algorithms that are trained on highly unbalanced data tend to bias the prediction in favor

of the majority group leading to low sensitivity (e.g., Van Hulse et al., 2007). To avoid this statistical artifact, we used up-sampling (e.g., García et al., 2012), an often applied technique, in which observations in the minority group (i.e., careless respondents) are replicated to match the sample size of the majority group (i.e., regular respondents). Please note the test sample is unaffected by the up-sampling. The analyses were repeated 1,000 times for different random samples generated in the simulation study and randomly drawn samples in the empirical study (training vs. testing sample).

To evaluate the binary classification into careless respondents (CR) and regular respondents (RR), we report five performance metrics (for an overview see Tharwat, 2018) based on the number of correctly identified CR (true positives, TP), incorrectly identified CR (false positives, FP), correctly identified RR (true negatives, TN), and incorrectly identified RR (false negatives, FN): (a) sensitivity or true positive rate or recall (= TP/(TP + FN)), (b) specificity or true negative rate (= TN/(FP + TN)), (c) precision or positive predictive value (= TP/(TP + FP)), (d) accuracy (= (TP + TN)/ (P + N)), and (e) the balanced accuracy, which is the mean of sensitivity and specificity. These indices were calculated in the test sample for each of the 1,000 iterations of the analyses.

Results

Study 1: Simulation Study

In the simulation study, we considered a 3×3 design, varying the amount of careless respondents in the overall sample (5%, 10%, 15%) and considering three types of response styles (random, midpoint, fixed pattern respondents). To reduce the complexity of the results, we only elaborate the results of the 10% condition because they were similar to the results of the other conditions (see Tables S1 and S2 in the online repository). Moreover, the size of the test sample ($n = 180$) in this condition is identical to the empirical data, which enables a comparison to the empirical data. Also, most frequently, prevalence rates between 10% and 15% are reported in the literature (Kurtz & Parrish, 2001; Meade & Craig, 2012). Table 2 summarizes the means and standard deviations for different performance metrics across 1,000 iterations for the three types of careless responding. The main findings of the simulation study are as follows: First, several detection methods performed well with respect to specificity; that is, regular respondents were in most cases correctly classified. In contrast, the rate of correctly identified careless respondents (i.e., sensitivity) is often low, which is in line with previous results (e.g., Niessen et al., 2016). Second, traditional detection mechanisms such as the Mahalanobis distance and the polytomous Z_h statistics performed excellent in terms of specificity and sensitivity for random respondents, whereas the sensitivity was low for the midpoint and pattern respondents. Third, although for the midpoint and pattern respondents the specificity for the traditional detection mechanisms (with the exception of the Z_h statistic) can be considered good, only the IRV yielded acceptable values for sensitivity. Fourth, the GBM performed

Table 2. Classification Accuracy of Traditional and Machine Learning Algorithms With Simulated Data.

Item	Maha.	Antonyms	EvenOdd	Longstring	IRV	Z_h	GBM
Random respondents							
Accuracy	1.00 (.00)	.87 (.02)	.88 (.02)	.83 (.02)	.96 (.01)	.92 (.02)	.96 (.01)
Sensitivity	1.00 (.01)	.11 (.07)	.49 (.12)	.01 (.03)	.80 (.07)	1.00 (.00)	.74 (.11)
Specificity	1.00 (.00)	.96 (.02)	.93 (.02)	.93 (.02)	.98 (.01)	.91 (.02)	.99 (.01)
Precision	.97 (.04)	.22 (.15)	.43 (.09)	.02 (.04)	.80 (.07)	.55 (.05)	.90 (.08)
Balanced accuracy	1.00 (.01)	.53 (.04)	.71 (.06)	.47 (.02)	.89 (.04)	.95 (.01)	.86 (.05)
Midpoint respondents							
Accuracy	.91 (.01)	.90 (.02)	.88 (.02)	.87 (.02)	.94 (.02)	.74 (.03)	.98 (.01)
Sensitivity	.18 (.08)	.27 (.10)	.50 (.12)	.37 (.11)	.68 (.09)	.51 (.12)	.88 (.08)
Specificity	.99 (.01)	.97 (.02)	.93 (.02)	.93 (.02)	.96 (.01)	.77 (.02)	.99 (.01)
Precision	.65 (.22)	.48 (.16)	.43 (.09)	.36 (.10)	.68 (.09)	.20 (.04)	.90 (.07)
Balanced accuracy	.59 (.04)	.62 (.05)	.71 (.06)	.65 (.06)	.82 (.05)	.64 (.06)	.94 (.04)
Pattern respondents							
Accuracy	.87 (.01)	.80 (.06)	.86 (.02)	.85 (.02)	.95 (.02)	.72 (.03)	.96 (.02)
Sensitivity	.03 (.04)	.08 (.13)	.10 (.08)	.20 (.09)	.76 (.10)	.50 (.11)	.93 (.08)
Specificity	.96 (.01)	.88 (.06)	.93 (.02)	.93 (.02)	.97 (.01)	.75 (.03)	.96 (.02)
Precision	.09 (.11)	.06 (.10)	.11 (.08)	.23 (.10)	.76 (.10)	.18 (.04)	.74 (.10)
Balanced accuracy	.50 (.02)	.48 (.07)	.51 (.04)	.56 (.05)	.87 (.06)	.63 (.06)	.95 (.04)

Note. Results are means and standard deviations across 1,000 simulated data sets (n_{test} = 180 with 10% careless respondents). Maha. = Mahalanobis distance; Antonyms = psychometric antonyms; EvenOdd = even–odd consistency; Longstring = Longstring Index; IRV = intraindividual response variability; Z_h = polytomous IRT person-fit statistic; GBM = gradient boosting machine. IRT = item response theory.

excellent with balanced accuracies of .86 for random respondents, .94 for midpoint respondents, and .95 for pattern respondents. Also, the precision was high, indicating that only a small fraction of the flagged respondents were false positives (see Table 2). Please note that the high classification accuracy of the gradient boosted trees indicate that the up-sampling procedure mostly compensated for potential adverse effect of unbalanced groups in the training-validation sample. As the number of careless respondents in the training data set increases, the performance only slightly improves (see Tables S1 and S2 in the online repository for the 5% and 15% careless respondents condition).

Study 2: Empirical Study

To compare the detection mechanisms for the empirical study, we first present the different performance metrics (see Table 3). To illustrate the variability of the empirical results, we drew 1,000 test samples from the empirical data (i.e., 30% of the overall sample) that comprises 90% regular and 10% careless respondents (i.e., 162 and 18 respondents, respectively, in absolute numbers) and report means and standard deviations of the respective classification indices. In addition to a GBM that relied on the response vectors (GBM_{Res}), we included two models that used response times only (GBM_{RT}) and a combination of both information (GBM_{Res+RT}), respectively. Since the sample size of the holdout sample is identical to the setup of the simulation study, the results can be compared with the simulation results reported above. In comparison to the simulated conditions, all detection mechanisms performed worse in the empirical study, demonstrating that it is much harder to flag aberrant responses. For the traditional detection methods, the picture that has already emerged for the midpoint and the pattern respondents in the simulation study was replicated with good values for specificity, but low values for sensitivity. Importantly, sensitivity and specificity were only mediocre for the machine learning approach: The average *false positive rate* (= 1 − specificity) was .29 for the most complex model and the average *miss rate* (or false negative rate = 1 − sensitivity) was .40. The *precision* (or positive predictive value) as the proportion of correctly identified careless respondents out of all flagged respondents is only .19, which means that the flagging of participants is false in 81% of the cases.

   Sensitivity, as the proportion of careless respondents that are correctly identified, and *specificity*, which denotes the proportion of regular respondents that are correctly identified, are mutually dependent—the increase of one leads to a reduction of the other (i.e., sensitivity-specificity trade-off). In the present context, we used an optimization function for the GBM algorithm that took into account both performance metrics equally. Depending on the research question and the consequences of misclassification other priorities might be better suited. Thus, researchers should ask whether Type I and Type II errors are equally unfavorable for their research or application, a question that can only be answered depending on the consequences of incorrect classifications. Improperly flagging responses as careless (Type I error)

Table 3. Classification Accuracy of Traditional and Machine Learning Algorithms With Empirical Data (10% Prevalence).

Item	Maha.	Antonyms	EvenOdd	Longstring	IRV	Z_h	GBM$_{Res}$	GBM$_{RT}$	GBM$_{Res + RT}$
Accuracy	.80 (.02)	.86 (.02)	.84 (.02)	.82 (.02)	.83 (.01)	.67 (.02)	.59 (.04)	.70 (.05)	.70 (.04)
Sensitivity	.18 (.08)	.11 (.08)	.19 (.09)	.18 (.09)	.14 (.07)	.38 (.11)	.61 (.12)	.56 (.12)	.60 (.12)
Specificity	.87 (.02)	.94 (.02)	.91 (.02)	.90 (.02)	.90 (.01)	.70 (.02)	.58 (.04)	.71 (.05)	.71 (.05)
Precision	.13 (.06)	.17 (.11)	.19 (.08)	.16 (.07)	.14 (.07)	.12 (.03)	.14 (.02)	.18 (.04)	.19 (.04)
Balanced accuracy	.53 (.04)	.53 (.04)	.55 (.04)	.54 (.04)	.52 (.04)	.54 (.06)	.60 (.06)	.64 (.06)	.66 (.06)

Note. Results are means and standard deviations across 1,000 random test samples of the empirical data (n_{test} = 180 with 10% careless respondents). Maha. = Mahalanobis distance; Antonyms = psychometric antonyms; EvenOdd = even–odd consistency; Longstring = Longstring Index; IRV = intraindividual response variability; Z_h = polytomous IRT person-fit statistic; GBM = gradient boosting machine; Res = responses; RT = response times. IRT = item response theory.

may unduly reduce the sample size, increase the survey costs, and bias the results, while not detecting real careless respondents (Type II error) might also bias and invalidate test results (also known as the MTurk quality crisis, e.g., Kennedy et al., 2020). In our analysis, we used a negative logistic loss function for the gradient boost algorithm to make the model more sensitive to the class of careless respondents, resulting in higher balanced accuracy rates (i.e., arithmetic mean of sensitivity and specificity) in comparison to the traditional detection mechanisms. Adopting a more conservative approach, that is, raising the bar to flag potential careless respondents, the accuracy or the receiver operating characteristic curve metric would be a better metric to optimize.

In addition to the *sensitivity* and *specificity*, we present the *balanced accuracy*— that is, the arithmetic mean of sensitivity and specificity—as a compromise between both performance measures of a binary classification test (see Figure 2). The main findings concerning balanced accuracy can be summarized as follows. First, both the traditional detection mechanisms and the machine learning algorithm performed considerably worse in the realistic setting versus the simulated data condition. This is particularly evident in the *precision*, which in the empirical study reached only about one fourth of the values from the simulation study. Second, the GBM achieved a slightly better trade-off between sensitivity and specificity. The balanced accuracy was best if the responses and the (aggregate) response times were included in the model. It should be mentioned that adding further predictors such as the number of response option changes, or the traditional detection measures did not lead to a substantial increment in the prediction quality in our study. The values had a large variation; that is, the performance metrics varied considerably depending on the specific sample. This is most likely due to the small sample size (for machine learning standards). It can also be taken as evidence that careless response processes under real-world conditions are much more heterogeneous than artificial ones (see also Denison & Wiernik, 2020) and, therefore, require a larger learning pool.

Discussion

Comparing the results of the simulation and the empirical study showed substantial differences in almost all performance measures for the detection of careless respondents (cf., Hong et al., 2020; Meade & Craig, 2012; Niessen et al., 2016). Whereas the simulation study highlighted substantial advantages of the studied machine learning algorithm, respective benefits were more modest in the empirical application. One reason for this discrepancy might be that some ''regular respondents'' in the empirical study might have (at least in part) responded carelessly, even though they claimed to adhere to the instructions. In turn, there also might have been ''careless respondents'' that actually paid attention out of curiosity, reactance, or misunderstanding. Both effects most likely restricted the maximum obtainable sensitivity in the empirical study. Thus, the best specificity value found was .71 for the GBM, which might constitute an upper bound that can be expected in applied settings.
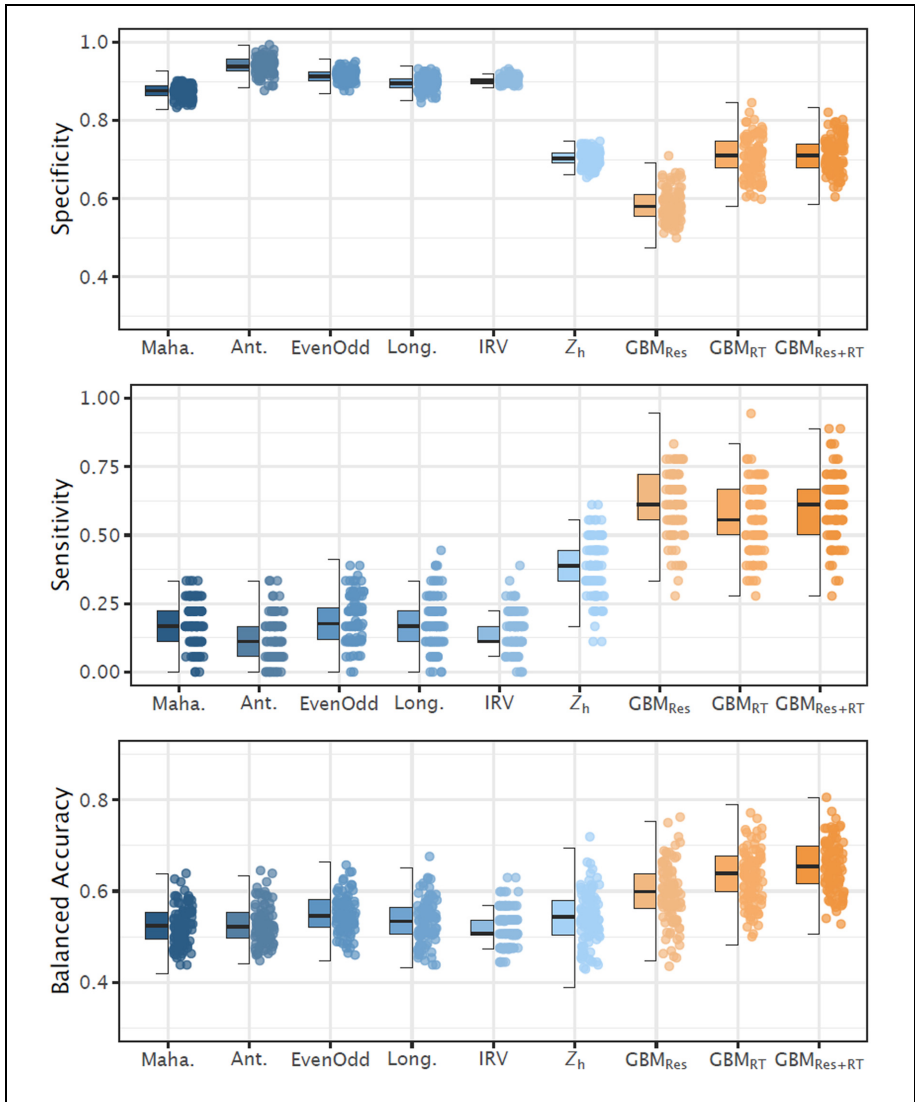
Figure 2. Specificity, sensitivity, and balanced accuracy across detection mechanisms. Maha. = Mahalanobis distance; Ant. = psychometric antonyms; EvenOdd = even–odd consistency; Long. = Longstring Index; IRV = intraindividual response variability; Z_h = polytomous IRT person-fit statistic; GBM = gradient boosting machine; Res = responses; RT = response times. Left side: The boxplot reflects the interquartile range, the solid line represents the median, and the whiskers represent minimum/maximum values within 1.5 times the interquartile range. Right side: Jittered point plot of 100 randomly drawn values. IRT = item response theory.

Another explanation for the observed discrepancy between the two studies might be that the simulated types of careless responding are inferior characterizations of actual careless responding observed in the real world. For example, humans are usually not capable of generating completely random answers as assumed in many previous simulation studies; a circumstance that is incidentally also exploited in the detection of malingering (Orthey et al., 2019). Apparently, human's response behavior is much more erratic than what is assumed in rather straightforward, pattern-like responses of simulation studies. A somewhat disappointing conclusion, therefore, is that findings of simulation studies do not directly translate into an ecologically valid context.

The present study falls in line with a series of critical voices on the hype about machine learning algorithms in fields such as personality science (e.g., Stachl et al., 2020) or clinical research (Cearns et al., 2019). In psychology, data quality and structure are frequently not comparable to computer sciences. We often rely on small sample sizes, which require special attention to avoid overfitting (Vabalas et al., 2019). Moreover, typical indicators in psychological research such as questionnaire or test data are subject to measurement errors, which might obscure the true relationship underlying the data (Jacobucci & Grimm, 2020). Careless (and also regular) responding might also manifest in many idiosyncratic ways that simply do not follow a clear predictable pattern. All these circumstances might account for the sobering results regarding the usefulness of machine learning in psychology.

The Potential of Machine Learning Algorithms

Gradient boosted trees are effective in cases where patterns in relation to a labeled outcome can be derived. In the simulation study, the GBM could ''learn'' not only because the pattern of careless (and regular) responding was rule-based and separable from each other but also because the outcome was provided. This knowledge of the outcome is used to optimally set up the predictive model. In contrast, traditional detection mechanisms are unsupervised algorithms that operate without such information, which makes it an almost unfair comparison. However, in real settings, information on the extent of regular responding is not given and one has to rely on other (partially unreliable) information such as asking participants if they answered diligently. One advantage of the GBM is that they can easily deal with plenty of features and split them optimally. In case of response times, GBM can flexibly handle this additional source of information, whereas traditional detection mechanisms have to work with a predefined cutoff. The importance of response times in the context of detecting aberrant responses has also been stressed elsewhere (e.g., Leiner, 2019). Usually, detection mechanisms based on response times focus on extremely rapid responses (e.g., 2 seconds/item; Huang et al., 2012; Wise & Kong, 2005). However, a speedy response is not necessarily synonymous with careless responding. In turn—although the instruction of the careless responding condition specifically requested to get through the questionnaire ''as quickly as possible''—it is also conceivable that a respondent is unusually slow (see also Niessen et al., 2016), for example, if they are

distracted. In principle, the GBM algorithm can detect either very fast or extremely slow respondents. In the present study, however, we observed a huge overlap of the response time distributions across conditions and larger differences at the lower tail of the distribution (see Figure S2 in the online repository).

Arguably, there is still potential to improve the classification accuracy of the GBM since even the "best" model is not optimal for practical purposes. For example, the predication model could incorporate additional information such as scales measuring test-taking motivation, conscientiousness, or diligence as additional predictors. It might also be worthwhile to use bogus items, instructed response items, or even paradata that are much harder to fake (e.g., Buchanan & Scofield, 2018; Steger, Schroeders, & Wilhelm, 2020). One promising candidate for such a piece of information would be the item number indicating the progress in the completion of the questionnaire. Careless responding is not dichotomous, rather insufficient effort can take place sooner or later depending on the commitment or fatigue of the test-taker and the length of the questionnaire (Gibson & Bowling, 2020). Costs (e.g., time or cognitive effort) and benefits (e.g., financial gratification or intrinsic motivation) are set off against each other to evaluate the willingness to further participate at a specific point of time. That is, not only the willingness to answer superficially increases but also the probability to abort the questionnaire (Bowling et al., 2020). Thus, it seems plausible to consider the phenomena of attrition and careless responding in tandem (Meade et al., 2017). Yet even in a complete data set, items at the beginning of the questionnaire might be more decisive for distinguishing careless responding than those items at the end, because it is likely that the motivation of even engaged participants decreases till the end of the questionnaire (see also Wise & Kong, 2005).

Disadvantages of Available Detection Indices

A serious drawback of several traditional detection indices (e.g., Longstring Index, psychometric antonyms) is that the cutoffs are usually set manually and data-driven without controlling for the risk of overfitting, which is why recommendations for cutoffs vary in the literature depending on the data set used to derive them (e.g., Niessen et al., 2016). Nonetheless, traditional indicators to detect careless responding—whether statistical outlier methods, consistency analyses, and response pattern functions—have the advantage that they can be applied with minimal prerequisites to new questionnaires. In contrast, supervised machine learning algorithms such as gradient boosted trees require labeled data sets, need an extensive training phase, and the prediction models are bound to a specific set of items. Generalizations to other data sets, samples, and situations are *not* possible, because every examination is highly specific in terms of items and persons. These are major drawbacks. In principle, it would be possible to follow the design we outlined in the study—that is, implement an experimental prestudy to build a "detection model" and to subsequently clean the data of the main study. However, given the present set of results, we are skeptical that such an approach will provide satisfying results

because careless responding in real life is apparently a lot harder to detect than obvious misconduct.

We would also like to caution that decisions of traditional detection approaches, which test-takers respond carelessly, are based not only on the gathered data but also on theoretical assumptions of the data generation process itself. Model-based approaches rely on the assumption that normality can be captured sufficiently well by a given model, so that non-matching observations are considered outliers. For example, in the even–odd consistency analyses, participants whose answers to similar questions are contradictory or inconsistent are more conspicuous. However, such a categorization involves certain risks, since it decides a priori on the correctness of data, and flags data that are not in line with a certain model or deviate the most. However, these model assumptions can be erroneous themselves, bearing the risk of shaping the data until it fits the model assumption, which might render the cure worse than the disease. Also, using machine learning algorithms comes at a cost because they are more data-driven and are considered black box approaches. As an outlook for future research, it seems promising to reunite the approaches of explanation and prediction (Yarkoni & Westfall, 2017) by enriching machine learning with assumptions about the process (and ultimately the causes) of careless responding (e.g., the risk to engage in careless responding should increase with test length and cognitive demand, which could be accounted for in the modeling procedure).

Authors' Note

Declaration of Conflicting Interests

Funding

ORCID iDs

Ulrich Schroeders https://orcid.org/0000-0002-5225-1122
Timo Gnambs https://orcid.org/0000-0002-6984-1276

Supplemental Material

Supplemental material for this article is available online at https://osf.io/mct37.

Notes

1. The differences between the detection measures that belong to the group of consistency analysis versus the response pattern functions, thus, is the point of comparison: In the former, it is mainly on the individual response vector, while in the latter, it also takes into account the responses of the sample. Please note that the distinction is not clear-cut and some mechanisms can be assigned differently (e.g., psychometric antonyms).
2. The random respondents correspond to the condition of ''uniformly random careless data'' in Meade and Craig (2012), whereas midpoint respondents correspond to the condition of ''normally distributed random careless data.'' The fixed pattern respondents were not previously simulated.
3. We used the R code rep(sample.int(5,1):sample.int(5,1), length.out=60) to generate the sequence.
4. The design also included a third condition that instructed respondents to deliberately respond carelessly with the intention not to be discovered. This induced response style resembles faking behavior (Geiger et al., 2018), but without impression management. Instead, test-takers were instructed to complete the questionnaire quickly without being recognized as an inattentive respondent. To streamline the present article, this condition has not been examined further.

References

Arias, V. B., Garrido, L. E., Jenaro, C., Martínez-Molina, A., & Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-020-01401-8

Ashton, M., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment*, *91*(4), 340-345. https://doi.org/10.1080/00223890902935878

Berk, R. A. (2017). *Statistical learning from a regression perspective*. Springer.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment*, *4*(3), 340-345. https://doi.org/10.1037/1040-3590.4.3.340

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2020). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*. Advance online publication. https://doi.org/10.1177/1094428120947794

Bowling, N. A., & Huang, J. L. (2018). Your attention please! Toward a better understanding of research participant carelessness. *Applied Psychology*, *67*(2), 227-230. https://doi.org/10.1111/apps.12143

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32. https://doi:10.1023/A:1010933404324

Buchanan, E. M., & Scofield, J. E. (2018). Methods to detect low quality data and its implication for psychological research. *Behavior Research Methods*, *50*(6), 2586-2596. https://doi.org/10.3758/s13428-018-1035-6

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, *9*, Article 271. https://doi.org/10.1038/s41398-019-0607-2

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review*, *23*(4), 486-501. https://doi.org/10.1177/0894439305278972

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement*, *70*(4), 596-612. https://doi.org/10.1177/0013164410366686

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

Curran, P. G., & Hauser, K. A. (2019). I'm paid biweekly, just not by leprechauns: Evaluating valid-but-incorrect response rates to attention check items. *Journal of Research in Personality*, *82*, 103849. https://doi.org/10.1016/j.jrp.2019.103849

Denison, A., & Wiernik, B. (2020). Careless response processes are heterogeneous: Comment on Goldammer et al. (2020). *PsyArXiv Preprints*. https://psyarxiv.com/9rk4w/

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67-86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Dunn, A. M., Heggestad, E. D., Shanock, L. R., & Theilgard, N. (2018). Intra-individual response variability as an indicator of insufficient effort responding: Comparison to other indicators and relationships with individual differences. *Journal of Business and Psychology*, *33*(1), 105-121. https://doi.org/10.1007/s10869-016-9479-0

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement*, *32*(3), 224-247. https://doi.org/10.1177/0146621607302479

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293-314. https://doi.org/10.1093/nsr/nwt032

Ferrari, P. A., & Barbiero, A. (2012). Simulating ordinal data. *Multivariate Behavioral Research*, *47*(4), 566-589. https://doi.org/10.1080/00273171.2012.692630

Figurska, M., Stańczyk, M., & Kulesza, K. (2008). Humans cannot consciously generate random numbers sequences: Polemic study. *Medical Hypotheses*, *70*(1), 182-185. https://doi.org/10.1016/j.mehy.2007.06.038

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

García, V., Sánchez, J. S., & Mollineda, R. A. (2012). On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, *25*(1), 13-21. https://doi.org/10.1016/j.knosys.2011.06.013

Geiger, M., Olderbak, S., Sauter, R., & Wilhelm, O. (2018). The ''g'' in faking: Doublethink the validity of personality self-report measures for applicant selection. *Frontiers in Psychology*, *9*, 1-15. https://doi.org/10.3389/fpsyg.2018.02153

Gibson, A. M., & Bowling, N. A. (2020). The effects of questionnaire length and behavioral consequences on careless responding. *European Journal of Psychological Assessment*, *36*(2), 410-420. https://doi.org/10.1027/1015-5759/a000526

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika*, *72*(2), 159-180. https://doi.org/10.1007/s11336-003-1081-5

Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods*, *47*, 1237-1259. https://doi.org/10.3758/s13428-014-0533-4

Gnambs, T., & Schroeders, U. (2020). Cognitive abilities explain wording effects in the Rosenberg Self-Esteem Scale. *Assessment*, *27*(2), 404-418. https://doi.org/10.1177/1073191117746503

Göritz, A. S., Borchert, K., & Hirth, M. (2019). Using attention testing to select crowdsourced workers and research participants. *Social Science Computer Review*. Advance online publication. https://doi.org/10.1177/0894439319848726

Greenwell, B., Boehmke, B., Cunnigham, J., & GBM Developers. (2019). *gbm: Generalized boosted regression models* (Version 2.1.5) [Computer software]. https://CRAN.R-project.org/package=gbm

Gummer, T., Roßmann, J., & Silber, H. (2018). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*. Advance online publication. https://doi.org/10.1177/0049124118769083

Hong, M., Steedle, J. T., & Cheng, Y. (2020). Methods of detecting insufficient effort responding: Comparisons and practical recommendations. *Educational and Psychological Measurement*, *80*(2), 312-345. https://doi.org/10.1177/0013164419865316

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. *Journal of Business and Psychology*, *27*(1), 99-114. 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*(3), 828-845. https://doi.org/10.1037/a0038510

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809-816. https://doi.org/10.1177/1745691620902467

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning*. Springer.

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

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education*, *16*(4), 277-298. https://doi.org/10.1207/S15324818AME1604_2

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614-629. https://doi.org/10.1017/psrm.2020.6

Kieslich, P. J., & Henninger, F. (2017). Mousetrap: An integrated, open-source mouse-tracking package. *Behavior Research Methods*, *49*(5), 1652-1667. https://doi.org/10.3758/s13428-017-0900-z

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(15), 5802-5805. https://doi.org/10.1073/pnas.1218772110

Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, *45*, 527-563. https://doi.org/10.1007/s41237-018-0063-y

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1-26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M. (2020). *Caret* (Version 6.0-85) [Computer software]. https://cran.r-project.org/web/packages/caret

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. https://doi.org/10.1207/S15327752JPA7602_12

Leiner, D. J. (2019). Too fast, too straight, too weird: Non-reactive indicators for meaningless data in internet surveys. *Survey Research Methods*, *13*(3), 229-248. https://doi.org/10.18148/srm/2019.v13i3.7403

Leiner, D. J. (2020). *SoSci survey* (Version 3.2.05-i) [Computer software]. http://www.soscisurvey.com

Litman, L., & Robinson, J. (2020). *Conducting online research on Amazon Mechanical Turk and beyond*. Sage.

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

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61-83. https://doi.org/10.1016/j.jrp.2013.09.008

Marjanovic, Z., Holden, R., Struthers, W., Cribbie, R., & Greenglass, E. (2015). The inter-item standard deviation (ISD): An index that discriminates between conscientious and random responders. *Personality and Individual Differences*, *84*, 79-83. https://doi.org/10.1016/j.paid.2014.08.021

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*(3), 450-470. https://doi.org/10.1037/a0019216

Meade, A. W., Allred, C. M., Pappalardo, G., & Stoughton, J. W. (2017). Careless response and attrition as sources of bias in online survey assessments of personality traits and performance. *Computers in Human Behavior*, *76*, 417-430. https://doi.org/10.1016/j.chb.2017.06.032

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

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*(2), 111-120. https://doi.org/10.1177/014662169401800202

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, *9*(1), 3-8. https://doi.org/10.1207/s15324818ame0901_2

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exeggeration. *Journal of Clinical and Experimental Neuropsychology*, *24*(8), 1094-1102. https://doi.org/10.1076/jcen.24.8.1094.8379

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. https://doi.org/10.1002/1097-4679 (198903)45:2<239::AID-JCLP2270450210 > 3.0.CO;2-1

Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016). Detecting careless respondents in web-based questionnaires: Which method to use? *Journal of Research in Personality*, *63*, 1-11. https://doi.org/10.1016/j.jrp.2016.04.010

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In F. Kreuter (Ed.), *Improving surveys with paradata* (pp. 43-72). Wiley. https://doi.org/10.1002/9781118596869.ch3

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. https://doi.org/10.1016/j.jesp.2009.03.009

Orthey, R., Vrij, A., Meijer, E., Leal, S., & Blank, H. (2019). Eliciting response bias within forced choice tests to detect random responders. *Scientific Reports*, *9*, Article 8724. https://doi.org/10.1038/s41598-019-45292-y

Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions: Items as personality markers. *European Journal of Personality*, *32*(3), 186-201. https://doi.org/10.1002/per.2147

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289-310. https://doi.org/10.1214/10-STS330

Soderberg, C. K. (2018). Using OSF to share data: A step-by-step guide. *Advances in Methods and Practices in Psychological Science*, *1*(1), 115-120. https://doi.org/10.1177/2515245918757689

Sotaridona, L. S., & Meijer, R. R. (2002). Statistical properties of the K-index for detecting answer copying. *Journal of Educational Measurement*, *39*(2), 115-132. https://doi.org/10.1111/j.1745-3984.2002.tb01138.x

Stachl, C., Pargent, F., Hilbert, S., Harari, G. M., Schoedel, R., Vaid, S., Gosling, S. D., & Bühner, M. (2020). Personality research and assessment in the era of machine learning. *European Journal of Personality*. Advance online publication. https://doi.org/10.1002/per.2257

Steger, D., Schroeders, U., & Gnambs, T. (2020). A meta-analysis of test scores in proctored and unproctored ability assessments. *European Journal of Psychological Assessment*, *36*(1), 174-184. https://doi.org/10.1027/1015-5759/a000494

Steger, D., Schroeders, U., & Wilhelm, O. (2020). Caught in the act: Predicting cheating in unproctored knowledge assessment. *Assessment*. Advance online publication. https://doi.org/10.1177/1073191120914970

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software*, *74*(5), 1-27. https://doi.org/10.18637/jss.v074.i05

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*. Advance online publication. https://doi.org/10.1016/j.aci.2018.08.003

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS One*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Van Hulse, J., Khoshgoftaar, T. M., & Napolitano, A. (2007, June). Experimental perspectives on learning from imbalanced data. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning* (pp. 935-942). https://doi.org/10.1145/1273496.1273614

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, *68*(3), 456-477. https://doi.org/10.1111/bmsp.12054

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469-501. https://doi.org/10.3102/1076998618767123

Weiner, S. P., & Dalessio, A. T. (2006). Oversurveying: Causes, consequences, and cures. In A. I. Kraut (Ed.), *Getting action from organizational surveys: New concepts, technologies, and applications* (pp. 294-311). Wiley.

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

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. https://doi.org/10.1007/s10862-005-9004-7

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. https://doi.org/10.1177/1745691617693393

Yentes, R., & Wilhelm, F. (2018). *Careless* (Version 1.1.3) [Computer software]. https://cran.r-project.org/web/packages/careless

Zopluoglu, C. (2019). Detecting examinees with item preknowledge in large-scale testing using Extreme Gradient Boosting (XGBoost). *Educational and Psychological Measurement*, *79*(5), 931-961. https://doi.org/10.1177/0013164419839439