Received: 14 December 2022 | Revised: 16 November 2023 | Accepted: 22 January 2024

DOI: 10.1111/spc3.12941

REVIEW ARTICLE

WILEY

Open science perspectives on machine learning for the identification of careless responding: A new hope or phantom menace?

Andreas Alfons[1] | **Max Welz**[1,2]

[1]Department of Econometrics, Erasmus University Rotterdam, Rotterdam, Netherlands

[2]Department of Public Health, Erasmus MC—University Medical Center Rotterdam, Rotterdam, Netherlands

Correspondence
Max Welz, Department of Econometrics, Erasmus University Rotterdam, PO Box 1738, 3000 DR Rotterdam, Netherlands.
Email: welz@ese.eur.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: VI.Vidi.195.141

Abstract

Powerful methods for identifying careless respondents in survey data are not just important to ensure the validity of subsequent data analyses, they are also instrumental for studying the psychological processes that drive humans to respond carelessly. Conversely, a deeper understanding of the phenomenon of careless responding enables the development of improved methods for the identification of careless respondents. While machine learning has gained substantial attention and popularity in many scientific fields, it is largely unexplored for the detection of careless responding. On the one hand, machine learning algorithms can be highly powerful tools due to their flexibility. On the other hand, science based on machine learning has been criticized in the literature for a lack of reproducibility. We assess the potential and the pitfalls of machine learning approaches for identifying careless respondents from an open science perspective. In particular, we discuss possible sources of reproducibility issues when applying machine learning in the context of careless responding, and we give practical guidelines on how to avoid them. Furthermore, we illustrate the high potential of an unsupervised machine learning method for the identification of careless respondents in a proof-of-concept simulation experiment. Finally, we stress the necessity of building an open data repository

with labeled benchmark data sets, which would enable the evaluation of methods in a more realistic setting and make it possible to train supervised learning methods. Without such a data repository, the true potential of machine learning for the identification of careless responding may fail to be unlocked.

KEYWORDS
careless responding, guidelines, machine learning, open science, reproducibility, unsupervised learning

1 | INTRODUCTION

Participants in surveys may not comply with survey instructions due to, for instance, fatigue, lack of motivation, or failure to correctly understand the instructions. This phenomenon is known as *careless responding* in the psychology literature and has been identified as a major threat to the validity of research results (e.g., Arias et al., 2020; Huang, Liu, & Bowling, 2015; McGrath et al., 2010; Meade & Craig, 2012; Woods, 2006). If sufficiently many responses are careless, any knowledge drawn from these responses is potentially biased and at worst inaccurate. Considering the ubiquity of survey data in psychology together with the societal impact of the field, this can have dire consequences for future research, policy decisions, and societal outcomes. For example, careless responding may inflate or deflate the effect size of an experimental manipulation, or even reverse the sign of the effect. Such over- or underestimation of the effectiveness of the manipulation may result in type I or type II errors in hypothesis testing (cf. Arias et al., 2020), possibly leading to inefficient or downright harmful policies, or effective policies not being implemented. Since careless responding is widely prevalent and suspected to be present in all survey data (Ward & Meade, 2023), it is broadly recommended to screen survey data for careless responses (e.g., Arias et al., 2020; Huang, Liu, & Bowling, 2015; Meade & Craig, 2012), for which a plethora of different methods exist (see, e.g., tab. 1 in Arthur et al., 2021, for an overview). Effective screeners for careless responding are of particular relevance in social and personality psychology, as accurately identifying careless respondents is essential for understanding the psychological processes that drive participants to respond carelessly. For instance, careless responding has been linked to personal characteristics (e.g. Kim, Dykema, et al., 2018) and personality traits (e.g., Bowling et al., 2016). In turn, enhanced psychological understanding of careless responding can lead to further improvement of screeners for careless responding. While machine learning methods have not received much consideration in the literature on the identification of careless responding, their potential for this task has recently been pointed out in a review paper by Arthur et al. (2021), with pioneering studies being conducted by Schroeders et al. (2022) and Welz and Alfons (2023).

In machine learning, there is a distinction between *supervised* and *unsupervised* learning (e.g., Hastie et al., 2009). Supervised learning algorithms aim to predict a response variable of interest. The data are split into a training set and a test set, with the algorithm learning predictive relationships on the training set, and the test set being used to evaluate predictive performance (i.e., to compare the predictions with the observed outcomes). Unsupervised learning, on the other hand, is exploratory in nature, and corresponding algorithms are trained on the full data set. There is no observed response variable, hence it is not possible to evaluate predictions on a test set. For instance, regression and classification are supervised learning tasks, whereas clustering is an unsupervised learning task. In practice, identifying careless responding is typically an unsupervised learning problem, since a researcher does not know in advance which respondents in a given survey are careless and which ones are not. Supervised learning techniques are only suitable subject to availability of training data with high-quality labels for careless respondents, for example, from an earlier survey containing the same items. Then, a supervised learning algorithm could be trained on the earlier survey

using the carelessness labels as the response variable, and the trained algorithm could be applied to the new survey to identify careless respondents. In a recent study, Schroeders et al. (2022) collect training data for supervised learning via an experimental manipulation, in which some participants are instructed to respond carelessly, whereas others are instructed to respond truthfully. However, Schroeders et al. (2022) themselves question whether participants complied with those instructions, and Ulitzsch et al. (2022a, 2022b) further criticize that respondents who are being instructed to respond carelessly may not behave in a comparable manner to those displaying careless responding outside of such an experiment.

Machine learning in psychology has seen increasing popularity, for example, in personality assessment (Miotto et al., 2022), measuring emotions (Kleinberg et al., 2020), predicting depression (Fokkema & Strobl, 2020), and detecting differential item functioning (Strobl et al., 2015). Then again, other studies across various fields have criticized science based on machine learning for a lack of reproducibility (e.g., Kapoor & Narayanan, 2022, and references therein). Even before widespread adoption of machine learning, the reproducibility crisis in the social sciences has rightfully received widespread attention in the literature (e.g., Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012). This begs the following fundamental question: If machine learning may amplify reproducibility issues, is it reasonable and responsible to explore this approach for the identification of careless responding? On the other hand, various open science practices have been proposed to address the reproducibility crisis (e.g., Miguel et al., 2014), such as registering preanalysis plans or sharing data and code from analyses, and we can draw from this experience.

In this paper, we discuss the benefits and risks of using machine learning for the detection of careless responding, and we outline how open science practices mitigate reproducibility issues with machine learning. We provide a proof-of-concept simulation experiment that illustrates the potential of machine learning in the context of careless responding, and we highlight relevant directions for future research. Moreover, we argue how open science practices, most notably building an open data repository of labeled benchmark data sets, are at the heart of solving persistent issues due to careless responding. Finally, we provide researchers with recommendations for "good enough" practices.

2 | THE PHENOMENON OF CARELESS RESPONDING

Numerous definitions of careless responding exist in the literature, with likely the most commonly used definition being *"a response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses"* (Huang et al., 2012). An alternative definition is given by Ward and Meade (2023): *"careless responding occurs when participants are not basing their response on the item content, and it can occur when a respondent does not read an item, does not understand an item, or is unmotivated to think about what the item is asking."*

The latter definition highlights that careless responding may not be intentional (e.g., Huang et al., 2012; Ward et al., 2017; Ward & Pond, 2015), such as participants misunderstanding an item due to poor or ambiguous wording. For instance, items that combine multiple statements into one item (e.g., *"How important is it to you to get good grades and please your parents?"*; Gehlbach, 2015) or negatively-worded items may confuse survey participants due to increased difficulty of cognitive processing (Chyung et al., 2018; Gehlbach, 2015; Swain et al., 2008; Weijters & Baumgartner, 2012), resulting in a higher risk of careless responses. A more fundamental issue with traditional rating-scale items was raised by Uher (2023), arguing that rating scales might be interpreted differently by each participant, which can lead to misinterpretation and, in turn, responses that fall under the definition of careless responses even though the intention behind such responses is not actually careless. In general, deficiencies in survey design can precipitate careless responding, therefore survey designers should write precise, clear, and unambiguous questions, avoid vagueness, use simple, clear, and neutral language (e.g., Stantcheva, 2022), and avoid excessively lengthy questionnaires to prevent carelessness due to survey fatigue (Berry et al., 1992; Bowling et al., 2021; Ward

et al., 2017). Besides survey design, the literature has identified that careless responding is influenced by participant personality. For instance, Bowling et al. (2016) find that conscientiousness, agreeableness, extraversion, and emotional stability are each negatively related to carelessness, while respondent disinterest (Meade & Craig, 2012) correlates positively with carelessness.

Careless responding is not to be confused with other types of response bias such as response faking (also known as dissimilation; Nichols et al., 1989), malingering (Berry et al., 1992), or socially desirable responding (Paulhus, 2002). These types of response bias are characterized by a respondent's intention to systematically misrepresent their true score for a certain scale—which requires careful content-dependent responding—while careless respondents do not have this intention since their responses are content-independent (cf. Ward & Meade, 2023).

To briefly address terminology, the term *careless responding* has first been used in Haertzen and Hill (1963), with frequently used synonyms being *insufficient effort responding* (Huang et al., 2012), *content nonresponsitivity* (Nichols et al., 1989), *participant inattention* (Maniaci & Rogge, 2014), *inconsistent responding* (Greene, 1978), *protocol invalidity* (Johnson, 2005), and *random responding* (e.g., Beach, 1989, although Schroeders et al., 2022, criticize this term for being a misnomer since carelessness can also emerge in non-random patterns).

2.1 | Characterization of careless responding

Ward and Meade (2023) categorize three major ways in which carelessness manifests: *invariability*, *inconsistency*, and *fast responses*. Invariability is characterized by identical patterns of responses, for example, 1-2-3-1-2-3. In the most extreme case, this boils down to straightlining, that is, always giving the same response (also known as longstring; Johnson, 2005). Inconsistent careless responses *"do not match patterns that would be expected based on theoretical/logical grounds or trends in the data"* (Ward & Meade, 2023). Hence, such responses fail to meet an expected level of consistency. Often, inconsistent responding is characterized by choosing answer categories near-randomly, for instance near the scale endpoints (extreme responding; Bachman & O'Malley, 1984) or with equal probability across all categories. Fast responses are responses that have been given at such a speed that renders it arguably impossible that a respondent has *"read, understood, and responded accurately to the survey items"* (Ward & Meade, 2023). Indeed, impossibly fast responses have been found to be indicative of carelessness (e.g., Bowling et al., 2016, 2023; Huang et al., 2012; Meade & Craig, 2012; Wise & Gao, 2017). However, response times may only be indicative of careless responses if they are impossibly fast, since also attentive respondents may have relatively fast response times, thereby posing a risk of misclassification (Curran, 2016; Meade & Craig, 2012; Ulitzsch et al., 2022a).

Importantly, the line between attentive and careless responding can be blurry, in particular regarding inconsistent responses. For instance, so-called *misresponses* may occur in negatively-worded items if a participant inadvertently chooses an answer category opposite to their true beliefs due to superficial reading (e.g., Baumgartner et al., 2018). Another example is that participants may be attentive for the first few items but subsequently respond based on their overall impression of what the survey is measuring rather than the specific item content (e.g., Weijters et al., 2013).

2.2 | Prevalence and effects of careless responding

While careless responding is considered to be widely prevalent (Bowling et al., 2016; Curran, 2016; Meade & Craig, 2012; Ward et al., 2017; Ward & Meade, 2023; Ward & Pond, 2015), there does not seem to be a consensus about the level of carelessness prevalence. For instance, Curran (2016), Huang et al. (2012, 2015b), and Meade and Craig (2012) estimate the prevalence to be between 10% and 15% of survey participants, whereas others estimate it to be 3.5% (Johnson, 2005) or 46% (Oppenheimer et al., 2009). Either way, Ward and Meade (2023) conjecture that careless responding is likely present in all survey data. There is evidence that already a small proportion of careless respondents of 5%–10% can jeopardize the validity of research findings through a variety of psychometric issues (Arias et al., 2020; Credé, 2010; Schmitt & Stults, 1985; Woods, 2006), such as reduced scale reliability (Arias et al., 2020) and construct

validity (Kam & Meyer, 2015), attenuated factor loadings, improper factor structure, and deteriorated model fit in factor analyses (Arias et al., 2020; Huang, Bowling, et al., 2015; Woods, 2006), as well as inflated type I or type II errors in hypothesis testing (Arias et al., 2020; Huang, Liu, & Bowling, 2015; Maniaci & Rogge, 2014; McGrath et al., 2010; Woods, 2006). These psychometric issues are particularly relevant for social and personality psychology due to the widespread use of lengthy Likert-scale questionnaires, psychometric techniques such as reliability estimation, and statistical models such as factor models. For instance, fundamental instruments like the NEO PI-R (Costa & McCrae, 1992) comprise 240 items, although Bowling et al. (2021) estimate via a latent growth curve analysis that if a researcher wants a minimum of 90% of participants to respond carelessly to no more than 5% of all items, the maximum number of items they should include in an online survey may be as small as 79. Therefore, it is plausible that carelessness prevalence in such lengthy questionnaires exceeds the levels of 5%–10% that are deemed problematic. Consequently, empirical research using such lengthy instruments should very carefully screen the collected data for careless responding.

2.3 | Identifying careless responding

One may distinguish between *a-priori* and *post-hoc* methods for the identification of careless responding. A-priori methods are based on certain items that are included in the survey before administration. The rationale is that attentive participants respond in a very specific manner to such items, and deviating responses are suspected to be careless. Examples include self-report items (*"Did you respond accurately and truthfully to all questions?"*), instructed items (*"Choose the middle answer category"*), and bogus items (*"I am paid biweekly by leprechauns"*; Meade & Craig, 2012). Details on a-priori methods can be found in Meade and Craig (2012), while an evaluation and further discussion is provided in Curran and Hauser (2019).

A large number of post-hoc detection methods have been proposed in the literature, for instance consistency indicators such as psychometric synonyms and psychometric antonyms (Meade & Craig, 2012), longstring indices (Johnson, 2005), multivariate outlier analyses (e.g., Curran, 2016), or threshold values for response times (Bowling et al., 2023; Huang et al., 2012). Other detection methods rely on theoretical models, such as person-fit statistics that are based on item response theory (e.g., Drasgow et al., 1985; see Karabatsos, 2003, for an empirical comparison of person-fit statistics), structural equation models (e.g., Kim, Reise, & Bentler, 2018; Reise et al., 2016), or mixture models (e.g., Arias et al., 2020; Steinmann et al., 2022; Van Laar & Braeken, 2022). Specifically, Ulitzsch et al. (2022a) propose a mixture model based on item response theory that incorporates response times, and extend this approach in Ulitzsch et al. (2022b) to identify careless responding without requiring response times. The latter two methods are designed for identifying various careless response styles but are computationally very intensive.

Detailed overviews of common methods for the detection of careless responding, together with their strengths and weaknesses, are given in Arthur et al. (2021), Curran (2016), DeSimone et al. (2015), and Ward and Meade (2023).

3 | MACHINE LEARNING AS REMEDY TO CARELESS RESPONDING?

Machine learning is largely unexplored for the detection of careless responding, but its high potential for this purpose has been acknowledged in the literature. For instance, Arthur et al. (2021) envision that machine learning could detect careless responding in real-time in the foreseeable future.

In one of the few works available on the detection of careless responding via machine learning, Schroeders et al. (2022) study a gradient boosting approach (Friedman, 2002). As a supervised learning technique, gradient boosting requires a-priori labels regarding which respondents are careless and which are not. In their experiment, Schroeders et al. (2022) therefore instructed one part of the participants to respond carelessly and another to respond truthfully. They find that neither gradient boosting nor any of the considered traditional methods achieves satisfactory performance in distinguishing these two groups, as for instance at most 19% of the flagged careless respondents were in fact instructed to be careless. In their discussion, the authors express doubts whether participants in both groups complied with the instructions. This highlights one limitation of using an experimental manipulation to collect training data for

supervised learning techniques for the detection of careless respondents, as incorrectly labeled participants are detrimental to the performance of such algorithms (cf. Cannings et al., 2020). Ulitzsch et al. (2022a, 2022b) discuss further limitations of this experimental approach, most notably that it hinges on the assumption that respondents instructed to perform carelessly behave comparably to careless respondents in out-of-lab conditions. Indeed, it seems plausible that careless responding stems from an intrinsic state-of-mind rather than being a conscious action that can be instructed.

In a typical survey, it is not known a-priori which respondents are careless, so that the identification of careless respondents constitutes an unsupervised learning problem. As careless responding manifests in various distinct ways (see Section 2.1), different unsupervised learning techniques may be necessary to detect different types of careless responses. For instance, autoassociative neural networks (autoencoders; Kramer, 1992) were originally developed to separate signal from noise in electrical engineering and signal processing applications. Autoencoders compress the data to a lower-dimensional representation while preserving as much information as possible, so that random noise is filtered out in the compression. They may therefore be particularly suited to identify inconsistent careless responses, which Ward and Meade (2023) hypothesize to be the most prevalent form of careless responding.

The flexibility of machine learning methods is the key for their often excellent performance, but this flexibility comes at a cost. Machine learning approaches rely on hyperparameters, which need to be carefully selected by the researcher prior to training the algorithm, or whose optimal values are determined during training by searching over a pre-specified grid of candidate values. The use of machine learning methods without due care has therefore been criticized in the literature for yielding results that are not reproducible. An overview of relevant studies is provided by Kapoor and Narayanan (2022). While these studies investigate machine learning in other contexts, similar concerns apply when using machine learning techniques for the detection of careless responding. However, much of the literature on reproducibility of machine learning is focused on supervised learning. Kapoor and Narayanan (2022) trace reproducibility issues back to incorrect use of machine learning techniques, resulting in so-called *data leakage*. Most forms of data leakage identified by those authors stem from issues with the test set, for example, the training and test sets not being completely disjoint samples (which causes inflated estimates of prediction performance). Such concerns are not applicable in unsupervised learning. A form of data leakage that is also relevant for unsupervised learning is the use of variables that are not legitimate, for example, variables that are not available to the researcher at the time of conducting the analysis. For detecting careless responses, such issues can be avoided by using only the survey responses to train the algorithm, as well as legitimate auxiliary information like response times.

While data leakage is less of an issue, the flexibility of machine learning methods through the selection of hyperparameters remains a concern for unsupervised learning techniques (see also, e.g., Valtonen et al., 2024). It is well known that flexibility in data analysis can be abused by researchers to show whatever they want to show (Simmons et al., 2011). In line with, for example, Simmons et al. (2011) and Valtonen et al. (2024), the key to overcoming potential reproducibility issues is transparency. Researchers should be transparent in reporting, even if no careless responses are detected, and provide replication files that include all preprocessing for computational reproducibility. Regarding the former, researchers should clearly report which methods are used for detecting careless respondents, how those methods were trained, and what they did with the respondents that were detected as careless (cf. Valentine et al., 2021, for a discussion on the reporting of outliers). Some studies suggest the adoption of checklists (e.g. Kapoor & Narayanan, 2022; Mitchell et al., 2019; Mongan et al., 2020), in which authors are asked to address specific points or questions in order to ensure correct and reproducible use of machine learning. Regarding computational reproducibility, machine learning methods are well suited. Although some machine learning methods can be accessed via point-and-click interfaces in software such as SPSS (IBM Corp, 2022), most machine learning implementations require the use of code for programming languages such as R (R Core Team, 2022) or Python (Van Rossum and Drake, 2009). Sharing such code in the form replication files on platforms such as the Open Science Framework (https://osf.io/) or GitHub (https://github.com/) is minimal effort for researchers.

Finally, when applying machine learning for detecting careless responses, researchers should focus on methods that are appealing from a conceptual point of view, and they should give sufficient thought to the selection of hyperparameters. Specifically, we recommend researchers to consider whether certain hyperparameters can be selected based

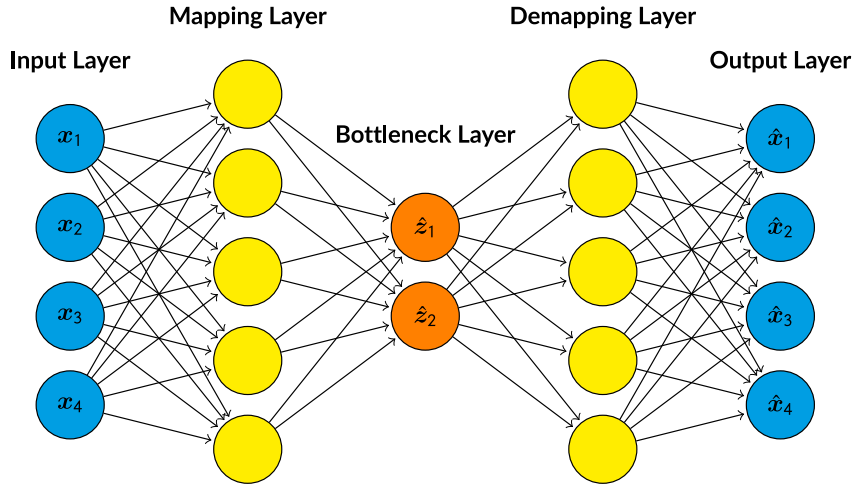

FIGURE 1 Illustrative example of an autoassociative neural network (autoencoder; Kramer, 1992) with input variables $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_4$. In such a network, the outcome variables are identical to the input variables. The network architecture consists of several layers, with each layer containing a certain number of nodes. In each node, a so-called activation function is applied to a linear combination of the nodes from the previous layer. The nodes in the bottleneck layer yield low-dimensional representations (here $\hat{\boldsymbol{z}}_1$ and $\hat{\boldsymbol{z}}_2$) of the data from the input layer, while the output layer provides reconstructions in the original dimension (here $\hat{\boldsymbol{x}}_1,\ldots,\hat{\boldsymbol{x}}_4$). The network is fitted by minimizing the reconstruction error according to a certain loss function.

on subject-matter knowledge rather than via computational procedures based on trial-and-error of several values. For instance, surveys in psychology typically collect information on various constructs, with the number of included scales being known to the survey designer (e.g., 30 facet scales in the revised NEO Personality Inventory; Costa & McCrae, 1992). Consider the aforementioned example of autoencoders (Kramer, 1992), which are illustrated in Figure 1. Autoencoders are conceptually appealing for multi-scale surveys, as they are based on a low-dimensional representation of the data and can be viewed as a non-linear generalization of principal component analysis (Kramer, 1991). The dimension of this low-dimensional representation is a hyperparameter of the autoencoder (i.e., the number of nodes in the bottleneck layer shown in Figure 1), which could be set equal to the number of scales included in the survey (cf. Welz & Alfons, 2023).

4 | PROOF-OF-CONCEPT

In order to gain insight into whether machine learning is a promising approach for identifying careless respondents, we conduct a small simulation experiment. The computations are performed in R version 4.2.2 (R Core Team, 2022) and Python version 3.8.10 (Van Rossum and Drake, 2009). Replication files are available from https://github.com/mwelz/OpenScienceML_Replication.

4.1 | Data generation

We simulate rating-scale data sets consisting of responses of n = 400 participants to p = 240 items, with each item providing 5 Likert-type answer categories (anchored by 1 = "strongly disagree" and 5 = "strongly agree"). This hypothetical survey measures q = 30 constructs, each of which is measured by a scale of eight items. The correlations between items within the same scale are randomly drawn from the interval [0.4, 0.6], while items from different scales are uncorrelated.[1] Based on these random correlation matrices, the Cronbach's α values of the scales range between 0.869 and 0.905 across the 1000 repetitions.[2] All eight items within a given scale follow the same response probability distribution. We consider three distinct types of distributions for the different scales: centered

TABLE 1 Response probability distributions of items within the corresponding scales.

Type of distribution	$P[X = 1]$	$P[X = 2]$	$P[X = 3]$	$P[X = 4]$	$P[X = 5]$
Centered	0.15	0.20	0.30	0.20	0.15
Skewed	0.10	0.15	0.20	0.25	0.30
Polarizing	0.30	0.175	0.05	0.175	0.30

Note: The random variable X denotes the response to an item with five Likert-type answer categories (1 = "strongly disagree", 2 = "disagree", 3 = "neither agree nor disagree", 4 = "agree", 5 = "strongly agree").

about the midpoint, skewed towards agreeing, and polarizing (likely to agree or to disagree). Table 1 lists the specific response probability distributions. Each of the three distributions is used in 10 scales, resulting in the aforementioned total of 3 × 10 × 8 = 240 items. We reverse half the items (four negatively-worded items per scale) and randomize the order of the items, with the same order being used for all participants.

Subsequently, we select a certain percentage of participants to be careless respondents. We refer to this percentage as the *carelessness prevalence*, which we set to 5%, 10%, 15%, 20%, 25%, and 30% of respondents, respectively. For each of the selected careless respondents, we replace the responses from a certain item onward by careless responses. This reflects the sentiment that careless respondents tend to be attentive at first but become careless at some point in the survey due to, for example, boredom or fatigue (Bowling et al., 2021; Galesic & Bosnjak, 2009; Gibson & Bowling, 2020; Ward & Meade, 2023). More precisely, for each careless respondent, we randomly select a carelessness onset item between the first and 192nd item (which corresponds to 80% of all items). From this onset item onward, we draw the responses with equal probability from the five answer categories. That is, we focus on a specific type of inconsistent careless responses (see Section 2.1).

4.2 | Methods

Among machine learning methods, autoencoders (Kramer, 1992) are conceptually appealing for the detection of inconsistent careless responding, as they are designed to filter out noise from the data (see Section 3). The network architecture of an autoencoder is illustrated in Figure 1. An autoencoder learns patterns in the data, such as consistent response patterns of attentive respondents. That is, an autoencoder compresses the observed responses to a latent low-dimensional representation and reconstructs the responses again via the learned patterns.[3] Responses that are reconstructed well indicate that the autoencoder has successfully learned the internal structure that underlies these responses. Conversely, responses that cannot be reconstructed well indicate that such responses are inherently different from the underlying structures learned by the autoencoder. This applies to inconsistent responding, for which internal structure is largely absent. Consequently, poor reconstruction performance is expected for inconsistent careless responding. To map the reconstructed responses into participant-level scores, we follow Hawkins et al. (2002) and compute the mean squared reconstruction error (MSRE) for each participant by averaging the squared reconstruction errors (scaled by the range of the answer categories) over the p items:

$$MSRE_i = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{x_{ij} - \hat{x}_{ij}}{L_j - 1}\right)^2, \qquad i = 1,\ldots,n,$$

where x_{ij} denotes the response of the ith participant to the jth item, $\hat{x}_{ij}$ the corresponding reconstruction from the autoencoder, and L_j the number of answer categories for the jth item. A large MSRE may be an indication that the participant engaged in inconsistent careless responding.

We compare the autoencoder to several benchmark methods that are recommended in the behavioral science literature (Arthur et al., 2021; Curran, 2016; DeSimone et al., 2015; Goldammer et al., 2020; Ward and Meade, 2023): Mahalanobis distances (Mahalanobis, 1936), personal reliability (Jackson, 1976; see Johnson, 2005, for a description), psychometric synonyms (Meade & Craig, 2012), intra-individual response variability (IRV; Dunn et al., 2018;

Marjanovic et al., 2015), and the l_z person-fit statistic for polytomous items (Drasgow et al., 1985).[4] These methods are comparable to the autoencoder in that they also compute certain scores for the respondents that are intended to reflect the level of carelessness, which we call *carelessness scores*.

Typically, an observation is considered a potential careless respondent if its carelessness score exceeds a certain threshold value. However, we evaluate the methods in a way that provides more complete insights into the intrinsic ability of methods to distinguish careless respondents from the remaining observations, irrespective of specific choices for the threshold values (cf. Curran & Denison, 2019, who argue that one should avoid that a method is favored simply because better threshold values have been developed). Intuitively, a method performs well if the most extreme carelessness scores are found predominantly in true careless respondents. This notion is operationalized in so-called *recall curves*, which visualize how many true careless respondents are recovered among the first k observations with the highest or lowest carelessness scores (depending on the method), for varying values of k. That is, the observations are first sorted according to the carelessness scores in descending order (autoencoder, Mahalanobis distances, IRV) or ascending order (personal reliability, psychometric synonyms, l_z person-fit). Then we compute how many careless respondents are included among the first k observations, and we divide by the total number m of careless respondents. To construct the recall curve, this is computed for all values of $k = 0, \ldots, n$. Recall curves are monotonically increasing and will for some k attain the maximum value of 1, as increasing k eventually exhausts the entire sample, which naturally includes all m careless respondents. The quicker a recall curve reaches a value close to 1, the better the corresponding method performs.

For a given data set, we apply the six considered methods and compute the respective recall curves. We repeat this procedure for the 1000 simulated data sets and report average recall curves across repetitions.

4.3 | Results

Figure 2 shows the recall curves of the six compared methods for various prevalence levels of careless responding. For comparison, a solid black reference line illustrates an ideal recall curve for hypothetical carelessness scores that rank the m true careless respondents before any other participants. The closer a method comes to that reference line, the better its detection performance. We stress that our findings are limited to the specific simulation design with only one type of inconsistent careless respondents.

The autoencoder yields the best performance with recall curves being quite close to the respective ideal recall curves, which indicates that it succeeds in assigning the highest carelessness scores predominantly to the careless respondents. Mahalanobis distances are the closest competitor, although there is a clear drop in performance, followed by the l_z person-fit and personal reliability. Psychometric synonyms perform rather poorly, with performance further deteriorating for increasing prevalence level. Even though the careless responses are generated by selecting response categories completely at random, that variability is not high enough to get picked up by IRV. Actually, the fact that the recall curves of IRV lie below a hypothetical diagonal line is an indication that IRV performs worse than randomly labeling respondents as careless.

4.4 | Additional simulations, discussion, and limitations

Since the design of our simulation experiment is somewhat stylized, we investigated variations of this baseline design, in which we (i) use various correlation structures between the items, (ii) draw the onset item for careless responding only from a later part in the survey, and (iii) generate different types of careless respondents. The simulation designs and results are described in detail in our GitHub repository containing the replication files: https://github.com/mwelz/OpenScienceML_Replication.

In the following summary of our findings, we focus on the autoencoder and its closest competitors Mahalanobis distances and the l_z person-fit. Across the investigated correlation settings, the autoencoder performs outstandingly,

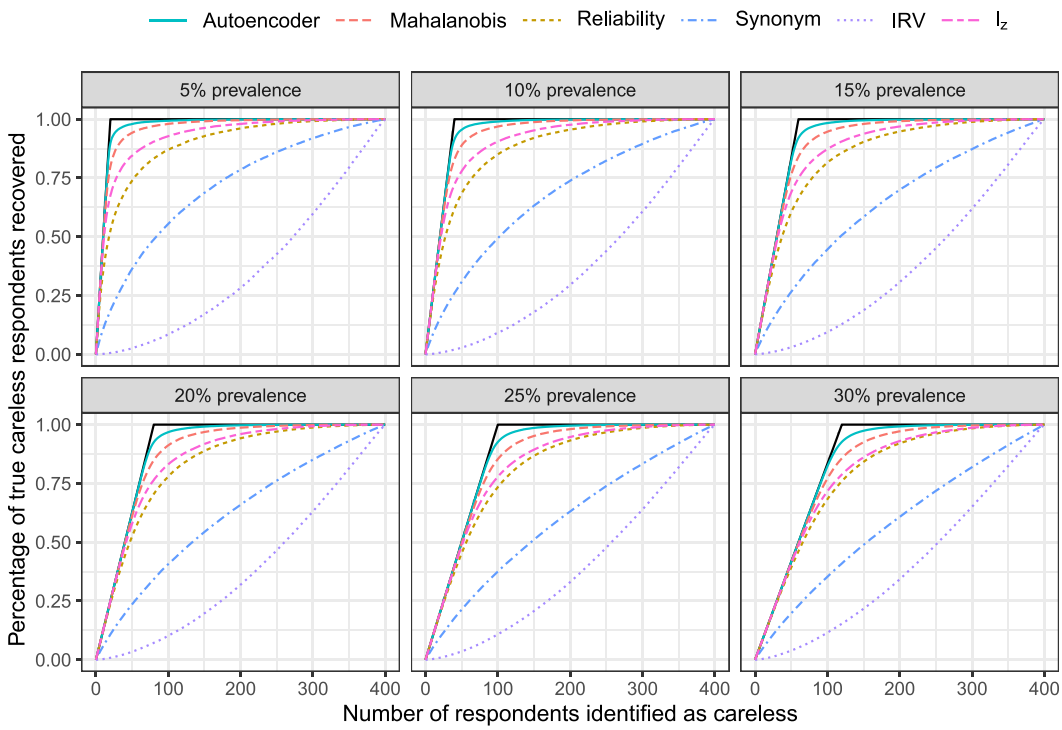

FIGURE 2 Recall curves of the six compared methods for various prevalence levels of careless responding in a simulated survey with n = 400 participants and p = 240 items, averaged across 1000 repetitions. A solid black reference line illustrates the best possible recall curve that can be achieved in each setting.

although it is not always the best method. In a setting with low correlations within the scales, the l_z person-fit performs best, but it is clearly outperformed by the autoencoder for moderate to high correlations within the scales. In a setting with nonzero correlations between scales, Mahalanobis distances perform best, but the gap to the autoencoder decreases with increasing carelessness prevalence. Also in other designs, the performance of Mahalanobis distances deteriorates with increasing prevalence, while the autoencoder remains stable across prevalence levels. In the variation of the design with a careless onset later in the survey, the autoencoder yields the best performance by a considerable margin. In the variation of the design with another type of inconsistent careless responding, the autoencoder performs near perfectly, as do Mahalanobis distances. As can be expected since the autoencoder is designed to filter out random noise from the data, it does not work well for the two investigated types of invariable careless respondents, but neither do Mahalanobis distances or the l_z person-fit.[5]

To summarize, the autoencoder outperforms the benchmark methods regarding the identification of inconsistent careless respondents, whereas the benchmark methods exhibit more variability in their performance across simulation designs. Moreover, the excellent performance of the autoencoder remains stable for relatively high prevalence of careless responding. This is a desirable property, as the level of prevalence may increase with survey length (cf. Bowling et al., 2021). Nevertheless, we emphasize that these results should not be mistaken for conclusive evidence that autoencoders are superior to existing methods, as the limited number of simulation designs does not suffice to draw general conclusions. We do not aim to make a specific methodological contribution about the use of autoencoders, rather we use the autoencoder as an example to highlight the potential of machine learning for the detection of careless responding. Furthermore, it is worth pointing out that the sample size of n = 400 is quite typical nowadays for surveys in psychology, and that our results indicate that machine learning methods may work well in such sample sizes, even for a relatively high ratio of items to observations (p = 240). Finally, a general limitation of our simulations is that it is unclear how our findings generalize to empirical settings, where data are more noisy and

careless responding is more fuzzy, making the latter harder to detect (cf. Meade & Craig, 2012). Accordingly, our simulation experiment should be viewed as a proof-of-concept that machine learning is a promising direction for further research on the identification of careless responding.

5 | WHERE DO WE GO FROM HERE?

In the previous section, we compared various methods for identifying careless respondents on simulated data. While such data have the attractive property that the researcher knows the ground truth, they are necessarily rather stylized. That is, in the context of careless responding, the researcher needs to make simplified assumptions on the data generating processes of attentive responses and careless responses. While one should try to be somewhat realistic in those assumptions, it is impossible to translate the complex and intricate nature of human responses into tractable code for a simulation experiment. In empirical data collected from surveys, on the other hand, the researcher simply does not know the ground truth as to which participants responded carelessly. This begs the question: How can we study methods for identifying careless respondents in a more realistic setting?

5.1 | Building an open data repository

The answer may be simple, but its implementation certainly is challenging: building an open data repository with data sets in which careless respondents are labeled. An example of a popular benchmark data base in empirical psychology is the Eugene-Springfield Community sample (ESCS; Goldberg, 2008), which is publicly available in the Harvard Dataverse (https://dataverse.harvard.edu/dataverse/ESCS-Data). The fact that the ESCS is publicly available and easily accessible has arguably contributed to it being a popular benchmark data base for testing novel psychometric methods. In the same spirit, we envision an easily accessible repository of benchmark data sets for testing methods for the identification of careless respondents. For common instruments such as the revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992), sharing data allows to establish specific benchmark data sets from different populations. Ideally, labels for careless respondents are then established through scientific consensus. That is, if researchers are transparent about the identification of careless respondents and share the relevant code, over time meta-analyses can be conducted, which together with manual inspection by subject-matter experts may lead to a consensus about which respondents can be considered careless in certain benchmark data sets.

Nevertheless, as scientific consensus is a lengthy process, we advocate for a more concerted approach. To emulate the process, many labs studies (Klein et al., 2014) may be fruitful for creating labels for careless respondents in specific data sets: many teams of researchers screen the same data set for careless respondents yielding many sets of labels, and a smaller team of subject-matter experts performs the meta-analysis and manual inspection to determine a final set of labels (e.g., to reduce likely false positives). For smaller teams of researchers, a multiverse analysis (Steegen et al., 2016) may be a suitable alternative for constructing labels in a similar fashion. Although the resulting labels should not be confused with the ground truth about which respondents are careless, we conjecture that labels of a high enough quality can be established to allow for the evaluation of newly-developed methods on realistic data sets in future research.

In addition to labeled benchmark data sets, this open data repository may also contain a collection of simulation protocols from which researchers can choose for further evaluation of methods. While simulation experiments permit researchers to know the true careless respondents but on stylized data, benchmark data sets would provide realistic data, but with some uncertainty left regarding the labels of careless respondents. Evaluating methods via simulation experiments and on benchmark data sets may therefore offer complementary insights in order to find the best performing methods. Similarly, recommendations for the values of hyperparameters (or suitable ranges of values) can be developed for specific machine learning methods through such a thorough evaluation. This reduces

researcher degrees of freedom in the application of machine learning, which may increase reproducibility (cf. Simmons et al., 2011).

5.2 | Supervised learning revisited

While benchmark data sets with high-quality labels allow researchers to evaluate methods for identifying careless respondents on realistic data, we argue that they offer another, far greater benefit: they make it possible to train supervised learning methods. As discussed in Section 3, a previous study on supervised learning in the context of careless responding by Schroeders et al. (2022) collected labeled training data via an experiment in which some participants were instructed to respond carelessly and others to respond truthfully. Besides relying on participants to follow instructions, such an experimental manipulation has the further drawback that it may influence the observed response behavior: it is doubtful whether participants who are instructed to respond carelessly behave in a way that resembles intrinsic careless responding behavior (Ulitzsch et al., 2022a, 2022b). Hence, such an experimental manipulation does not seem to be a suitable approach for obtaining labeled training data for supervised learning. On the contrary, careless responding behavior is not influenced when labels are created post-hoc as outlined in the previous section.

When trained on data with high-quality labels, supervised learning methods offer various advantages over unsupervised learning methods. First, in principle any supervised learning method can be trained to detect various types of careless responding, possibly all types that are present in the training data. This is not evident for unsupervised learning methods, and certain unsupervised methods may only be able to identify specific types of careless responding (cf. our results for the autoencoder from Section 4.4). Second, through the use of additional (expert) knowledge in the form of labels, supervised learning methods may require fewer observations in training than unsupervised learning methods to effectively distinguish between careless and attentive respondents.

As outlined in Section 5.1, subject-matter knowledge and manual inspection may play an important role in constructing labels of high quality, which is time consuming and may not always be feasible. Even without such a step to construct a final set of labels, the availability of various sets of labels from a many labs study (Klein et al., 2014) or a multiverse analysis (Steegen et al., 2016) offers other possibilities in that it allows to study supervised learning methods under uncertainty with respect to the labels. That is, a supervised learning method could be trained on different labels and the resulting predictions for classifying new observations could be aggregated across the different trained algorithms to accommodate label uncertainty.

Many supervised learning methods, such as gradient boosting (Friedman, 2002), random forests (Breiman, 2001), or basic forms of neural networks (e.g., Hastie et al., 2009, Chapter 11), require that all observations have information on the same input variables. Consequently, such methods need to be trained for specific surveys. For instance, if labeled benchmark data sets are available for the NEO PI-R instrument (Costa & McCrae, 1992), they can be used to train a supervised learning algorithm. This pretrained algorithm can then be made publicly available. If a researcher collects new data for the NEO PI-R instrument, they can easily apply the pretrained algorithm to identify careless respondents. Note that applying a pretrained algorithm to new data from a potentially different domain is known as *transfer learning* (e.g. Weiss et al., 2016), and that using a pretrained algorithm implies that the researcher no longer has any degrees of freedom regarding the hyperparameters. Furthermore, it avoids any reproducibility issues that stem from data leakage (see Section 3, as well as Kapoor & Narayanan, 2022).

This approach may be suitable for commonly used instruments such as NEO PI-R, but it has practical limitations given the vast number of instruments used in psychology. However, there exist advanced types of neural networks that can work with different input dimensions (e.g., Hochreiter & Schmidhuber, 1997; Vaswani et al., 2017), which have been highly successful, for example, in the field of natural language processing. One example is the popular text generation tool ChatGPT (OpenAI, 2023), which can take prompts of any length to generate the requested text. It may be possible to train such advanced neural networks on labeled benchmark data sets from a variety of

instruments, so that the pretrained algorithm can then be used to identify careless respondents in newly collected data from surveys of any number of items.

While the possibilities are enticing, there are limitations to using pretrained supervised learning algorithms that need to considered. As careless and attentive response behavior may differ between populations, labeled training data from various populations are necessary. Otherwise, the response behavior on the newly collected data (in which careless respondents should be identified) may be too different from the training data to achieve good performance. In the latter example of an advanced neural network that is trained on various surveys, further limitations apply. Response behavior may not only differ between populations, but they may also depend on context and survey characteristics such as length and item homogeneity (Ulitzsch et al., 2022b). Hence, the training data may need to come from a relatively large number of surveys to cover enough range of possible response behaviors.

Although these limitations imply that pretrained supervised learning algorithms for the identification of careless responding are likely still years away from practical use in empirical research, we argue that supervised learning is nonetheless an exciting playground for future methodological research. Pretrained supervised learning algorithms could be revolutionary for identifying careless respondents, but their potential can only be unlocked through an extensive repository of open benchmark data sets.

6 | "GOOD ENOUGH" PRACTICES FOR RESEARCHERS

Similar to general guidelines on data analysis (Simmons et al., 2011) and outlier handling (Valentine et al., 2021), transparency is key for avoiding pitfalls with respect to reproducibility, even if no careless respondents are identified. Accordingly, Table 2 provides a minimal checklist for "good enough" practices in using machine learning to identify careless respondents. Where possible, we recommend motivating choices regarding the specific machine learning methods and their hyperparameters by subject-matter knowledge. If the latter is not feasible, hyperparameter values may be set based on recommendations in the literature, but cross-validation (e.g., Hastie et al., 2009, Chapter 7) may be preferable.[6] If cross-validation is used, details such as the number of folds and the loss function should be reported. In addition, researchers should clearly describe the decision rules together with the threshold values that will be used to identify respondents as careless, and how identified careless respondents will be treated for subsequent data analysis. Ideally, these strategies are detailed during preregistration of the study (see Hardwicke & Wagenmakers, 2022,

TABLE 2 Checklist for "good enough" practices in machine learning for the identification of careless respondents.

Stage	Recommendation
Study planning	Choose methods that are conceptually appealing and report the motivation for your choices. *Example*: Autoencoders are well suited to detect inconsistent careless responding in surveys containing multiple scales, as they are designed to filter out noise by compressing the data to a low-dimensional representation.
Study planning	If possible, select values of hyperparameters by subject-matter knowledge and report your reasoning. Otherwise provide justification based on literature, or report the cross-validation strategy that will be used to determine the hyperparameter values. *Example*: The number of nodes in the bottleneck layer of the autoencoder corresponds to the dimension of the low-dimensional representation of the data, hence it is set to the number of scales in the survey.
Study planning	Report decision rules and threshold values for identifying careless respondents.
Study planning	Report how any identified careless respondents will be treated for further analysis.
Study planning	Preregister the study or submit it as a registered report.
Study execution	Report which (or how many) respondents are identified as careless, or if no careless respondents have been found.
Post-study	Share replication files and report which software versions have been used in the analysis.

for an introduction to preregistration) or, alternatively, in a registered report (see, e.g, Kiyonaga & Scimeca, 2019, for practical considerations on registered reports). After the analysis, researchers should report which (or how many) participants were identified as careless respondents, or if no careless respondents have been found. If space in the paper is limited, all of these details should be given, for example, in an online supplement.

Moreover, the use of machine learning methods should always go hand in hand with sharing data and code in the form of replication files. We thereby emphasize an old idea of Knuth (1984) with respect to writing code: *"Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do."* While researchers may be reluctant to share their code if they feel that the code is not efficient or pretty enough, we argue that such fears are unwarranted. As psychologists or data analysts, we are not expected to be expert programmers—the code just needs to get the job done. Sharing code in any form is always preferred to not sharing code at all. Rather than being concerned with efficiency or prettiness, researchers should focus their efforts on making the code more understandable for fellow researchers (or their future selves, if they need the code again after a number of months or years). We recommend interspersing code with plenty of comments that explain in easily understandable language *what* is done and, more importantly, *why* it is done. Furthermore, authors should state which software versions they used for the analysis, as software may change over time.

On a final note, as this paper is focused on the identification of careless respondents, we mostly refer to Arthur et al. (2021) for recommendations on how to handle respondents that are identified as careless. Nevertheless, we emphasize that binary decisions by the researcher on whether or not to remove identified careless respondents are bound to cause issues, not just regarding reproducibility but also with respect to the validity of the subsequent statistical analyses. In particular in social and personality psychology, removing careless respondents leads to a biased sample, as careless responding is linked to person characteristics and personality (see Ulitzsch et al., 2022a, for a detailed discussion and references). In the more general context of outliers, removing outliers has been shown to invalidate statistical theory, therefore distorting confidence intervals and inflating rejection rates in hypothesis testing (Chen & Bien, 2020; Karch, 2022). Alternatively, robust methods should be considered, which are designed to be influenced less by individual observations and to give reliable results even when outliers are present in the sample (see, e.g. Maronna et al., 2018, for a technical overview of robust methods). An outlook on robust methods in psychometrics is given in Mair (2018), while a specific example in the context of mediation analysis is provided by Alfons et al. (2022). However, while robust methods for general types of outliers are well studied, further research is necessary regarding their behavior in the presence of careless responding. Hence, a more detailed discussion on robust statistical methods is out of scope for this paper. Another alternative to binary decisions about excluding careless respondents is to use weighted two-step estimation approaches (e.g., Hong & Cheng, 2019; Ulitzsch et al., 2023b). In the first step, a weight is determined for each respondent such that a low weight reflects a higher confidence that the respondent is careless. In the second step, these weights are applied in subsequent analysis so that careless respondents contribute less to the estimates. For a general discussion on such procedures for downweighting careless respondents, we refer to Ulitzsch et al. (2023a, 2023b). A disadvantage of this approach is that standard errors and hypothesis tests in the second step do not incorporate the uncertainty from obtaining the weights in the first step. On the other hand, the weights in the first step can be based on carelessness scores from machine learning methods,[7] which has the advantage that no threshold value needs to be selected to determine which respondents are considered careless. Hence, such a weighted two-step estimation approach may still be appealing when detecting careless responding via machine learning. Nevertheless, we are not aware of any studies on using carelessness scores from machine learning methods in two-step downweighting procedures, hence further research is needed on this approach.

7 | DISCUSSION

Taking an open science perspective, we discussed the potential and the pitfalls of machine learning approaches for the identification of careless respondents. We analyzed possible causes of reproducibility issues, and we stressed the importance of transparency and open science practices as a remedy. Results from a proof-of-concept simulation

experiment indicate that this may be a fruitful direction for further research. Looking ahead, we outlined how building an open data repository of labeled benchmark data sets is instrumental in advancing this area of research. Finally, we provided a checklist for "good enough" practices in the application of machine learning for the detection of careless respondents, with a focus on how to avoid reproducibility issues.

While a large part of the literature on careless responding is focused on the identification of careless responding as a preprocessing step for further analysis of survey data (e.g., Arias et al., 2020; Arthur et al., 2021; Ward & Meade, 2023), careless responding is an interesting phenomenon to study in its own right. As such, machine learning methods for identifying careless responding can also be viewed as tools for exploratory data analysis rather than preprocessing tools. We believe that it is in this role that machine learning methods can lead to important contributions to the field of social and personality psychology. Accurate tools for identifying careless respondents open the door for building new theory and designing new experiments, for instance linking careless responding to personality traits (e.g. Bowling et al., 2016). In turn, a better understanding of the phenomenon of careless responding can help further improve methods for its identification.

Moreover, the existing literature is largely concerned with identifying which respondents are careless, and this paper is no exception. However, for instance Ward and Meade (2023) deem it rare that carelessness occurs for all items. As careless responding is likely linked to survey fatigue (e.g., Bowling et al., 2021), careless responding may occur only in a subset of items, most notably in items towards the end of the survey. For this reason, Welz and Alfons (2023) use machine learning to identify the onset of careless responding, that is, the point in a survey from which onward participants respond carelessly (if such a point exists). This approach can be developed further into a tool for monitoring in real-time whether participants become careless, which could be directly incorporated into survey software. For example, the corresponding participants could then be given an intervention in order to refocus their attention, which may improve the quality of the collected data. Detecting careless responding in real-time could furthermore help identify potential flaws in the survey design during pretesting. For instance, items at which careless responses accumulate might be improperly worded, or having many participants start responding carelessly from some point onward could indicate excessive survey length. Consequently, not only could data quality be improved, but also the quality of survey design by detecting and fixing potential design flaws. It follows that machine learning methods may not only serve as potentially powerful screeners for careless responding, but may also offer solutions to persisting problems with data quality in survey-based research in psychology.

Arthur et al. (2021) already envisaged that artificial intelligence and machine learning tools to detect careless responding in real-time could exist in the foreseeable future. In this paper, we draw up a roadmap for the required research. Crucially, it all hinges on open science practices: without transparency and open data, we may not get there.

ACKNOWLEDGMENTS

We are grateful to the participants of our session at the *Society for the Improvement of Psychological Science (SIPS) 2022 Meeting* for the fruitful discussion on the topic of careless responding. In addition, we thank the editor Sofia Persson and three anonymous reviewers for their constructive and helpful remarks. This work is supported by a grant of the Dutch Research Council (NWO), research program Vidi, project number VI.Vidi.195.141.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict interests.

ORCID

Andreas Alfons https://orcid.org/0000-0002-2513-3788
Max Welz https://orcid.org/0000-0003-2945-1860

ENDNOTES

[1] If the resulting matrix is not positive semidefinite, we find the nearest positive definite matrix via function nearPD() in the R package Matrix (Bates et al., 2023).

[2] We use function genOrdCat() from the R package simstudy (Goldfeld & Wujciak-Jens, 2022) to generate the data. With this function, supplied target values for the correlations only hold approximately on the population level, hence the same applies to the Cronbach's α values.

[3] Regarding the hyperparameters, we set the number of nodes in the mapping and demapping layers equal to $1.5 \times p$, where $p = 240$ is the number of items in the survey. The number of nodes in the bottleneck layer corresponds to the dimension of the latent low-dimensional representations and is therefore set to the number of scales $q = 30$ in the survey. Following Kramer (1992), we use nonlinear activation functions for the mapping and demapping layers, namely the hyperbolic tangent $\tanh(x) = 2/(1 + \exp(-2x))$, and the linear activation function $identity(x) = x$ in the bottleneck layer. For training the algorithm, we use the robust pseudo Huber loss function $loss(x) = \delta^2\left(\sqrt{1 + (x/\delta)^2} - 1\right)$ with $\delta = 1$, which is optimized via stochastic gradient descent (Goodfellow et al., 2016, Chapter 8) with learning rate 0.0001, batch size 10, and 100 epochs. We implement the autoencoder with the R package keras (Allaire & Chollet, 2022), which is an R interface to the Python library Keras (Chollet, 2022), which in turn is an interface to the TensorFlow library (Abadi et al., 2015).

[4] We use the R package careless (Yentes & Wilhelm, 2021) to compute the following benchmark methods: function mahad() for (squared) Mahalanobis distances, evenodd() for personal reliability, psychsyn() for psychometric synonyms, and irv() for intra-individual response variability. In addition, we compute the polytomous l_z person-fit statistic via function lzpoly() in the R package PerFit (Tendeiro, 2021). We use all functions with their default values, with two exceptions: (i) we set plot = FALSE in mahad() to suppress plotting the Mahalanobis distances, and (ii) if the default value of argument critval in psychsyn() results in fewer than 6 item pairs to be considered psychometric synonyms, we set the value so that at least 6 item pairs remain in consideration. This is motivated by Goldammer et al. (2020), who recommend using psychometric synonyms only with more than 5 item pairs of sufficient correlation. Furthermore, we do not consider psychometric antonyms, as these are conceptually equivalent to psychometric synonyms (Meade & Craig, 2012).

[5] Note that for invariable careless responding, low scores can be expected for the autoencoder, personal reliability, psychometric synonyms, IRV, and the l_z person-fit, while high scores can be expected for Mahalanobis distances. For computing the recall curves, the observations are sorted accordingly.

[6] Note that cross-validation is not only applicable to supervised learning methods. Despite the lack of a response variable or observed labels, hyperparameters of unsupervised learning methods can typically be selected via cross-validation with a suitable objective function. Consider the autoencoder from Section 4, where we use the reconstruction error as measured by the pseudo Huber loss function to train the algorithm. To select hyperparameters via cross-validation, the out-of-sample pseudo Huber reconstruction error across participants can be minimized.

[7] If necessary, the carelessness scores can be transformed such that they are positive and that lower scores are indicative of carelessness.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software version 2.9. https://www.tensorflow.org/

Alfons, A., Ateş, N., & Groenen, P. (2022). A robust bootstrap test for mediation analysis. *Organizational Research Methods*, *25*(3), 591–617. https://doi.org/10.1177/1094428121999096

Allaire, J., & Chollet, F. (2022). keras: R interface to 'Keras'. R package version 2.9.0. https://CRAN.R-project.org/package=keras

Arias, V. B., Garrido, L., 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*, *52*(6), 2489–2505. https://doi.org/10.3758/s13428-020-01401-8

Arthur, W., Hagen, E., & George, F. (2021). The lazy or dishonest respondent: Detection and prevention. *Annual Review of Organizational Psychology and Organizational Behavior*, *8*(1), 105–137. https://doi.org/10.1146/annurev-orgpsych-012420-055324

Bachman, J. G., & O'Malley, P. M. (1984). Yea-saying, nay-saying, and going to extremes: Black-white differences in response styles. *Public Opinion Quarterly*, *48*(2), 491–509. https://doi.org/10.1086/268845

Bates, D., Maechler, M., & Jagan, M. (2023). Matrix: Sparse and dense matrix classes and methods. R package version 1.5-4.1. https://CRAN.R-project.org/package=Matrix

Baumgartner, H., Weijters, B., & Pieters, R. (2018). Misresponse to survey questions: A conceptual framework and empirical test of the effects of reversals, negations, and polar opposite core concepts. *Journal of Marketing Research*, *55*(6), 869–883. https://doi.org/10.1177/0022243718811848

Beach, D. A. (1989). Identifying the random responder. *Journal of Psychology*, *123*(1), 101–103. https://doi.org/10.1080/00223980.1989.10542966

Berry, D. T., 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://psycnet.apa.org/doi/10.1037/1040-3590.4.3.340

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the questions ever end? Person-level increases in careless responding during questionnaire completion. *Organizational Research Methods*, *24*(4), 718–738. https://doi.org/10.1177/1094428120947794

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*(2), 218–229. https://psycnet.apa.org/doi/10.1037/pspp0000085

Bowling, N. A., Huang, J. L., Brower, C. K., & Bragg, C. B. (2023). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. *Organizational Research Methods*, *26*(2), 323–352. https://doi.org/10.1177/10944281211056520

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cannings, T. I., Fan, Y., & Samworth, R. J. (2020). Classification with imperfect training labels. *Biometrika*, *107*(2), 311–330. https://doi.org/10.1093/biomet/asaa011

Chen, S., & Bien, J. (2020). Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, *29*(2), 323–334. https://doi.org/10.1080/10618600.2019.1660180

Chollet, F. (2022). Keras. Python library version 2.9. https://keras.io

Chyung, S. Y. Y., Barkin, J. R., & Shamsy, J. A. (2018). Evidence-based survey design: The use of negatively worded items in surveys. *Performance Improvement*, *57*(3), 16–25. https://doi.org/10.1002/pfi.21749

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual*. 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*(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., & Denison, A. J. (2019) Creating carelessness: A comparative analysis of common techniques for the simulation of careless responder data. https://doi.org/10.31234/osf.io/ge6fa

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

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

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

Fokkema, M., & Strobl, C. (2020). Fitting prediction rule ensembles to psychological research data: An introduction and tutorial. *Psychological Methods*, *25*(5), 636–652. https://doi.org/10.1037/met0000256

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

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly*, *73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Gehlbach, H. (2015). Seven survey sins. *The Journal of Early Adolescence*, *35*(5–6), 883–897. https://doi.org/10.1177/0272431615578276

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

Goldammer, P., Annen, H., Stöckli, P. L., & Jonas, K. (2020). Careless responding in questionnaire measures: Detection, impact, and remedies. *The Leadership Quarterly*, *31*(4), 101384. https://doi.org/10.1016/j.leaqua.2020.101384

Goldberg, L. R. (2008). The Eugene-Springfield community sample: Information available from the research participants. *Technical Report*, *48*(1). Oregon Research Institute. https://ipip.ori.org/ORI_TechnicalReport_ESCS_Mar08.pdf

Goldfeld, K., & Wujciak-Jens, J. (2022). simstudy: Simulation of study data. R package version 0.5.0. https://CRAN.R-project.org/package=simstudy

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. http://www.deeplearningbook.org

Greene, R. L. (1978). An empirically derived MMPI carelessness scale. *Journal of Clinical Psychology*, *34*(2), 407–410. https://doi.org/10.1002/1097-4679(197804)34:2%3C407::AID-JCLP2270340231%3E3.0.CO;2-A

Haertzen, C. A., & Hill, H. R. (1963). Assessing subjective effects of drugs: An index of carelessness and confusion for use with the Addiction Research Center Inventory (ARCI). *Journal of Clinical Psychology*, *19*(4), 407–412. https://doi.org/10.1002/1097-4679(196310)19:4%3C407::AID-JCLP2270190410%3E3.0.CO;2-N

Hardwicke, T. E., & Wagenmakers, E.-J. (2022) Reducing bias, increasing transparency, and calibrating confidence with preregistration. https://doi.org/10.31222/osf.io/d7bcu

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer Series in Statistics. (2nd ed.). Springer.

Hawkins, S., He, H., Williams, G., & Baxter, R. (2002). Outlier detection using replicator neural networks. In Y. Kambayashi, W. Winiwarter, & M. Arikawa (Eds.), *Data warehousing and knowledge discovery* (pp. 170–180). Springer.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hong, M. R., & Cheng, Y. (2019). Robust maximum marginal likelihood (RMML) estimation for item response theory models. *Behavior Research Methods*, *51*(2), 573–588. https://doi.org/10.3758/s13428-018-1150-4

Huang, J. L., Bowling, N. A., Liu, M., & Li, Y. (2015). Detecting insufficient effort responding with an infrequency scale: Evaluating validity and participant reactions. *Journal of Business and Psychology*, *30*(2), 299–311. https://doi.org/10.1007/s10869-014-9357-6

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

IBM Corp. (2022). IBM SPSS statistics, version 29.0. https://www.ibm.com/products/spss-statistics

Jackson, D. N. (1976). The appraisal of personal reliability. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

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

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

Kapoor, S., & Narayanan, A. (2022) Leakage and the reproducibility crisis in ML-based science. https://doi.org/10.48550/arXiv.2207.07048

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

Karch, J. D. (2022) Outliers may not be automatically removed. https://doi.org/10.31234/osf.io/47ezg

Kim, D. S., Reise, S. P., & Bentler, P. M. (2018). Identifying aberrant data in structural equation models with IRLS-ADF. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 343–358. https://doi.org/10.1080/10705511.2017.1379881

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2018). Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys. *Social Science Computer Review*, *37*(2), 214–233. https://doi.org/10.1177/0894439317752406

Kiyonaga, A., & Scimeca, J. M. (2019). Practical considerations for navigating registered reports. *Trends in Neurosciences*, *42*(9), 568–572. https://doi.org/10.1016/j.tins.2019.07.003

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, v., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Kleinberg, B., van der Vegt, I., & Mozes, M. (2020). Measuring emotions in the COVID-19 real world worry dataset. In *Proceedings of the 1st workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics. https://aclanthology.org/2020.nlpcovid19-acl.11

Knuth, D. E. (1984). Literate programming. *The Computer Journal*, *27*(2), 97–111. https://doi.org/10.1093/comjnl/27.2.97

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, *37*(2), 233–243. https://doi.org/10.1002/aic.690370209

Kramer, M. A. (1992). Autoassociative neural networks. *Computers & Chemical Engineering*, *16*(4), 313–328. https://doi.org/10.1016/0098-1354(92)80051-A

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India*, *2*, 49–55. https://doi.org/10.1007/s13171-019-00164-5

Mair, P. (2018). *Modern psychometrics with R*. Springer.

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. Theory and Measurement in Personality and Individual Differences. https://doi.org/10.1016/j.paid.2014.08.021

Maronna, R. A., Martin, R. D., Yohai, V. J., & Salibián-Barrera, M. (2018). *Robust statistics: Theory and methods* (2nd ed.). John Wiley & Sons.

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

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., & Van der Laan, M. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31. https://doi.org/10.1126/science.1245317

Miotto, M., Rossberg, N., & Kleinberg, B. (2022). Who is GPT-3? An exploration of personality, values and demographics. In *Proceedings of the fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)* (pp. 218–227). Association for Computational Linguistics. https://aclanthology.org/2022.nlpcss-1.24

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19* (pp. 220–229). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287596

Mongan, J., Moy, L., & Kahn, C. E. J. (2020). Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, *2*, e200029. https://doi.org/10.1148/ryai.2020200029

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%3C239::AID-JCLP2270450210%3E3.0.CO;2-1

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, 6251. https://doi.org/10.1126/science.aac4716

OpenAI (2023). ChatGPT: Conversational AI by OpenAI. Retrieved August 17, 2023, from https://openai.com

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

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Paulhus, D. (2002). Socially desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Erlbaum.

R Core Team. (2022). *R: A Language and Environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? Application of iteratively reweighted least squares to the Rosenberg self-esteem scale. *Multivariate Behavioral Research*, *51*, 818–838. https://doi.org/10.1080/00273171.2016.1243461

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*(4), 367–373. https://doi.org/10.1177/014662168500900405

Schroeders, U., Schmidt, C., & Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. *Educational and Psychological Measurement*, *82*(1), 29–56. https://doi.org/10.1177/00131644211004708

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stantcheva, S. (2022). *How to run surveys: A guide to creating your own identifying variation and revealing the invisible*. Working Paper 30527. National Bureau of Economic Research. https://www.nber.org/papers/w30527

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Steinmann, I., Strietholt, R., & Braeken, J. (2022). A constrained factor mixture analysis model for consistent and inconsistent respondents to mixed-worded scales. *Psychological Methods*, *27*(4), 667–702. https://doi.org/10.1037/met0000392

Strobl, C., Kopf, J., & Zeileis, A. (2015). A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289–316. https://doi.org/10.1007/s11336-013-9388-3

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, *45*(1), 116–131. https://doi.org/10.1509/jmkr.45.1.116

Tendeiro, J. N. (2021). PerFit: Person fit. R package version 1.4.6. https://CRAN.R-project.org/package=PerFit

Uher, J. (2023). What's wrong with rating scales? Psychology's replication and confidence crisis cannot be solved without transparency in data generation. *Social and Personality Psychology Compass*, *17*(5), e12740. https://doi.org/10.1111/spc3.12740

Ulitzsch, E., Domingue, B. W., Kapoor, R., Kanopka, K., & Rios, J. A. (2023a). A probabilistic filtering approach to non-effortful responding. *Educational Measurement: Issues and Practice*, *42*(3), 50–64. https://doi.org/10.1111/emip.12567

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022a). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika*, *87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., Shin, H. J., & Lüdtke, O. (2023b). Accounting for careless and insufficient effort responding in large-scale survey data—development, evaluation, and application of a screen-time-based weighting procedure. *Behavior Research Methods*, 1–22. https://doi.org/10.3758/s13428-022-02053-6

Ulitzsch, E., Yildirim-Erbasli, S. N., Gorgun, G., & Bulut, O. (2022b). An explanatory mixture IRT model for careless and insufficient effort responding in self-report measures. *British Journal of Mathematical and Statistical Psychology*, *75*(3), 668–698. https://doi.org/10.1111/bmsp.12272

Valentine, K. D., Buchanan, E. M., Cunningham, A., Hopke, T., Wikowsky, A., & Wilson, H. (2021). Have psychologists increased reporting of outliers in response to the reproducibility crisis? *Social and Personality Psychology Compass*, *15*(5), e12591. https://doi.org/10.1111/spc3.12591

Valtonen, L., Mäkinen, S. J., & Kirjavainen, J. (2024). Advancing reproducibility and accountability of unsupervised machine learning in text mining: Importance of transparency in reporting preprocessing and algorithm selection. *Organizational Research Methods*, *27*(1), 88–113. https://doi.org/10.1177/10944281221124947

Van Laar, S., & Braeken, J. (2022). Random responders in the TIMSS 2015 student questionnaire: A threat to validity? *Journal of Educational Measurement*, *59*(4), 470–501. https://doi.org/10.1111/jedm.12317

Van Rossum, G., & Drake, F. L. (2009). Python 3 reference manual. CreateSpace. https://www.python.org/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30).

Ward, M., & Meade, A. W. (2023). Dealing with careless responding in survey data: Prevention, identification, and recommended best practices. *Annual Review of Psychology*, *74*(1), 577–596. https://doi.org/10.1146/annurev-psych-040422-045007

Ward, M., 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

Ward, M., & 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

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737–747. https://doi.org/10.1509/jmr.11.0368

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*(3), 320–334. https://doi.org/10.1037/a0032121

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 9. https://doi.org/10.1186/s40537-016-0043-6

Welz, M., & Alfons, A. (2023) I don't care anymore: Identifying the onset of careless responding. https://doi.org/10.48550/arXiv.2303.07167

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, *30*(4), 343–354. https://doi.org/10.1080/08957347.2017.1353992

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://psycnet.apa.org/doi/10.1007/s10862-005-9004-7

Yentes, R. D., & Wilhelm, F. (2021). careless: Procedures for computing indices of careless responding. R package version 1.2.1. https://CRAN.R-project.org/package=careless

AUTHOR BIOGRAPHIES

Andreas Alfons is an associate professor of statistics at Erasmus School of Economics, Erasmus University Rotterdam. He obtained his doctorate from Vienna University of Technology. His main area of research is the development of statistical methods that are robust against outliers and deviations from model assumptions. He is interested in computational statistics, machine learning, statistical software, psychometrics, as well as applications in the behavioral sciences. Recently, his research is focused on robust methods for rating-scale data and the identification of careless responding in surveys.

Max Welz is a PhD student in statistics and econometrics at Erasmus School of Economics, Erasmus University Rotterdam. His main area of research is developing statistical methods that are robust against violations of model

assumptions. Besides methodological work, he is interested in applications in behavioural sciences, medicine, and economics. His recent work is focused on robust methods for ordinal data obtained from surveys.

How to cite this article: Alfons, A., & Welz, M. (2024). Open science perspectives on machine learning for the identification of careless responding: A new hope or phantom menace? *Social and Personality Psychology Compass*, e12941. https://doi.org/10.1111/spc3.12941