# Can Autoencoders Replace Attention Checks to Detect Inattentive Survey Respondents?

*Completed Research Paper*

**Ilias Triantafyllopoulos**
New York University
ilias.triantafyllopoulos@nyu.edu

**Panos Ipeirotis**
New York University
panos@stern.nyu.edu

## Abstract

*Ensuring data quality is a persistent challenge in survey-based research, particularly with the rise of online participant pools prone to inattentiveness and random responding. Traditional quality control methods, such as attention checks and response pattern analyses, add cognitive load and are often domain-specific. In this paper, we explore the use of autoencoders - unsupervised neural networks that learn to reconstruct structured data - as a scalable, domain-agnostic alternative for detecting inattentive survey respondents. Autoencoders can effectively identify response patterns that deviate from typical behavior without requiring labeled data or explicit participant intervention. Across nine real-world survey datasets, our experiments demonstrate that autoencoders consistently improve over baseline predictors, achieving notable reconstruction ability (average Lift > 1.4) and strong inattentiveness detection performance (AUC up to 0.79). We further introduce a modified loss function tailored to survey structures and explore Percentile Loss to enhance detection in challenging cases. These results suggest that autoencoders offer a flexible and automated solution for improving behavioral data reliability, complementing traditional survey quality control techniques.*

**Keywords:** Autoencoders, Inattentiveness, Attention Checks, Behavioral Research, Surveys, Unsupervised Learning

## Introduction

Behavioral, social, and political scientists rely on surveys. This kind of research requires high data quality to be effective and produce valid outcomes. However, the high data quality is often violated due to the presence of random or inattentive responses, which can distort findings, add noise, and reduce statistical power. Researchers call this phenomenon *'content nonresponsivity,'* which is defined as a loss in the consistency of responses to the data items (Nichols et al., 1989). Respondents may provide random answers due to a lack of engagement, survey fatigue, or just unwillingness, thereby affecting the validity of research outcomes (Meade and Craig, 2012). This issue has been further exacerbated by the increasing reliance on online crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) and Prolific to find survey participants. Even though these platforms offer researchers a fast, simple, and cheaper way of collecting data from diverse participants, there are some concerns about the produced data quality. The workers are usually more likely to multitask during the studies (Goodman et al., 2013) and respond randomly to them. Studies show that inattentive and fraudulent responses are more common on online platforms, particularly when participants have the opportunity to cheat (Peer et al., 2021).

Current techniques of mitigating these issues involve attention checks (Berinsky et al., 2014), metrics of the individual responses' variation (Curran, 2016), and the Bayesian Truth Serum mechanism (Weaver and Prelec, 2013). These methods, while useful, introduce additional cognitive load, increase the survey completion time, and tend to be domain-specific. Given these limitations, there is a growing need for automated, data-driven approaches capable of identifying random responses without direct participant intervention.

These limitations can be significantly overcome by leveraging machine learning (ML) and artificial intelligence (AI) advancements. When collecting respondents' data for surveys, we cannot ensure who answered randomly or if they gave fabricated information, as we do not have a ground truth to compare against. Creating datasets that include this information would require a lot of costs and time, but even if we were able to provide these, they would include biases and inconsistencies, as definitions of inattentiveness can vary across studies. Consequently, unsupervised learning techniques offer a promising alternative. In this work, we utilize the autoencoders (Wang et al., 2016). Autoencoders are a type of neural network architecture designed to learn compact representations of data by reconstructing input patterns with minimal error. When applied to survey responses, an autoencoder learns the underlying distribution of typical responses, making it able to detect responses that significantly deviate from expected patterns as potential outliers. Unlike traditional methods, this approach does not require predefined rules or labeled training data, making it a scalable, adaptable, and domain-agnostic solution for detecting inattentive or random respondents.

Moreover, autoencoders can handle high-dimensional data efficiently, making them particularly useful when survey responses involve multiple variables. This is not rare if we think of how many different options usually accompany every question. Mapping the high-dimensional survey data into low-dimensional spaces preserves the structural and relational patterns inherent in the dataset. In this way, autoencoders are able to capture key characteristics and patterns of the respondents, while the inattentive users would be considered as noisy samples and difficult to reconstruct. As a result, valid responses tend to cluster naturally within the latent space, whereas inattentive or random responses appear dispersed, making them detectable as anomalies (Xu et al., 2018).

In this work, we explore the usage of autoencoders to detect random or inattentive responses in survey research. In particular, the central hypothesis is that the internal properties of this model can effectively distinguish random from valid responses by learning latent patterns in datasets and identifying outliers based on the reconstruction error. Unlike traditional quality control methods, which rely on attention checks or truth-inducing mechanisms, this approach operates without requiring participant engagement, predefined rules, or labeled data.

Our contribution is twofold. First, we review autoencoders as a general-purpose tool for spotting inattentive survey participants and introduce methodological refinements that adapt them specifically for categorical survey data. Second, we provide the first large-scale empirical validation of autoencoders for inattentiveness detection, systematically benchmarking performance across nine diverse real-world survey datasets.

# Related Work

### Autoencoders for Anomaly Detection

Autoencoders were originally introduced for unsupervised anomaly detection because they can reconstruct typical inputs while producing higher errors for unusual ones (Hawkins et al., 2002). This principle has been extended with variants such as Denoising Autoencoders (Vincent et al., 2010), Variational Autoencoders (Kingma & Welling, 2013), and Robust Deep Autoencoders (Zhou & Paffenroth, 2017), each aiming to improve robustness to noise or explicitly model outliers. These innovations established autoencoders as a flexible foundation for anomaly detection in diverse domains (Goodge et al., 2021; Rubio et al., 2020).

### Survey Inattentiveness Detection

Traditional approaches to inattentiveness rely on embedded design features such as attention checks, response time thresholds, or pattern-based indicators like straightlining (Kim et al., 2019). More advanced statistical and machine learning approaches include EM-based truth discovery (Dawid & Skene,

1979), unsupervised credibility modeling (Yin et al., 2008), and supervised classification using behavioral features (Schroeders et al., 2022; Ozaki, 2024). The Bayesian Truth Serum (Weaver & Prelec, 2013) represents another strand, incentivizing truthful responses through mechanism design.

**Autoencoders for Survey Inattentiveness**

Only recently have autoencoders been applied specifically to inattentive responding. Alfons & Welz (2024) introduced autoencoders for classifying inattentive responses, but their evaluation was limited to synthetic data. Welz & Alfons (2023) proposed CODERS, which combines autoencoder reconstruction errors with changepoint detection to identify when a respondent begins answering carelessly within a survey, demonstrating proof-of-concept in one real and several synthetic datasets. Our work differs in three key respects. First, we adapt autoencoders for categorical survey data, introducing variable-level loss weighting and exploring Percentile Loss (PL) to address the reconstruction–detection trade-off. Second, instead of focusing on within-survey onset, we target the identification of inattentive respondents as a whole. Third, we validate our approach on nine diverse, real-world survey datasets that include both attentive and inattentive participants, establishing large-scale empirical benchmarks. In doing so, we position autoencoders as a scalable and domain-agnostic tool for inattentiveness detection in survey practice.

## Data

We describe the datasets that we used for our study and give the main points of each dataset, along with how the data were collected and what attention checks were included. In Table 1, we summarize the statistics of all datasets, including the number of samples, the number of variables, the number of features, and the average number of features per variable.

We used https://datasetsearch.research.google.com/ to find publicly available survey datasets that include attention checks and use mainly structured/categorical responses (as opposed to textual or other forms of unstructured data). We identified nine datasets which differ substantially in topic, respondent population, and quality-control mechanisms to ensure that the evaluation reflects the method's robustness across diverse contexts. This heterogeneity spans (1) respondent types (such as adolescents (Robinson-Cimpian, 2014), MTurk workers (Moss et al., 2023), and nationally representative adult samples (Mastroianni & Dana, 2022)); (2) survey topics (ranging from political attitudes to misinformation susceptibility); and (3) attention check designs (ranging from none (Robinson-Cimpian, 2014) to multiple embedded checks (Pennycook et al., 2020)).

Beyond diversity, we applied two inclusion criteria: (a) datasets had to contain attention checks, and (b) they had to retain the responses of participants who failed these checks. These criteria were highly exclusive because most published works release only "cleaned" datasets with inattentive respondents removed, making large-scale evaluation of inattentiveness detection challenging.

| Dataset | Samples | Variables | Features | AFV |
|---|---|---|---|---|
| Robinson-Cimpian (2014) Mischievous Respondents | 14,765 | 98 | 619 | 6.32 |
| Pennycook et al. (2020) COVID-19 Misinformation | 853 | 188 | 708 | 3.75 |
| Condition 1 | 212 | 98 | 404 | 4.12 |
| Condition 2 | 206 | 98 | 358 | 3.65 |
| Condition 3 | 220 | 98 | 376 | 3.84 |
| Condition 4 | 215 | 98 | 400 | 4.08 |
| Alvarez et al. (2019) Inattentive | 2,725 | 39 | 196 | 5.03 |
| Uhalt (2020) Attention Checks and Response Quality | 308 | 60 | 337 | 5.62 |
| O'Grady et al. (2019) Moral Foundations | 355 | 72 | 322 | 4.47 |

| | | | | |
|---|---|---|---|---|
| Buchanan and Scofield (2018) Low-Quality Data | 1,038 | 23 | 159 | 6.91 |
| Moss et al. (2023) Ethical Data | 2,277 | 51 | 332 | 6.51 |
| Mastroianni and Dana (2022) Attitude Change | 1,036 | 51 | 322 | 6.31 |
| Ivanov et al. (2021) Racial Resentment | 860 | 67 | 310 | 4.63 |

**Table 1. Summary of Datasets used in this study. Every sample consists of a number of variables. Variables can be either questions or demographic records. Every variable consists of features in the way it is explained in the Method. The Average Number of Features per Variable (AFV) is also given in the last column.**

**(Robinson-Cimpian, 2014) Mischievous Respondents**: This dataset comes from the 2012 Dane County Youth Assessment (DCYA), an anonymous web-based survey of 14,765 U.S. high school students. The study investigated "mischievous responders", adolescents who intentionally gave extreme or implausible answers (e.g., exaggerated reports of health behaviors) that distort between-group disparity estimates across categories such as sexual orientation, gender identity, and disability. Outcomes examined included suicidal ideation, school belongingness, and substance use. Unlike other datasets in our study, this survey contained no embedded attention checks; inattentive cases were instead identified through domain knowledge and responses to unrelated questions. This makes the dataset distinctive, as it provides a large-scale setting where inattentiveness is inferred from patterns of extreme or inconsistent responding rather than explicit screening items.

**(Pennycook et al., 2020) COVID-19 Misinformation**: This study examined how attention and cognitive reflection affect the spread of COVID-19 misinformation. A total of 853 U.S. respondents evaluated 30 headlines (15 true, 15 false) presented in a social media format, with ground truth verified by fact-checking sources. The central aim was to test whether misinformation sharing occurs because people fail to consider accuracy rather than because they truly believe false claims. To measure this, the study introduced four experimental conditions that varied both the framing of the question (accuracy vs. sharing intention) and the ordering of response options. Data quality was assessed with multiple attention checks: two multiple-choice instruction items, one embedded Likert-scale item requiring selection of a specific value ("3"), and an additional self-report item asking whether participants had responded randomly. This combination of structured conditions and varied attention checks makes the dataset particularly useful for evaluating inattentiveness detection under different operational definitions.

**(Alvarez et al., 2019) Inattentive**: This survey collected 2,725 responses from California adults through Qualtrics' e-Rewards panel to study how inattentiveness affects political attitude measures. The study examined whether inattentive respondents introduce noise, satisficing, or misreporting, and whether they differ demographically from attentive participants. Data quality was monitored with three trap questions: two multiple-choice items requiring a specific answer and one open-text item requiring the word "government." Importantly, the full dataset includes both attentive and inattentive respondents, making it a strong benchmark for evaluating detection methods in a domain with coherent, multi-item political constructs.

**(Uhalt, 2020) Attention Checks and Response Quality**: This Qualtrics survey included 308 participants and focused on self-assessment of personality traits using Likert-scale items. To evaluate response quality, the instrument embedded multiple explicit attention checks that instructed participants to select specific options (e.g., "I see myself selecting 'Agree Strongly' if I'm paying attention to the survey"). These checks were distributed throughout the questionnaire to monitor engagement. The dataset provides a useful test case because it combines a structured personality scale, where attentive responses should show internal consistency, with clear ground-truth labels from embedded checks.

**(O'Grady et al., 2019) Moral Foundations**: This study collected 355 valid responses from U.S. undergraduate business students to examine how moral foundations predict prosocial behavior. Participants completed the Moral Foundations Questionnaire (MFQ), which measures individualizing

foundations (care, fairness) and binding foundations (loyalty, authority, purity). Data quality was monitored with a multiple-choice instructional manipulation check (IMC) embedded in the survey to ensure that respondents read instructions carefully. The dataset offers a structured attitudinal battery with a single attention check for identifying inattentiveness.

**(Buchanan and Scofield, 2018) Low-Quality Data**: This online Qualtrics survey collected 1,038 valid responses and examined methods for detecting low-quality data in psychological research. The study explored inattentive, low-effort, and automated responses (e.g., survey bots) using behavioral and statistical indicators such as response times, click counts, and distributional anomalies. Data quality was also monitored with an embedded attention check instructing participants to "Please mark strongly agree for this question." The dataset is distinctive in combining traditional survey responses with metadata-based quality indicators, making it useful for testing detection methods that rely solely on response patterns.

**(Moss et al., 2023) Ethical Data**: Study 1 surveyed 2,277 active U.S.-based MTurk workers to examine financial dependence on MTurk, time investment, and perceptions of fairness. The dataset is heterogeneous, covering multiple facets of workers' experiences. Data quality was monitored with two screens: one required selecting the correct summary of a prior item, while the other directly asked participants to indicate agreement with the statement "I am not reading the questions in this survey." This combination provides both indirect and self-report measures of inattentiveness.

**(Mastroianni and Dana, 2022) Attitude Change**: Study 1 collected 1,036 responses from a nationally representative U.S. adult sample via Prolific to examine perceptions of long-term societal attitude change. Participants compared their estimates of historical public opinion trends to actual polling data, with particular interest in whether people systematically overestimate liberalization. Data quality was ensured with an embedded attention check requiring respondents to type the number "1" in both blanks of a survey item. This dataset provides a high-quality, representative benchmark with a single IMC for inattentiveness.

**(Ivanov et al., 2021) Racial Resentment**: This MTurk survey collected 860 responses in April 2020 to study attitudes toward decarceration during COVID-19. The study examined how information about prison health risks, racial resentment, and empathy influenced support for release, and whether support varied by crime type, age, or health status of incarcerated individuals. Data quality was monitored with two embedded attention checks. The dataset offers a heterogeneous attitudinal instrument with multiple checks, useful for assessing inattentiveness in politically sensitive contexts.

## A Review of Autoencoders for Inattentiveness Detection

Autoencoders are a class of neural networks designed to learn compact representations of data in an unsupervised manner by encoding inputs into a lower-dimensional latent space and reconstructing them from this compressed representation. An autoencoder consists of two parts: an encoder, which learns an efficient representation of the data in a lower dimension, and a decoder, which learns how to create the initial data from this low-dimensional representation (Figure 1).

**Encoder:** The encoder function, denoted as $f_\theta$, maps an input response vector $x \in \mathbb{R}^d$ ($d$ is the number of features) to a lower-dimensional latent representation $z \in \mathbb{R}^m$ (with $m \ll d$):

$z = f_\theta(x) = \sigma(W_e x + b_e)$, where $W_e \in \mathbb{R}^{mxd}$ is the weight matrix, $b_e \in \mathbb{R}^m$ is the bias term, and $\sigma(\cdot)$ is a non-linear activation function such as ReLU or sigmoid.

**Decoder:** The decoder function, denoted as $g_\varphi$, reconstructs the original response vector from the latent representation $\hat{x} = g_\varphi(z) = \sigma(W_d z + b_d)$ , where $W_d \in \mathbb{R}^{dxm}$ is the decoder weight matrix, $b_d \in \mathbb{R}^d$ is the bias term, and $\hat{x} \in \mathbb{R}^d$ is the reconstructed response vector. The number of latent variables $m$, as well as the depth of the encoder and decoder, the number of units per layer, and the choice of activation functions are all hyperparameters that can be tuned for optimal performance. To prevent overfitting and

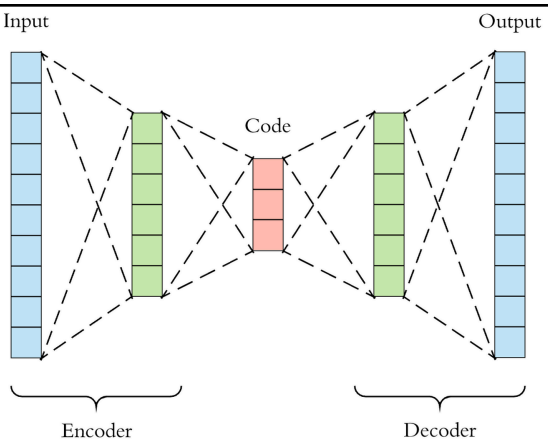


**Figure 1. Architecture of a simple Autoencoder. The autoencoder consists of two parts: (a) the Encoder, which encodes the information into latent variables, and (b) the Decoder, which decodes the information to the initial input. (Source)**

enhance generalization, L2 regularization (also known as weight decay) is applied to the weights of the network (Krogh and Hertz, 1991). The regularized objective function is given by $L_{reg} = L(x, \hat{x}) + \lambda \sum ||W||^2_2$ where λ is a regularization hyperparameter that controls the penalty on large weight values. Additionally, we use dropout in hidden layers to randomly deactivate neurons during training (Srivastava et al., 2014). During training, we drop each unit is with a probability $p_{drop}$: $h_{drop}^{\ l} = r \odot h^l, \quad r_i \sim Bernoulli(1 - p_{drop})$ where $h^l$ is the activation vector in the layer $l$, and $\odot$ denotes element-wise multiplication with the dropout mask $r$. $p_{drop}$ is also a hyperparameter to be tuned.

**Data Preparation**: The datasets we use in this study are surveys, where each participant provides answers to a series of usually multiple-choice questions. Since the majority of survey items follow a closed-form structure, the data is inherently *categorical*. Autoencoders, as well as all machine learning algorithms, function with numeric data, and thus, we transform the categorical variables into a numerical representation proper for the network. Each survey question (variable) is encoded using *one-hot encoding*, where each possible response choice is represented as a separate binary feature. Specifically, for a question with $k$ possible answer choices, a vector of length $k$ is created (or $k + 1$ if there exist users who passed the question), where only the selected response is marked as 1. At the same time, the remaining elements are set to 0. This transformation results in a feature space where the number of input dimensions is significantly larger than the number of original survey items, as each categorical variable expands into multiple binary features. The model treats each response as a set of features, learning the underlying distribution of the entire response set. We train the network so as to reconstruct the original input from its learned latent representation, capturing patterns across all answers. For further details on dataset statistics and preprocessing steps, refer to the Data Section.

**Loss Function**: For categorical survey responses, which is our case, the Binary Cross-Entropy (BCE) loss is used for Autoencoders training $BCE(x, \hat{x}) = -\sum_{i=1}^{d} [x_i \, log(\hat{x}_i) + (1 - x_i) \, log(1 - \hat{x}_i)]$. As explained earlier, our datasets consist of variables (questions) that each contain multiple binary features. Thus, we modify the standard BCE loss to take into consideration this "nested" structure. Instead of treating every response option independently, we compute the loss for each variable and normalize it by the logarithm of its number of features. Formally, our loss function is:

$$L = (1 / |\mathcal{V}|) \sum_{\{v \in \mathcal{V}\}} [1 / log(|F_v|)] \sum_{\{f \in F_v\}} BCE(x_f, \hat{x}_f)$$

where $\mathcal{V}$ is the set of all survey variables (questions), $F_v$ is the set of features (possible responses) for variable v. This variable-level weighting ensures that survey questions with more response options do not disproportionately dominate the total loss function.

**Randomness Detection**: As already explained, autoencoders are suitable models for capturing the internal correlations between features in the dataset and constructing a structured latent representation that preserves these dependencies. Given their ability to learn compact and meaningful representations of valid response patterns, we hypothesize that an autoencoder can effectively differentiate between attentive and inattentive respondents based on reconstruction loss. After training our model on each survey, we can then run an additional evaluation, using the same training data, and compute the reconstruction loss for each individual row (participant). Attentive users should create the same patterns in surveys, in contrast to the random responses, resulting in lower reconstruction losses. Therefore, we rank the reconstruction errors in descending order, identifying the respondents with the highest losses as potential inattentive participants. Our hypothesis is based on extensive prior research in anomaly and outlier detection, where autoencoders and similar deep learning architectures have been successfully employed to identify data points that deviate from expected patterns; in Computer Vision (An and Cho, 2015; Zhou and Paffenroth, 2017), time series (Xu et al., 2018), and tabular data (Eduardo et al., 2020). Bringing these examples to our use case, we predict that respondents who answered randomly will be flagged as outliers due to their poor reconstruction quality, favoring the broader application of autoencoder-based anomaly detection in behavioral data analysis. To the best of our knowledge, we are the first to apply this methodology to the analysis of inattentiveness in behavioral studies.

Our approach builds on the assumption that inattentive respondents lack consistent response patterns, producing randomness that the autoencoder cannot reconstruct. By contrast, minority groups with coherent but distinctive perspectives remain reconstructible because their internal consistency anchors them within the learned manifold. This distinction mirrors earlier dimensionality reduction techniques such as Principal Component Analysis (PCA), where structured minority patterns are captured by weaker components, whereas purely random responses cannot be represented. Thus, the autoencoder primarily flags incoherence rather than legitimate minority viewpoints, clarifying an important boundary condition of our method.

**Method**: For our experiments, we focus exclusively on categorical variables. However, we also take into consideration the numeric variables: if they have fewer than 20 distinct values, we treat them as categorical. Otherwise, we discretize them into predefined categories based on their standardized values. First, we apply *standard normalization* to each numeric variable $z = (x - \mu) / \sigma$ where x represents the raw value, μ is the mean, and σ is the standard deviation of the variable. Afterwards, we categorize the values into six discrete bins, as defined in Table 2.

| Category | Standardized Value Range |
|---|---|
| Bottom-extreme | $z < -1.4$ |
| Low | $-1.4 \leq z < -0.7$ |
| Normal | $-0.7 \leq z < 0.7$ |
| High | $0.7 \leq z < 1.4$ |
| Top-extreme | $z > 1.4$ |
| Missing Data | N/A |

**Table 2. Categorization of numeric variables based on standardized values.**

Admittedly, many questions were open-ended and required text as a response. We focus exclusively on structured categorical and categorized numerical variables, thereby excluding open-ended responses from our analysis. Datasets with a high proportion of such text fields were intentionally avoided to maintain consistency and not affect our findings.

For each dataset (survey), we separately perform a hyperparameter tuning phase before training the autoencoder. We employ *Bayesian Optimization* to efficiently explore the hyperparameter space. This optimization is conducted using the *KerasTuner* package within the Keras framework. Table 3 provides an overview of the hyperparameters explored during tuning, along with their respective search ranges. The optimization process evaluates various configurations over 30 trials, using a validation split of 20% and training each candidate model for up to 300 epochs, with an *Early Stopping* of 10 epochs. The best-performing configuration is selected based on validation loss minimization.

| **Hyperparameter** | **Values Explored** |
|---|---|
| Learning Rate | {0.0001, 0.001, 0.01} |
| Encoder Layers | {1, 2, 3} |
| Encoder Units | {64, 96, 128, 160, 192, 224, 256} |
| Encoder Activation[1] | {ReLU} |
| Encoder Regularization | {0.0, 0.001, 0.01} |
| Encoder Dropout | {0.0, 0.1, 0.2, 0.3, 0.4, 0.5} |
| Encoder Batch Normalization | {True} |
| Latent Space Dimensionality | {2, 3, …, 50} |
| Latent Activation | {ReLU} |
| Decoder Layers | {1, 2, 3} |
| Decoder Units | {64, 96, 128, 160, 192, 224, 256} |
| Decoder Activation | {ReLU} |
| Decoder Regularization (L2) | {0.0, 0.001, 0.01} |
| Decoder Dropout | {0.0, 0.1, 0.2, 0.3, 0.4, 0.5} |
| Decoder Batch Normalization | {True} |

**Table 3. Hyperparameter search space in our tuning experiments, using Bayesian Optimization. The Units, Activation, Regularization, Dropout, and Batch Normalization explored values were the same for each layer, regardless of how many layers were chosen each time for both the encoder and decoder.**

# Experimental Results

## *Metrics used for the Evaluation*

### Reconstruction Evaluation

First, we assess how accurately the model can reconstruct the original data. For each variable separately, we compute the accuracy, defined as the proportion of correctly reconstructed values across all samples. Given an original dataset $X \in \mathbb{R}^{nxd}$ and its reconstruction, $\widehat{X}$, the accuracy for a variable $v_i$ is defined as:

$$Accuracy(v_i) = (1 / N)\Sigma_{j=1}^{n} \mathbb{I}(x_{j,i} = \widehat{x_{j,i}})$$

[1] As a robustness check, we also extended the activation function search space beyond ReLU to include SELU, Swish, and GELU. The results were largely consistent with our main findings, with only minor, dataset-specific variations (e.g., modest improvements in Alvarez et al. (2019), slight decreases in Buchanan & Scofield (2018)).

where $x_{j,i}$ and $\widehat{x_{j,i}}$ represent the original and reconstructed values of the variable $v_i$ for the sample $j$, and $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the reconstruction is correct and 0 otherwise. To obtain a dataset performance measure, we compute the average accuracy across all $d$ variables:

$$Mean\ Accuracy\ =\ (1\ /\ d)\ \sum_{i=1}^{d} Accuracy(v_i)$$

To assess the performance of our model, we compare its accuracy to a baseline model that always predicts the majority class for each variable. The baseline accuracy for a variable $v_i$ is defined as:

$$Baseline\ Accuracy(v_i)\ =\ max_c\ (1\ /\ N)\ \sum_{j=1}^{n} \mathbb{I}(x_{j,i} =\ c),\qquad c \in \mathcal{C}_i,$$

where $\mathcal{C}_i$ is the set of unique categorical values for the variable $v_i$. The Mean Baseline Accuracy is defined similarly to Mean Accuracy. The Lift metric quantifies the improvement of our model over the baseline and is calculated as:

$$Lift\ =\ (1\ /\ d)\ \sum_{i=1}^{d}\ Accuracy(v_i)\ /\ Baseline\ Accuracy(v_i)$$

A Lift value greater than 1 indicates that the model performs better than the majority-class baseline, while a Lift value close to 1 suggests that the model offers little improvement over simple majority-based predictions. In addition to the accuracy-based evaluation, we leverage the One-vs-All (OVA) ROC AUC metric, denoted as ORA, to further assess the reconstruction quality per variable. Unlike traditional accuracy, which evaluates direct matches between predicted and actual values, ORA measures the model's ability to rank correct responses higher than incorrect ones across all possible categories.

For a given variable $v_i$ with categorical outcomes, we treat each possible category as a positive class, with all other categories as negative (One-vs-All approach). We then compute the ROC AUC score for each category and aggregate across all categories using a macro-average:

$$ORA(v_i)\ =\ (1\ /\ |\mathcal{C}_i|)\ \sum_{\{c\ \in\ \mathcal{C}_i\}} ROC\ AUC(c)$$

where $\mathcal{C}_i$ is the set of unique categorical values for the variable $v_i$, $ROC\ AUC(c)$ is the area under the ROC curve for class $c$, treating it as the positive class in a One-vs-All (OvA) manner.

To obtain the final ORA score, we compute the mean across all variables:

$$Mean\ ORA\ =\ (1\ /\ d)\ \sum_{i=1}^{d} ORA(v_i)$$

**Randomness Detection Evaluation**

Afterwards, we assess the model's ability to correctly detect inattentive users. We consider attention check responses as ground truth labels. We note that since some datasets contain multiple attention checks, we compute our metrics for each attention check separately.

After passing a dataset through the model, we obtain a reconstruction error for each sample. We then rank the errors in decreasing order, with the assumption that inattentive users should have higher errors. This allows us to frame the inattentiveness detection task as an information retrieval problem, where the goal is to rank inattentive users at the top. To evaluate how well the model ranks inattentive users, we compute recall at position h = total number of inattentive users:

$$Recall@k\ =\ |\{inattentive\ users\ in\ topk\}|\ /\ h$$

This metric tells us how well we perform if we knew a priori the number of inattentive users and selected exactly that many samples.

To assess precision at specific ranks, we compute:

$$Precision@k\ =\ |\{inattentive\ users\ in\ topk\}|\ /\ k$$

We report precision at $k$ = 10, 50, 100, which helps evaluate how well the model performs when selecting a fixed number of respondents. Note that $Precision@h\ =\ Recall@h$.

NDCG evaluates ranking quality while giving higher weight to correctly identified inattentive users appearing earlier in the ranking. The discounted cumulative gain at rank $h$ is:

$DCG@h = \sum_{i=1}^{h} \mathbb{I}(inattentive\ at\ rank\ i) / log_2(i + 1)$

We normalize by the ideal DCG (best possible ranking):

$NDCG@h = DCG@h / IDCG@h$

where IDCG@h is computed by sorting inattentive users in perfect ranking order.

Finally, we evaluate our model's ability to discriminate between attentive and inattentive users using the ROC curve. Here, reconstruction error is used as the decision threshold, and we compute:

$True\ Positive\ Rate\ (TPR) = True\ Positives / (True\ Positives + False\ Negatives)$

$False\ Positive\ Rate\ (FPR) = False\ Positives / (False\ Positives + True\ Negatives)$

By plotting TPR vs FPR across different thresholds, we obtain the ROC curve and compute AUC (Area Under Curve), which measures the model's ranking ability:

$AUC = \int_0^1 TPR(t)\, dFPR(t)$

Higher AUC values indicate that the model effectively distinguishes inattentive from attentive users.

## *Results*

### Reconstruction Performance

Reconstruction performance provides a measure of how accurately the model captures the typical patterns of the data. High reconstruction accuracy indicates that the autoencoder has successfully learned the underlying structure of the responses. We focus on whether the model meaningfully improves over a simple baseline predictor — one that always predicts the majority response for each survey question — by computing Lift scores. The reconstruction performance across all datasets is presented in Table 4. For the Pennycook et al. (2020) dataset, we also report results for each condition separately. Overall, we observe that the model achieves an accuracy of over 70% in all datasets except for the O'Grady et al. (2019) dataset, which records the lowest performance. Among all datasets, the highest Lift score is achieved in Uhalt (2020), where the model significantly outperforms a simple majority-class predictor in this dataset. While the model consistently improves over the baseline predictor, there is variability in Lift scores across datasets. The Alvarez et al. (2019) dataset also exhibits a high Lift score. This dataset is constructed with the goal of investigating inattentiveness, a task very close to ours. Conversely, O'Grady et al. (2019) dataset achieves the lowest Lift score, suggesting that its data structure presents challenges for the model. Potential factors affecting performance include feature sparsity, where datasets with a larger number of features may exhibit higher sparsity, making reconstruction more difficult, and sample size, since smaller datasets may lead to poorer generalizations.

| Dataset | Accuracy↑ | Baseline Acc | Lift↑ | ORA↑ |
|---|---|---|---|---|
| Robinson-Cimpian (2014) | 79.66 | 60.19 | 1.44 | 0.69 |
| Pennycook et al. (2020) | 76.71 | 66.22 | 1.15 | 0.71 |
| Condition 1 | 78.14 | 57.15 | 1.50 | 0.74 |
| Condition 2 | 76.49 | 56.51 | 1.46 | 0.74 |
| Condition 3 | 75.72 | 53.48 | 1.49 | 0.74 |
| Condition 4 | 75.99 | 52.75 | 1.53 | 0.75 |
| Alvarez et al. (2019) | 86.57 | 52.56 | 1.94 | 0.84 |
| Uhalt (2020) | 71.95 | 35.66 | 2.08 | 0.75 |
| O'Grady et al. (2019) | 63.52 | 58.78 | 1.13 | 0.53 |
| Buchanan and Scofield (2018) | 86.33 | 50.87 | 1.95 | 0.80 |
| Moss et al. (2023) | 77.31 | 59.78 | 1.37 | 0.68 |
| Mastroianni and Dana (2022) | 70.87 | 61.64 | 1.15 | 0.60 |

| | | | | |
|---|---|---|---|---|
| Ivanov et al. (2021) | 68.50 | 45.32 | 1.65 | 0.71 |

**Table 4. Evaluation of model performance across datasets when we try to "predict" an out-of-sample attribute value. Acc refers to the mean accuracy of the model in reconstructing the original data. Baseline Acc (Baseline Accuracy) represents the accuracy if we "guess" the majority class as the value for each attribute. Lift is computed as the ratio of Accuracy to Baseline Acc, indicating the improvement of our model over a naive predictor. ORA denotes the One-Vs-All ROC AUC metric, with 0.5 as the baseline performance.**

**Randomness Detection Performance**

Table 5 presents the performance of our model in detecting inattentive respondents across all datasets. For datasets containing multiple attention checks, we report results for each check separately, as well as for the union and intersection of all checks. In the union case, a respondent is considered inattentive if they fail at least one attention check, whereas in the intersection case, only those who fail all attention checks are classified as inattentive. For the case of Robinson-Cimpian (2014), where the survey lacks attention checks, we develop specific criteria based on the paper. These criteria include users who gave extreme answers to questions unrelated to the survey, e.g., eating carrots, salad, and fruits over 4 times per week each.

The AUC provides a robust measure of how well the reconstruction error differentiates inattentive from attentive respondents. We can observe that the highest AUC values are achieved in Uhalt (2020) (0.70), Alvarez et al. (2019) (0.79), and some attention checks in Pennycook et al. (2020), such as the third when considering all conditions (0.78), indicating strong discriminative power in these datasets. The lowest AUC values appear in Buchanan and Scofield (2018) (0.60) and Moss et al. (2023) (0.51–0.54), suggesting that inattentive responses in these datasets are harder to distinguish from attentive ones.

The union condition consistently yields higher recall values across datasets, as it captures a larger range of inattentive respondents. The intersection condition, while more restrictive, results in higher AUC values in some datasets, e.g., Pennycook et al. (2020), indicating that respondents failing all attention checks are stronger cases of inattentiveness.

While precision at different cutoffs (P@10, P@50, P@100) provides insights into how well the model ranks inattentive users at specific points, its utility depends on practical deployment scenarios. For instance, in real-world applications where manual validation of flagged respondents is feasible, high P@10 or P@50 is desirable, ensuring that the top-ranked inattentive cases are indeed errors. We can easily observe that the precision metrics are higher when the total number of errors is also high.

In the Robinson-Cimpian (2014) dataset, we observe notably low precision scores across all cutoff thresholds, despite a relatively high AUC (0.74). This discrepancy is primarily attributed to the small number of inattentive users (h = 230) relative to the total sample size (14,765), which can hurt precision scores.

For Pennycook et al. (2020), results demonstrate that aggregating all conditions yields substantially stronger performance metrics than analyzing each condition separately. Notably, Attention Check 3 consistently outperforms the others in terms of AUC, suggesting it is particularly effective at distinguishing inattentive respondents. We hypothesize that this is due to the nature of the check: embedded within a Likert-scale battery, it instructs participants to select a specific response ("neither agree nor disagree"), which attentive users are more likely to notice and comply with. In contrast, Attention Checks 1 and 2 involve preference-based questions that prompt participants to override their natural choice by selecting a predetermined option. Because these items' instructions are information-rich and potentially engaging, even attentive users may inadvertently fail them. Attention Check 4, which asks users if they responded randomly at any point during the survey, is highly generic and may capture a broader range of behaviors, including momentary lapses in attention or fatigue, even among otherwise attentive participants. This may explain its moderate performance and the overall variability across individual checks. The Alvarez et al. (2019) dataset yields the highest overall performance across both AUC and precision metrics. Upon closer examination, we observe that inattentive participants in this

dataset exhibit distinct response patterns, including widespread item nonresponse, making them easier to identify through reconstruction-based approaches.

In contrast, the Moss et al. (2023) dataset exhibits the weakest performance, with AUC scores close to random classification (≈0.51–0.54). An inspection of the training dynamics reveals that, despite early stopping being enabled, the autoencoder reaches a very low reconstruction loss that it maintains across many epochs, suggesting overfitting to patterns that are not informative for detecting inattentiveness. To address this limitation, we explore Percentile Loss (PL) (Merrill and Olson, 2020), which was originally proposed in the computer vision domain, where the goal was to model the common background rather than rare anomalies. The rationale behind PL is that it is statistically improbable for the 95th-percentile loss in a batch to correspond to an anomaly, assuming anomalies are rare. Accordingly, PL focuses learning on the lowest-error subset of a batch, thereby discouraging the model from overfitting to anomalous inputs. We adapt this approach to our context by calculating the reconstruction loss over a specified percentile of samples during training. When applying the 80th percentile loss, Moss et al. (2023) performance significantly improves, with AUC scores rising to 0.70–0.80 depending on the attention check configuration. This improvement, however, comes at the expense of reconstruction accuracy: the model's Lift score decreases by approximately 11%. This trade-off is expected, as the model learns to ignore some of the more irregular patterns (including those of inattentive users), thus improving its ability to identify inattentiveness through deviation. However, this improvement is not universal. For example, in Alvarez et al. (2019), the application of PL reduces AUC from 0.79 to 0.76, with only a marginal decrease in Lift (about 1%). This variability across datasets underscores the need for further investigation into the reconstruction–detection trade-off and the development of more adaptive strategies for selecting optimal percentile thresholds across varying attentional profiles.

Exploratory analysis across the nine datasets suggests that performance variation is driven less by size or sparsity and more by survey structure and label quality. Specifically, correlations indicate that Lift improvement over baseline relates moderately to AUC (Pearson r≈0.41), whereas sample size, number of variables, number of features, and AFV show near-zero associations (e.g., AFV r≈−0.10). Surveys with coherent, multi-item constructs (e.g., political attitudes, personality) yield stronger performance (AUCs ≈0.70–0.79), while heterogeneous instruments show weaker separability (≈0.51–0.54). Similarly, embedded instruction checks align closely with reconstruction-based detection (AUCs up to 0.84), whereas self-reports or single IMCs produce noisier labels (AUCs ≈0.60). This suggests that autoencoders are most effective for structured surveys with redundant item batteries, while heterogeneous instruments may require robust training strategies such as Percentile Loss.

| **Dataset** | **h** | **R@h↑** | **P@10↑** | **P@50↑** | **P@100 ↑** | **NDCG @h↑** | **AUC↑** |
|---|---|---|---|---|---|---|---|
| Robinson-Cimpian (2014) | 230 | 0.07 | 0 | 0.10 | 0.07 | 0.07 | 0.74 |
| Pennycook et al. (2020) | | | | | | | |
| All Conditions | | | | | | | |
| Attention 1 | 636 | 0.78 | 1.00 | 0.90 | 0.90 | 0.80 | 0.62 |
| Attention 2 | 236 | 0.38 | 0.70 | 0.56 | 0.45 | 0.42 | 0.58 |
| Attention 3 | 68 | 0.32 | 0.50 | 0.40 | 0.28 | 0.36 | 0.78 |
| Attention 4 | 172 | 0.38 | 0.70 | 0.50 | 0.42 | 0.43 | 0.62 |
| Union | 686 | 0.83 | 1.00 | 0.94 | 0.93 | 0.84 | 0.61 |
| Intersection | 28 | 0.25 | 0.30 | 0.26 | 0.17 | 0.29 | 0.84 |
| Condition 1 | | | | | | | |
| Attention 1 | 158 | 0.73 | 0.80 | 0.76 | 0.76 | 0.76 | 0.51 |
| Attention 2 | 56 | 0.27 | 0.30 | 0.24 | 0.27 | 0.24 | 0.52 |
| Attention 3 | 13 | 0.15 | 0.10 | 0.12 | 0.11 | 0.24 | 0.70 |
| Attention 4 | 44 | 0.30 | 0.40 | 0.26 | 0.22 | 0.32 | 0.53 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Union | 169 | 0.79 | 1.00 | 0.80 | 0.82 | 0.81 | 0.53 |
| Intersection | 8 | 0.13 | 0.10 | 0.06 | 0.06 | 0.25 | 0.65 |
| Condition 2 | | | | | | | |
| Attention 1 | 152 | 0.73 | 0.90 | 0.74 | 0.75 | 0.75 | 0.50 |
| Attention 2 | 63 | 0.41 | 0.50 | 0.44 | 0.38 | 0.46 | 0.61 |
| Attention 3 | 20 | 0.25 | 0.20 | 0.18 | 0.14 | 0.20 | 0.63 |
| Attention 4 | 42 | 0.29 | 0.40 | 0.26 | 0.22 | 0.37 | 0.54 |
| Union | 165 | 0.81 | 0.90 | 0.88 | 0.81 | 0.83 | 0.55 |
| Intersection | 6 | 0 | 0.10 | 0.04 | 0.03 | 0 | 0.59 |
| Condition 3 | | | | | | | |
| Attention 1 | 159 | 0.72 | 0.80 | 0.80 | 0.76 | 0.72 | 0.54 |
| Attention 2 | 54 | 0.35 | 0.50 | 0.36 | 0.30 | 0.40 | 0.60 |
| Attention 3 | 15 | 0.13 | 0.10 | 0.06 | 0.08 | 0.10 | 0.54 |
| Attention 4 | 41 | 0.34 | 0.50 | 0.36 | 0.26 | 0.43 | 0.65 |
| Union | 173 | 0.79 | 1.00 | 0.88 | 0.81 | 0.82 | 0.55 |
| Intersection | 4 | 0.25 | 0.10 | 0.02 | 0.02 | 0.17 | 0.57 |
| Condition 4 | | | | | | | |
| Attention 1 | 167 | 0.78 | 0.90 | 0.78 | 0.78 | 0.80 | 0.50 |
| Attention 2 | 63 | 0.32 | 0.20 | 0.30 | 0.31 | 0.28 | 0.51 |
| Attention 3 | 20 | 0.25 | 0.40 | 0.18 | 0.14 | 0.29 | 0.68 |
| Attention 4 | 45 | 0.20 | 0.40 | 0.18 | 0.21 | 0.24 | 0.52 |
| Union | 179 | 0.84 | 0.90 | 0.86 | 0.84 | 0.86 | 0.52 |
| Intersection | 10 | 0.30 | 0.30 | 0.06 | 0.06 | 0.33 | 0.57 |
| Alvarez et al. (2019) | 975 | 0.68 | 1.00 | 1.00 | 1.00 | 0.72 | 0.79 |
| Uhalt (2020) | 6 | 0.33 | 0.20 | 0.04 | 0.05 | 0.23 | 0.70 |
| O'Grady et al. (2019) | 20 | 0.52 | 0.10 | 0.12 | 0.08 | 0.13 | 0.63 |
| Buchanan and Scofield (2018) | 59 | 0.10 | 0.10 | 0.10 | 0.10 | 0.09 | 0.60 |
| Moss et al. (2023) | | | | | | | |
| Attention 1 | 161 | 0.54 | 0.30 | 0.14 | 0.17 | 0.15 | 0.53 |
| Attention 2 | 140 | 0.52 | 0.50 | 0.14 | 0.12 | 0.14 | 0.54 |
| Union | 248 | 0.54 | 0.50 | 0.22 | 0.24 | 0.21 | 0.54 |
| Intersection | 53 | 0.06 | 0.30 | 0.06 | 0.05 | 0.07 | 0.51 |
| Mastroianni and Dana (2022) | 60 | 0.47 | 1.00 | 0.56 | 0.29 | 0.60 | 0.65 |
| Ivanov et al. (2021) | | | | | | | |
| Attention 1 | 165 | 0.30 | 0.80 | 0.46 | 0.36 | 0.37 | 0.63 |
| Attention 2 | 55 | 0.09 | 0.20 | 0.08 | 0.09 | 0.11 | 0.60 |
| Union | 173 | 0.32 | 0.80 | 0.48 | 0.39 | 0.39 | 0.64 |
| Intersection | 47 | 0.06 | 0.20 | 0.06 | 0.06 | 0.09 | 0.58 |

**Table 5. Evaluation of inattentiveness detection across datasets. h represents the number of inattentive users identified as ground truth. R@h denotes Recall@h. At perfect ranking, where all inattentive users have a higher error than all attentive, R@h = 1. P@k denotes Precision@k. A separate evaluation is provided for the datasets where we had more attention checks. Union means we consider a sample as inattentive only when it failed in one of the attention checks. Intersection means we consider a sample as inattentive only when it fails in all attention checks.**

## Conclusions

In this study, we proposed and evaluated empirically the use of autoencoders as an unsupervised, general-purpose approach for detecting inattentive survey respondents, offering a scalable alternative to traditional attention checks and rule-based quality controls. By leveraging the model's ability to reconstruct structured survey responses, we detect inattentiveness through deviations in reconstruction quality without requiring explicit participant intervention, predefined screening rules, or labeled training data. Across nine real-world survey datasets, our results demonstrate that autoencoders can reliably capture latent response patterns and distinguish inattentive respondents with notable performance gains over baseline predictors. We further assume that for different use cases, the methodology facilitates flexibility to adapt correspondingly. More specifically, if, for example, we care for strong deviations from the underlying patterns, we should train the autoencoders for a longer time without restricting them. As shown in our empirical findings, this would lead the model to capture a higher percentage of patterns, moving the reconstruction–detection trade-off toward the reconstruction direction. Our metrics also provide a use case dependency. If we know a priori that a specific number of participants were inattentive, then we can rely on the Recall at a cutoff of this known number to conclude on how accurate we could be if we were to exclude them. On the other hand, if we want to conclude on how accurate we could be by excluding the number of participants we want, without a priori knowledge, then Precision at this number is the ideal metric.

Compared to prior related work, our contribution extends the literature on machine learning-based inattentiveness detection in important ways. Methodologically, we refine the standard autoencoder framework by tailoring the loss function to categorical survey structures and by experimenting with Percentile Loss (PL) as a strategy to improve inattentiveness detection in noisy settings. Empirically, we establish practical benchmarks through large-scale validation on nine diverse datasets, setting a foundation for future methodological comparisons. Alfons and Welz (2024), the closest work to ours, evaluated autoencoders in synthetic datasets to classify inattentive respondents, but their experiments were restricted to simulated random responding without validation against real-world survey structures. Moreover, their evaluation focused only on overall detection rates without investigating reconstruction performance, detection ranking quality, or optimization techniques.

Our work examines the performance of autoencoders in diverse real-world datasets, adapting loss structures to survey formats, tuning hyperparameters via Bayesian optimization, evaluating both reconstruction fidelity and inattentiveness ranking performance with ground truth attention checks, and proposing percentile loss for further consideration on the "difficult" cases. Welz and Alfons (2023) proposed CODERS, which also uses autoencoders but for a different task: detecting the onset of inattentive responding within a survey. Their method focuses on when a participant becomes inattentive during answering, using item-level change-point detection. Their evaluation was limited to one real dataset and simulations. In contrast, our method detects who among the full sample is inattentive overall, operating at the respondent level rather than longitudinally within a survey. Theoretically, while CODERS advances the understanding of partial or fatigue-driven carelessness, our contribution lies in establishing autoencoders as a general-purpose framework for respondent-level inattentiveness detection. Practically, our validation spans nine diverse, real-world datasets that include both attentive and inattentive respondents, which is rare in this domain. Together, these distinctions demonstrate that our work complements rather than duplicates CODERS, offering researchers tools for different levels of granularity in managing survey data quality.

Despite these contributions, several limitations should be acknowledged. First, our method assumes that attentive respondents exhibit structured patterns that the autoencoder can learn. In datasets with extremely heterogeneous populations or poorly designed surveys (e.g., without coherent constructs), reconstruction learning may struggle. Second, our approach focuses on categorical and discretized numeric variables. In the future, we want to generalize our analysis for all kinds of data. This would include the text data that exists in multiple surveys and can be encoded in the autoencoder paradigm. Incorporating open-ended text responses into the model, perhaps through the use of embedding techniques, could allow the method to handle a broader range of survey formats. Third, while our experiments cover diverse domains, we primarily evaluated surveys with relatively structured formats; future work is needed to test generalizability to highly unstructured surveys or different cultures/languages. Fourth, the choice of threshold for flagging inattentiveness based on reconstruction error remains somewhat heuristic, and further research could investigate dynamic or adaptive thresholding strategies. Although the application of Percentile Loss (PL) improved detection in some datasets, it also reduced reconstruction fidelity, revealing a trade-off that demands deeper theoretical and empirical study. It can be further investigated to improve robustness against reconstruction overfitting in datasets with complex noise patterns. This can include either a fixed percentile or a learned hyperparameter based on some measures. We believe that there is a lot of room for improvement in this direction, since there is evidence for promising results on the "difficult" cases. Finally, our evaluation relies on the basic assumption that attention checks are consistent and correct across the various datasets. There could be many different reasons to violate this hypothesis. Future studies could explore other metadata signals, such as response times or answer variance, to evaluate this methodology or even further enhance inattentiveness detection via hybrid models. This limited our analysis since the number of studies that included attention checks and made their full, non-filtered data publicly available was severely scarce.

## References

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, 18(2), e12941. https://doi.org/10.1111/spc3.12941

Alvarez, R. M., Atkeson, L. R., Levin, I., & Li, Y. (2019). Paying attention to inattentive survey respondents. Political Analysis, 27(2), 145–162. https://doi.org/10.1017/pan.2018.57

An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. Special Lecture on IE, 2(1), 1-18. http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. American Journal of Political Science, 58(3), 739–753. https://doi.org/10.1111/ajps.12081

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

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

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28(1), 20–28. https://doi.org/10.2307/2346806

Eduardo, S., Nazábal, A., Williams, C. K. I., & Sutton, C. (2020, June). Robust variational autoencoders for outlier detection and repair of mixed-type data. Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics, 108, 4056-4066. https://proceedings.mlr.press/v108/eduardo20a.html

Goodge, A., Hooi, B., Ng, S. K., & Ng, W. S. (2021, January). Robustness of autoencoders for anomaly detection under adversarial impact. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) (pp. 1244-1250). https://doi.org/10.24963/ijcai.2020/173

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. Journal of Behavioral Decision Making, 26(3), 213–224. https://doi.org/10.1002/bdm.1753

Hawkins, S., He, H., Williams, G. J., & Baxter, R. A. (2002, September). Outlier detection using replicator neural networks. In S. Kambayashi, W. Winiwarter, & M. Arikawa (Eds.), Data Warehousing and Knowledge Discovery (DaWaK 2002) (Lecture Notes in Computer Science, Vol. 2454, pp. 170–180). Springer. https://doi.org/10.1007/3-540-46145-0_17

Ivanov, S., Novisky, M. A., & Vogel, M. (2021). Racial resentment, empathy, and support for release during COVID-19: Results from a survey experiment. Socius, 7. https://doi.org/10.1177/23780231211005222

Kim, Y., Dykema, J., Stevenson, J., Black, P., & Moberg, D. P. (2019). 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

Kingma, D. P., & Welling, M. (2013, December). Auto-encoding variational Bayes [Preprint]. arXiv. https://arxiv.org/abs/1312.6114

Krogh, A., & Hertz, J. A. (1991). A simple weight decay can improve generalization. In J. E. Moody, S. J. Hanson, & R. J. Lippmann (Eds.), Advances in Neural Information Processing Systems (Vol. 4, pp. 950–957). Morgan Kaufmann. http://papers.nips.cc/paper/1991/hash/8eefcfdf5990e441f0fb6f3fad709e21-Abstract.html

Mastroianni, A. M., & Dana, J. (2022). Widespread misperceptions of long-term attitude change. Proceedings of the National Academy of Sciences, 119(11), e2107260119. https://doi.org/10.1073/pnas.2107260119

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

Merrill, N., & Olson, C. C. (2020, September). A new autoencoder training paradigm for unsupervised hyperspectral anomaly detection. 2020 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 3967–3970. https://doi.org/10.1109/IGARSS39084.2020.9324636

Moss, A. J., Rosenzweig, C., Robinson, J., Jaffe, S. N., & Litman, L. (2023). Is it ethical to use Mechanical Turk for behavioral research? Relevant data from a representative survey of MTurk participants and wages. Behavior Research Methods, 55(8), 4048–4067. https://doi.org/10.3758/s13428-022-02005-0

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

Ozaki, K. (2024). Detecting inattentive respondents by machine learning: A generic technique that substitutes for the directed questions scale and compensates for its shortcomings. Behavior Research Methods, 56(7), 7059–7078. https://doi.org/10.3758/s13428-024-02407-2

O'Grady, T., Vandegrift, D., Wolek, M., & Burr, G. (2019). On the determinants of other-regarding behavior: Field tests of the Moral Foundations Questionnaire. Journal of Research in Personality, 81, 224–237. https://doi.org/10.1016/j.jrp.2019.06.008

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. Behavior Research Methods, 54(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological Science, 31(7), 770-780. https://doi.org/10.1177/0956797620939054

Robinson-Cimpian, J. P. (2014). Inaccurate estimation of disparities due to mischievous responders: Several suggestions to assess conclusions. Educational Researcher, 43(4), 171–185. https://doi.org/10.3102/0013189X14534297

Rubio, J., Barucca, P., Gage, G., Arroyo, J., & Morales-Resendiz, R. (2020). Classifying payment patterns with artificial neural networks: An autoencoder approach. Latin American Journal of Central Banking, 1(1-4), 100013. https://doi.org/10.1016/j.latcb.2020.100013

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(56), 1929–1958. http://jmlr.org/papers/v15/srivastava14a.html

Uhalt, J. (2020). Attention checks and response quality survey data [Dataset]. OSF. https://osf.io/uz872/

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11(12), 3371–3408. http://jmlr.org/papers/v11/vincent10a.html

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. Neurocomputing, 184, 232–242. https://doi.org/10.1016/j.neucom.2015.08.104

Weaver, R., & Prelec, D. (2013). Creating truth-telling incentives with the Bayesian Truth Serum. Journal of Marketing Research, 50(3), 289–302. https://doi.org/10.1509/jmr.09.0039

Welz, M., & Alfons, A. (2023). When respondents don't care anymore: Identifying the onset of careless responding [Preprint]. arXiv preprint arXiv:2303.07167. https://arxiv.org/abs/2303.07167

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., ... & Qiao, H. (2018, April). Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. Proceedings of the 2018 World Wide Web Conference, 187-196. https://doi.org/10.1145/3178876.3185996

Yin, X., Han, J., & Yu, P. S. (2007, August). Truth discovery with multiple conflicting information providers on the web. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1048–1052). ACM. https://doi.org/10.1145/1281192.1281309

Zhou, C., & Paffenroth, R. C. (2017, August). Anomaly detection with robust deep autoencoders. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 665-674). ACM. https://doi.org/10.1145/3097983.3098052