Wiley
Human Behavior and Emerging Technologies
Volume 2024, Article ID 7267030, 15 pages
https://doi.org/10.1155/2024/7267030

WILEY

Research Article

A Step Forward in Identifying Socially Desirable Respondents: An Integrated Machine Learning Model Considering T-Scores, Response Time, Kinematic Indicators, and Eye Movements

Cristina Mazza,[1] **Irene Ceccato**,[2] **Loreta Cannito**,[3] **Merylin Monaro**,[4] **Eleonora Ricci**,[5] **Emanuela Bartolini**,[5] **Alessandra Cardinale**,[6] **Adolfo Di Crosta**,[2] **Matteo Cardaioli**,[7,8] **Pasquale La Malva**,[2] **Marco Colasanti**,[9] **Renata Tambelli**,[1] **Luciano Giromini**,[10] **Rocco Palumbo**,[2] **Riccardo Palumbo**,[5,11] **Alberto Di Domenico**,[2] **and Paolo Roma**[12]

[1]*Department of Dynamic and Clinical Psychology, and Health Studies, Sapienza University of Rome, Rome, Italy*
[2]*Department of Psychology, University "G. d'Annunzio" of Chieti-Pescara, Chieti, Italy*
[3]*Department of Social Sciences, University of Foggia, Foggia, Italy*
[4]*Department of General Psychology, University of Padua, Padua, Italy*
[5]*Department of Neuroscience, Imaging and Clinical Sciences, University "G. d'Annunzio" of Chieti-Pescara, Chieti, Italy*
[6]*National Nuclear Physics Institute, Rome, Italy*
[7]*Department of Mathematics, University of Padua, Padua, Italy*
[8]*GFT Italy, Milan, Italy*
[9]*Department of Psychological, Health and Territorial Sciences, University "G. d'Annunzio" of Chieti-Pescara, Chieti, Italy*
[10]*Department of Psychology, University of Turin, Turin, Italy*
[11]*Center for Advanced Studies and Technology (CAST), Chieti, Italy*
[12]*Department of Human Neuroscience, Sapienza University of Rome, Rome, Italy*

Correspondence should be addressed to Paolo Roma; paolo.roma@uniroma1.it

Received 3 May 2024; Accepted 3 September 2024

Academic Editor: Alhamzah Alnoor

Context: In high-stakes assessments, such as court cases or managerial evaluations, decision-makers heavily rely on psychological testing. These assessments often play a crucial role in determining important decisions that affect a person's life and have a significant impact on society.
Problem Statement: Research indicates that many psychological assessments are compromised by respondents' deliberate distortions and inaccurate self-presentations. Among these sources of bias, socially desirable responding (SDR) describes the tendency to provide overly positive self-descriptions. This positive response bias can invalidate test results and lead to inaccurate assessments.
Objectives: The present study is aimed at investigating the utility of mouse- and eye-tracking technologies for detecting SDR in psychological assessments. By integrating these technologies, the study sought to develop more effective methods for identifying when respondents are presenting themselves in a favorable light.
Methods: Eighty-five participants completed the Lie (L) and Correction (K) scales of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) twice: once answering honestly and once presenting themselves in a favorable light, with the order of conditions balanced. Repeated measures univariate analyses were conducted on L and K scale *T*-scores, as well as on mouse- and eye-tracking features, to compare the honest and instructed SDR conditions. Additionally, machine learning models were developed to integrate T-scores, kinematic indicators, and eye movements for predicting SDR.

Results: The results showed that participants in the SDR condition recorded significantly higher *T*-scores, longer response times, wider mouse trajectories, and avoided looking at the answers they intended to fake, compared to participants in the honest condition. Machine learning algorithms predicted SDR with 70%–78% accuracy.
Conclusion: New assessment strategies using mouse- and eye-tracking can help practitioners identify whether data is genuine or fabricated, potentially enhancing decision-making accuracy.
Implications: Combining self-report measures with implicit data can improve SDR detection, particularly in managerial, organizational, and forensic contexts where precise assessments are crucial.

1. Introduction

In high-stakes situations, such as legal or managerial evaluations where outcomes can significantly affect individuals and those around them, decision-makers often rely heavily on psychological testing. Unfortunately, research has shown that many psychological assessments are compromised by respondents' deception, distortions, and acquiescent response styles. The tendency to present oneself in an overly favorable light is known as socially desirable responding (SDR) [1]. Due to the social and economic consequences of SDR, detecting it is a critical area of research. Identifying valid predictors and implicit behavioral measures of SDR is particularly useful because they can provide practitioners with information about target behaviors independently of the subject's test responses, be used to validate self-report data, and are often difficult to manipulate intentionally. Furthermore, as the number of behavioral parameters increases, subjects' ability to monitor their response behavior decreases.

The present study is aimed at investigating the combined utility of mouse- and eye-tracking for SDR identification. A simulation design was employed in which each participant completed the experimental task under two sets of instructions, with the order balanced according to the group to which participants were randomly assigned. Participants either first responded honestly and then presented themselves as having perfect psychological health, omitting any criticality and impairment in psychological and behavioral functioning, or first responded in a favorable light and then responded honestly.

The research examined eye movements by comparing the areas of interest (AOIs) of selected versus unselected responses, providing insights into differences in visual exploration between SDR and honest conditions based on participants' actual responses rather than item content. The analysis also considered the Lie (L) and Correction (K) scales from the Minnesota Multiphasic Personality Inventory-2 (MMPI-2) as outcome measures. Building on the literature (see Section 2), the hypotheses were the following:

- H1: Participants in the SDR condition would score significantly higher on the L and K scales than those in the honest condition.
- H2: Participants in the SDR condition would have significantly slower mouse movements than those in the honest condition.
- H3: Participants in the SDR condition would have significantly wider and less stable mouse trajectories than those in the honest condition.
- H4: Eye movements would differ between the honest and SDR conditions. Due to conflicting findings from previous studies, no specific direction was defined for the effects.

The ultimate goal of the study was to use machine learning (ML) models to create an integrated model capable of accurately detecting SDR. This model would be based on explicit scale scores, temporal and spatial kinematic indicators, and eye movement data.

2. Literature Review

2.1. SDR. SDR is one of the most common and pervasive sources of bias in high-stakes evaluative settings, such as parenting skills evaluations, driving exams, and personnel selection. In fact, it is estimated that SDR occurs in approximately 30%–50% of personnel selection processes [2, 3] and 20%–74% of forensic settings, including child custody evaluations [4]. Despite its significant practical impact, the literature on SDR is not as extensive as it should be, and the instruments available to identify SDR are limited.

The Marlowe–Crowne Social Desirability Scale (MCSDS) [5] and the Balanced Inventory of Desirable Responding (BIDR), also known as the Paulhus Deception Scales (PDS) [6], are the most commonly used stand-alone scales for detecting SDR. In the context of personality inventories, the primary tools for identifying SDR are "embedded" validity scales or indicators. These are designed to assess the validity and interpretability of self-report questionnaires and help interpret test scores by accounting for SDR [7–11].

Traditionally, the key measures used to detect SDR, specifically "faking good" or underreporting on personality questionnaires, include the L and K scales of the MMPI-2 [12], the Virtuous Responding (VR) scale of the Psychopathic Personality Inventory-Revised (PPI-R) [13], and the Positive Impression (PIM) scale of the Personality Assessment Inventory (PAI) [14]. However, personality questionnaires often have high transparency, allowing subjects to easily discern what constructs a test or item is measuring and adjust their responses accordingly [15]. To address this issue, researchers have sought alternative, indirect methods for detecting SDR [16, 17], including reaction time (RT) [18–22], time pressure [15, 20, 23, 24], and mouse tracking [19, 25–28]. Many studies suggest that SDR takes longer to endorse because it is either more cognitively demanding [18, 20, 29] or because it increases arousal due to the fear of detection [22].

2.2. Detecting Socially Desirable Responses Using Kinematic Indicators. Recently, mouse tracking has emerged as a

valuable technique for detecting socially SDR [25, 28, 30]. This method involves capturing both the temporal and spatial aspects of mouse movement by recording cursor location at a high frequency (i.e., 60–75 times per second) [31]. Mouse trajectories provide insights into real-time mental processes during decision-making tasks, such as completing a personality questionnaire, because motor movements are continuously influenced by underlying cognitive processes [26, 32–36].

Mazza et al. [25] demonstrated that participants engaged in SDR exhibit longer RTs and greater maximum deviation (MD) times than honest respondents. The MD time refers to the time taken to reach the point of MD between the actual and ideal response trajectory using the mouse cursor. Additionally, SDR participants show wider mouse trajectories when responding to L scale items.

2.3. Eye Movements in SDR. Researchers have recently begun exploring the potential of eye-tracking to identify feigned responses, given that eye movements are physiological and not entirely under conscious control [37–43]. Eye-tracking technology records gaze location and eye movements over time and across tasks. For instance, during activities like reading or viewing images, eyes may fixate on specific AOIs or move rapidly in a motion known as saccades [44]. Visual fixations involve maintaining gaze on a target for approximately 130–330 ms, allowing the brain to begin processing visual information [44]. Saccades are the quick eye movements between fixation points, lasting about 30 ms during reading and 40–50 ms when viewing a scene [45, 46]. Blinking, defined as the rapid closing and opening of the eyelids, is another automatic ocular behavior that typically lasts 30–40 ms and occurs approximately every 2–30 s [47]. Research has shown that eye movements can reveal cognitive processing [48–50] and emotional activation [51–53]. Specifically, when individuals experience high cognitive load, their fixation duration and saccade speed increase, while their blink rate decreases [50, 54–56]. Additionally, saccades, blinks, and fixations can provide insights into physiological arousal, vigilance, and fatigue [57]. Several studies have applied eye-tracking technology in deception detection, revealing that when individuals lie, their fixation duration and saccadic movements increase, while their blink rate and duration decrease [50, 58–60].

To the best of our knowledge, few studies have specifically employed eye-tracking technology to detect SDR. In a within-subject study, Van Hooft and Born [43] asked 129 participants to complete the Five Factor Personality Inventory either honestly or with SDR in a personnel selection context. They found that all personality traits could be manipulated to create a more favorable impression. Additionally, SDR participants responded more quickly and exhibited nearly one fewer eye fixation per item on average compared to honest participants. Furthermore, SDR participants paid more attention to extreme response options (e.g., "much more/less (often) than others") than honest participants.

Logistic regression analyses revealed that test scores, RT, and the number of fixations effectively distinguished between honest and SDR participants, achieving an accuracy of 82.9%. Notably, the inclusion of eye-tracking data significantly improved the model's accuracy. These findings suggest that SDR is associated with a lower cognitive load and an altered focus of attention compared to honest responses.

More recently, Fang et al. [41] investigated the role of eye-tracking in detecting SDR and examined whether the eye movement patterns of individuals instructed to lie differ from those who lie spontaneously. The results indicated that eye movements can effectively distinguish between honest responses and SDR. Specifically, participants engaging in SDR exhibited dilated pupils and had more frequent and longer fixations. Additionally, these participants focused more on positive items (i.e., socially desirable answers), suggesting increased cognitive processing of such items. However, the study found inconsistent results with saccade and blink data, prompting the authors to recommend further research. Unlike Hooft and Born's findings, these results support the hypothesis that SDR is cognitively demanding and requires more effort than honest responses.

The present research falls within the research field of SDR detection.

3. Materials and Methods

3.1. Participants. One hundred Italian young adults voluntarily participated in the study. All were undergraduate students who were given extra credit for participating. Fifteen participants (15%) were excluded from the analysis due to technical problems that invalidated the procedure. The final sample was composed of 85 participants ($n = 60$ female, 70.6%; $n = 25$ male, 29.4%), aged 18–31 years ($M = 21.89$, $SD = 2.97$). Most participants were right-handed ($n = 80$, 94.1%), 31 wore glasses (36.5%), and 6 wore contact lenses (7.1%) (see Table 1). As detailed below, the order of the test conditions (SDR vs. honest responding) was balanced across participants. Half of the participants ($n = 43$; $M_{Age} = 21.93$, $SD = 3.16$) completed the questionnaire first honestly and then with SDR; the other half ($n = 42$; $M_{Age} = 21.86$, $SD = 2.79$) followed the opposite pattern. There were no significant differences between groups in the descriptive statistics (Table 1).

All participants provided informed consent prior to data collection. The experimental procedure was approved by the local ethics committee (Board of the Department of Human Neuroscience, Faculty of Medicine and Dentistry, Sapienza University of Rome), in accordance with the Declaration of Helsinki.

3.2. Materials

3.2.1. MMPI-2. The MMPI-2 [12] is a 51-scale self-report questionnaire that assesses personality and psychopathology. It comprises 567 items with dichotomous response options (i.e., *true*/*false*), and it is widely used in forensic and evaluative settings [61–64]. The present study analyzed the L and K underreporting scales of the Italian version of the inventory [65, 66]. The L scale, composed of 15 items, detects the acknowledgment of uncommon virtues and the tendency to offer a more socially acceptable self-image

Table 1: Descriptive statistics for the research sample.

Variable	Total ($n = 85$)	Order of instruction	
		Honest-SDR ($n = 43$)	SDR-honest ($n = 42$)	
Biological sex *n* (%)				
Female	60 (70.6)	31 (36.5)	29	34.1
Male	25 (29.4)	12 (14.1)	13	15.3
Manual dominance *n* (%)				
Right handed	80 (94.1)	39 (45.9)	41	48.2
Left handed	5 (5.9)	4 (4.7)	1	1.2
Visual correction *n* (%)				
None	48 (56.5)	23 (27.1)	25	29.4
Glasses	31 (36.5)	16 (18.8)	15	17.6
Contact lenses	6 (7.1)	4 (4.7)	2	2.4

(e.g., by asserting that the item "*I do not always tell the truth*" is false). The K scale, composed of 30 items, detects defensiveness through measures of adjustment and emotional control (e.g., "*criticism or scolding hurts me terribly*"). Higher scores on the L and K scales are associated with higher SDR.

3.3. Research Design and Experimental Procedure. A within-subjects design was implemented to control for the influence of individual and dispositional factors on eye and hand movements. The experimental task was completed individually in a neutral, quiet room in the Department of Psychological, Health and Territorial Sciences (DiSPuTer), University "G. d'Annunzio" of Chieti-Pescara, between May and December 2021. Participants, placed approximately 60 cm from the screen, completed the test on a 15.6" display laptop running Microsoft Windows, with an eye-tracker mounted (see below for details). After the initial reception, participants read and signed the informed consent form. Following this, they completed a sociodemographic questionnaire to provide data on their age, biological sex, manual dominance, and visual impairment/correction. Subsequently, the computer session started with an eye-tracker calibration procedure to ensure measurement precision. Participants were then introduced to the experimental task by a set of instructions that correlated with their first testing condition. All participants completed the experimental task twice (i.e., once for each testing condition). The Honest condition required participants to answer honestly, while the SDR condition required participants to promote an overly positive self-image. The order of conditions was randomly assigned, in alignment with previous studies [41]. The Honest condition instructions were as follows:

We are interested in some characteristics of your personality. We want you to take this test in a totally sincere fashion. Be careful, because the questionnaire contains some gimmicks to detect dishonesty. After reading each item, you should take all the time you need to respond most accurately.

The SDR condition instructions were as follows:

We are interested in learning about some of your personality traits. Imagine that you have to participate in a selection process for a job you want. In this situation, it would be advantageous for you to appear normal and in perfect psychological health. In other words, we ask you to fill out the test in such a way as to present a positive image of yourself. Be careful, because the questionnaire contains some gimmicks to detect dishonesty, and you must answer in such a way that you do not get caught. After reading the statement, use as much time as you want to respond, following these instructions.

After finishing the first experimental task, participants were shown an unrelated short video (i.e., filler task). Subsequently, participants were presented with the experimental task again, with the other set of instructions. Finally, participants were debriefed and given extra credit for their course. The full procedure lasted approximately 20 min. Figure 1 provides a general overview of the entire experimental workflow.

3.4. Experimental Task. Items were presented in the central portion of the computer screen. Participants had to click (with the mouse) a START button in the center of the screen to initiate the presentation of each item (see Figure 2(a)). They responded by clicking one of two response buttons (i.e., *true* vs. *false*) presented in the upper part of the screen: one in the upper-left corner and one in the upper-right corner (see Figure 2(b)). Items that belonged to more than one scale (e.g., L scale Item 2 and K scale Item 1) were shown to participants only once. Variables associated with these items (i.e., score, RT, MD, area under the curve (AUC), and gaze behavior features) were duplicated and integrated into all relative scales' metrics. The displayed order was consistent with the original protocols.

3.5. Collected Behavioral Measures

3.5.1. Mouse Dynamics. During the experimental task, several variables associated with mouse movement in spatial and temporal terms were automatically registered. The recorded mouse-related features were of two kinds: spatial, which included MD (i.e., maximum perpendicular distance between the actual and the ideal trajectory) and the AUC (i.e., geometric area between the actual and the ideal trajectory); and temporal, which included RT (i.e., the time between the presentation of the question and the click of the response button).

3.5.2. Eye Movements. The Tobii Pro Nano (Tobii, Karlsrovagen, Sweden) was employed for gaze sampling during

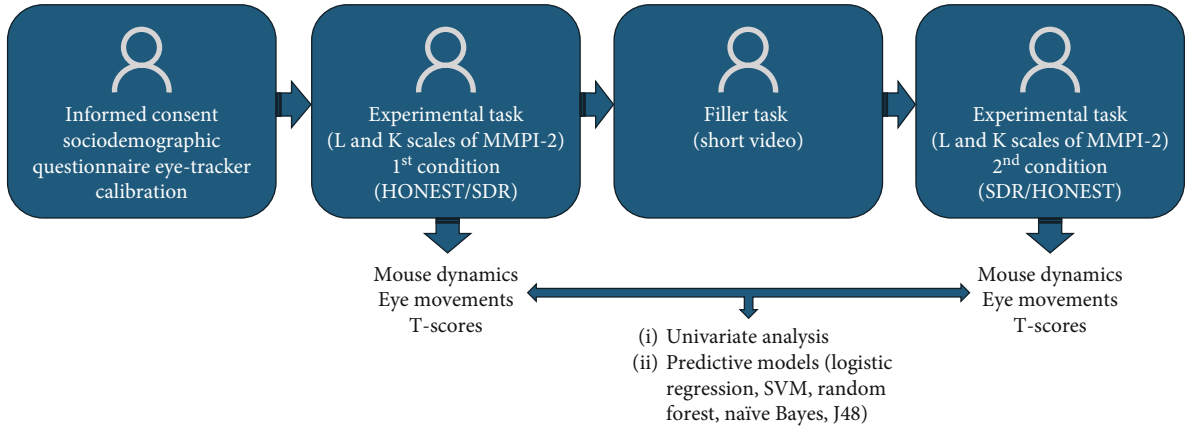

Figure 1: Experimental workflow.

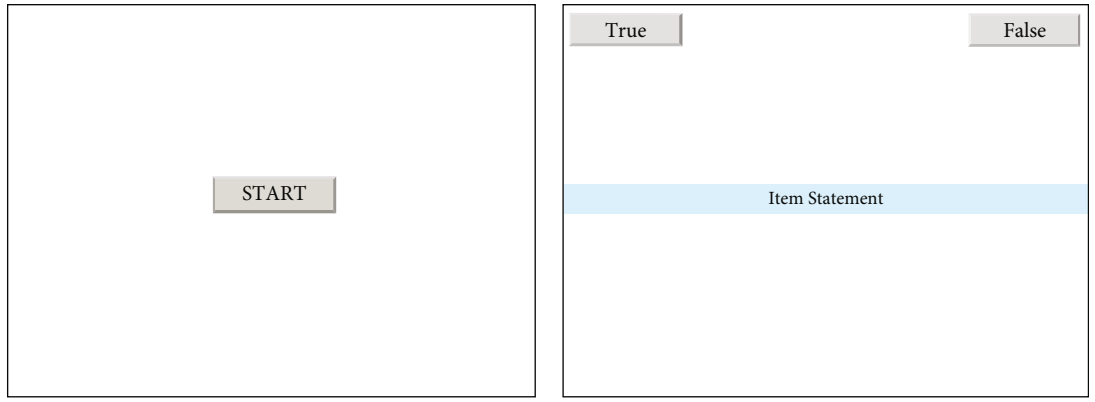

Figure 2: Example of an experimental trial as seen by participants. After the participant clicked on the (a) start button, (b) the stimulus appeared. *Note:* This is a prototypical image of how the items were presented on the screen. To ensure text security, we did not report an actual text item.

the task. Each participant was seated in front of a 15.6" display with an eye-tracking device positioned at the bottom. Within the experimental room, the lighting was stable, and the screen was adjusted to maintain constant brightness and contrast. The eye-tracking software Tobii Pro Lab, version 1.145 [67], was used to calibrate participants' gaze as they followed a white dot across the screen. If a participant was unable to calibrate, their seating posture was adjusted, and the calibration process was repeated until calibration was achieved.

The Tobii Pro Lab software was used also to preprocess eye-tracking data. Specifically, the Tobii "I-VT fixation" filter was used to analyze fixations. Velocity-threshold identification (I-VT), a velocity-based algorithm, defines fixations when eye movement velocity is below a specific threshold. In the current study, default filter values were maintained (i.e., max angle between fixations = 0.5 degrees; velocity threshold = 30 degrees/second; max time between fixations = 75 ms; minimum fixation duration = 60 ms). Parameters that have proven reliable in detecting deception [41, 43] were then computed and employed for the statistical analyses. Table 2 describes each of the considered parameters.

Gaze behavior was analyzed to assess differences in ocular activity between the two testing instructions (i.e., honest vs. SDR) on the response buttons. Initially, it was tested whether the instructions affected participants' visual exploration of the response options, independently of their actual response (i.e., response AOI). Analyses of these AOIs are reported in the Supporting Information (available here). The focus of interest was eye movements as a function of response selection. To this end, two AOIs were defined for both the L and the K scale, and gaze behavior was compared across the selected and unselected responses (i.e., selected response AOIs). Statistical testing was conducted separately for each scale.

Of note, a percentage (rather than raw count) was calculated for the fixations and saccades, to control for individual differences in the visual exploration of the screen. Also, blink frequency (i.e., count/s) was analyzed instead of blink count, to account for differences in trial duration.

3.6. Statistical Analysis

3.6.1. Univariate Analyses. To ensure that participants correctly understood and followed the instructions for each

Table 2: Gaze behavior features and indicators.

Feature	Indicator
Fixation percentage	Number of fixations during a specified time interval (i.e., onset to the end of each trial) within target AOIs, as a proportion of the total number of fixations during the same interval
Fixation average duration (ms)	Mean elapsed time between the first and the last gaze point in the sequence of gaze points in the fixation
Saccade percentage	Number of saccades during a specified time interval (i.e., onset to the end of each trial) within target AOIs, as a proportion of the total number of saccades during the same interval
Saccade average duration (ms)	Mean time required to move the fovea from the initial to the final position
Dwell time (ms)	Total elapsed time between the first and the last fixation inside the AOI
Blink frequency (count/s)	Number of blinks in the period of interest, adjusted for trial duration. Only data gaps > 100 ms considered

Note: Each indicator was measured within a trial (i.e., item) and then averaged across all trials composing a questionnaire.
Abbreviation: AOI = area of interest.

condition, scores on the outcome measures were compared via separate paired *t*-tests on the L and K scales *T*-score. Paired *t*-tests were also run to compare temporal and spatial features of mouse movement in the honest vs. SDR conditions. Separate analyses were conducted for the L and K scales. Repeated measure ANOVAs with instruction (SDR vs. honest) and response AOI (selected vs. unselected) were also conducted.

Effect sizes were interpreted as follows: for paired *t*-tests, Cohen's $d = 0.2$ was considered indicative of a small effect, $d = 0.5$ a medium effect, and $d = 0.8$ a large effect [68]. For repeated measure ANOVAs, $\eta^2 p = 0.01$ was considered indicative of a small effect, $\eta^2 p = 0.06$ a medium effect, and $\eta^2 p = 0.14$ a large effect [69]. The *p* value was considered significant at the 0.05 level. Analyses were performed using IBM SPSS v.25 [70].

3.6.2. Predictive Models. To investigate the effectiveness of *T*-scores and mouse and eye movements in SDR detection, a predictive statistical approach was adopted, implementing ML models. ML techniques have been recently applied to a broad range of domains [71, 72], including predicting human behavior and, specifically, over- and underreporting [29, 73, 74]. In the present study, ML analyses were run in WEKA 3.9 [75], following a best practice workflow comprised of (a) feature selection, (b) model training and validation, and (c) model testing using an out-of-sample group [76]. As ML models are built to fit data, their fit with new (i.e., unseen) data must be tested. A data training set is generally used to train and validate the model, while a data test set is used to test the model's accuracy on new data. This procedure guarantees generalization and increases the replicability of the results [77–79]. For this purpose, participants were randomly split into training ($n = 60$) and test ($n = 25$) sets. The training set consisted of 120 responses (60 honest and 60 SDR), while the test set included 50 responses (25 honest and 25 SDR).

First, feature selection was run with the aim of removing redundant and irrelevant features, and thereby increasing model generalization by reducing overfitting and noise in the data [80]. This was performed using a correlation-based feature selector (CFS) [81]. The CFS algorithm uses the "greedy stepwise" search method to evaluate a subset of features, in terms of the individual predictive ability of each feature and the redundancy with other predictors. It selects the subset with the highest correlation with the dependent variable (i.e., honest vs. SDR), but low intercorrelation. The predictors that resulted from this process were fed as inputs to several ML models.

These models were then trained and validated using 10-fold cross-validation [82]—a procedure that consists of repeatedly partitioning the sample into training and validation sets. Thus, the sample of 120 responses was randomly partitioned into 10 equal-size subsamples, or folds (i.e., 10 folds of 12 responses). One of the 10 folds was retained as validation data to test the model, and the remaining 9 folds were used as training data. This process was repeated 10 times, with each of the 10 folds used once as validation data. The results of the 10 folds were then averaged to produce a single estimation of prediction accuracy. Finally, to evaluate the accuracy of the validated models in classifying unseen participants as honest or SDR, they were tested on the out-of-sample test set of 50 responses. The predictive performance of the models was evaluated using accuracy, precision, recall, and *F*-measure (i.e., *F*1 score). Together, these metrics provide a comprehensive assessment of model performance. In certain contexts, such as forensics and managerial settings, it is crucial not to rely solely on overall accuracy but also to consider false positives and false negatives.

As described above, classification accuracy was assessed by applying different ML algorithms to determine whether the results remained consistent across different classifiers and were not influenced by specific model assumptions. The algorithms we selected represent a range of classification strategies, including regression, classification trees, and Bayesian statistics. Specifically, the following algorithms were used, based on relevant previous literature [25, 30]:

- Logistic regression [83]: This method evaluates the relationship between a categorical dependent variable and one or more independent variables, using a logistic function to estimate probabilities.

- Support vector machine (SVM) [84]: A binary linear classifier that organizes data in space and separates

categories by a margin that is maximized for greater classification accuracy.
- Random forest [85]: An ensemble learning technique that builds multiple decision trees and aggregates their results.
- Naïve Bayes [86]: A probabilistic classifier that applies Bayes' theorem, assuming independence between predictors.

In addition, the J48 tree model [87] was implemented to facilitate interpretability. J48, often referred to as an implementation of the C4.5 algorithm in Weka software, builds decision trees by selecting attributes that provide the highest normalized information gain. It is one of the simplest classifiers in terms of transparency, as it emphasizes the logic behind the classification [88]. It should be noted that all algorithms were run using the default parameters of WEKA 3.9 [75], without any fine-tuning to increase accuracy.

4. Results

4.1. L and K Scale T-Scores. A significant effect was found for instruction (i.e., honest vs. SDR) on L scale T-scores, $t_{(84)} = 11.15$, $p < 0.001$, with a large effect size, $d = 1.21$. As expected, participants in the SDR condition recorded significantly higher T-scores. A significant ($t_{(84)} = 9.43$, $p < 0.001$) and large effect ($d = 1.02$) was also found for instruction on K scale T-scores, with participants in the SDR condition scoring significantly higher (see Table 3 and Figure 3).

4.2. Mouse-Tracking Variables

4.2.1. Mouse-Tracking Variables on the L Scale. A significant and small effect of instruction (i.e., honest vs. SDR) was found for all mouse-tracking variables on the L scale: RT ($t_{(84)} = 2.32$, $p = 0.023$, $d = 0.25$), AUC ($t_{(84)} = 2.99$, $p = 0.004$, $d = 0.33$), and MD ($t_{(84)} = 2.94$, $p = 0.004$, $d = 0.32$). For each variable, the average was higher for participants in the SDR condition (see Figure 4(a) and Table 4).

4.2.2. Mouse-Tracking Variables on the K Scale. As shown in Table 4, a significant and small effect of instruction (i.e., honest vs. SDR) was found for both the AUC and MD on the K scale: AUC ($t_{(84)} = 2.11$, $p = 0.038$, $d = 0.23$) and MD ($t_{(84)} = 2.70$, $p = 0.009$, d=0.29). The average of these variables was higher for participants in the SDR condition (see Figure 4(b)).

4.3. Eye-Tracking Variables on Selected Response AOI. A repeated measures ANOVA with instruction (i.e., SDR vs. honest) and response AOI (i.e., selected vs. unselected) was conducted for all eye-tracking parameters (see Table 5 for descriptive statistics).

4.3.1. Eye Movements on the L Scale. For fixation percentage and average duration, the main effects of instruction and response AOI were not significant, $Fs_{(1,84)} \leq 1.22$, $ps \geq 271$. However, a significant interaction effect emerged for both fixation percentage, $F_{(1,84)} = 62.45$, $p < 0.001$, $\eta^2 p = 0.43$, and fixation average duration, $F_{(1,84)} = 65.92$, $p < 0.001$, $\eta 2p = 0.44$. Pairwise comparisons showed that the selected response received more and longer fixations in the honest condition. In contrast, the unselected response AOI was fixated on more frequently and for a longer duration in the SDR condition.

Similar results emerged for saccade percentage and average duration, which showed nonsignificant main effects for instruction and response AOI, $Fs_{(1,84)} \leq 2.29$, $ps \geq 0.134$, yet significant interaction effects. For saccade percentage, the interaction effect was significant, $F_{(1,84)} = 27.14$, $p < 0.001$, $\eta^2 p = 0.24$, and pairwise analyses indicated that the selected response received more saccades in the honest condition, while the unselected response received more saccades in the SDR condition. For saccade average duration, the interaction effect was significant, $F_{(1,84)} = 7.77$, $p = 0.007$, $\eta^2 p = 0.09$, and pairwise analyses indicated that in the unselected response AOI, saccades tended to be longer in the honest condition, while no difference between conditions emerged in saccade duration within the selected response AOI.

For dwell time, a main effect of instruction emerged, $F_{(1,84)} = 6.85$, $p = 0.010$, $\eta^2 p = 0.08$, with higher dwell time recorded by participants in the SDR condition. A significant interaction between instruction and response AOI was also found, $F_{(1,84)} = 61.48$, $p < 0.001$, $\eta 2p = 0.42$. Pairwise

Table 3: Results of the paired t-test on T-scores of the L and K scales.

Variable	Instruction	*M*	SD	Difference
L scale	SDR	64.42	13.17	**16.07**
	Honest	48.35	9.14	
K scale	SDR	53.84	9.14	**9.48**
	Honest	44.35	8.71	

Note: Statistically significant effects ($p < 0.05$) are in bold. The final column reports the difference between the two means (M_{SDR}-M_{Honest}).

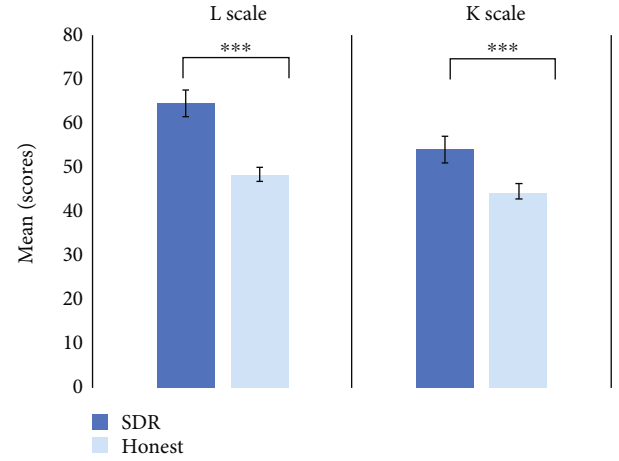

Figure 3: Effects of instruction (Honest vs. SDR) on T-scores of the L and K scales. *Note:* ***$p < 0.001$. Error bars indicate the standard error (SE).

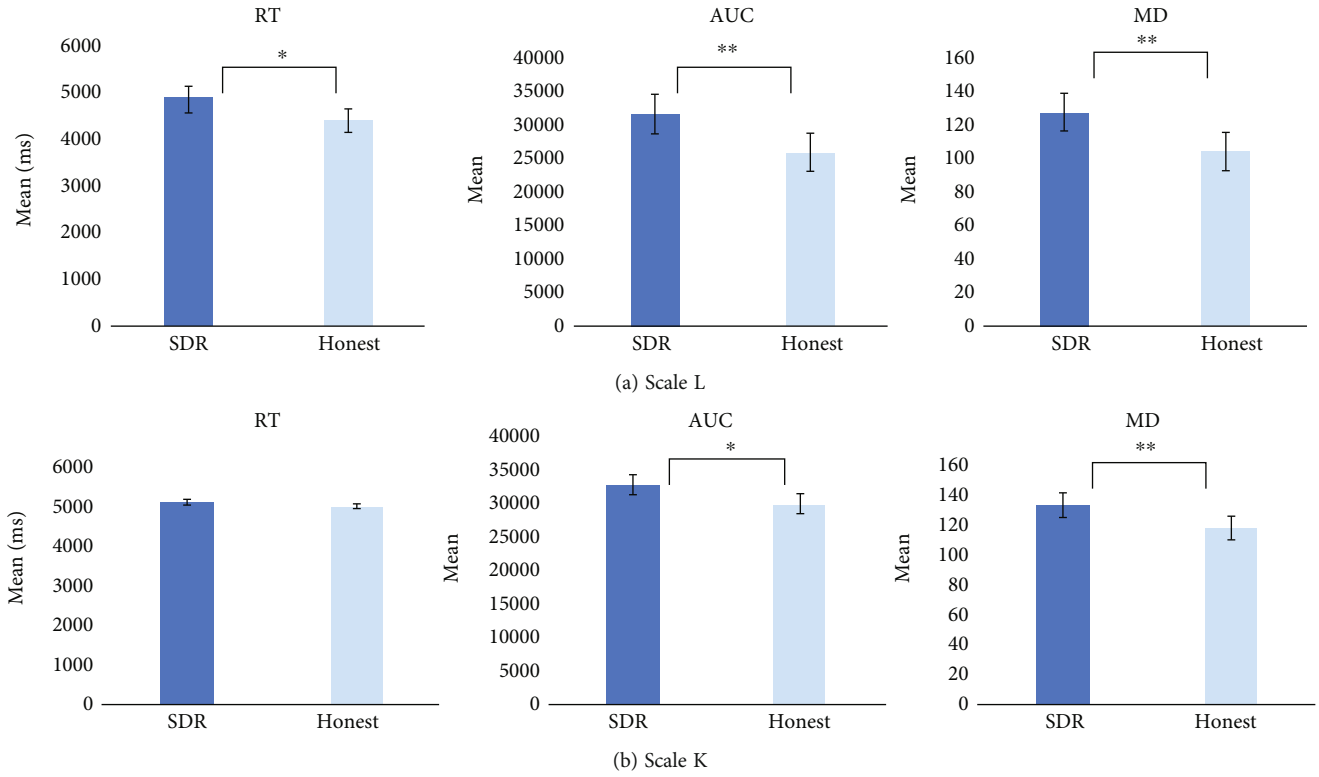

FIGURE 4: Effects of instruction (Honest vs. SDR) on mouse-tracking features. *Note*: $^{*}p < 0.05$; $^{**}p < 0.01$. Error bars indicate the standard error (SE).

TABLE 4: Results of the paired *t*-test on the mouse-tracking variables for the L and K scales.

Variable	Instruction	L scale			K scale		
		M	*SD*	Difference	*M*	*SD*	Difference
RT	SDR	4895.32	1864.38	**489.95**	5110.29	1503.61	76.29
	Honest	4405.37	1323.17		5033.99	1372.27	
AUC	SDR	32018.97	18629.06	**5928.41**	32792.72	17719.49	**3127.66**
	Honest	26090.57	16853.93		29665.05	16582.82	
MD	SDR	128.77	71.23	**23.61**	134.58	71.92	**15.86**
	Honest	105.16	66.22		118.73	62.19	

Note: The difference between the two means (M_{SDR}-M_{Honest}) is reported. Statistically significant effects ($p < 0.05$) are in bold.

comparisons indicated that dwell time on the selected response AOI was higher in the honest condition. In contrast, dwell time on the unselected response AOI was higher in the SDR condition.

Finally, for blink frequency, there was a significant main effect of instruction, $F_{(1,84)} = 22.05$, $p < 0.001$, $\eta^2 p = 0.21$, with higher blink frequency in the SDR condition. Also, the main effect of response AOI was significant, $F_{(1,84)} = 6.66$, $p = 0.012$, $\eta^2 p = 0.07$, indicating that blink frequency was higher for the unselected response AOI. Notably, the interaction effect was also significant, $F_{(1,84)} = 102.10$, $p < 0.001$, $\eta^2 p = 0.55$. Pairwise comparisons indicated that blinks were more frequent in the selected response AOI in the honest condition and in the unselected response AOI in the SDR condition (see Table 6).

4.3.2. Eye Movements on the K Scale. The main effect of instruction was not significant for all tested parameters (all $Fs_{(1,84)} \leq 3.41$, $ps \geq 0.07$). A main effect of response AOI was detected for fixation percentage ($F_{(1,84)} = 4.67$, $p = 0.033$, $\eta^2 p = 0.05$) and blink frequency ($F_{(1,84)} = 8.16$, $p = 0.005$, $\eta^2 p = 0.09$), with both higher in the SDR condition. No other tested parameter presented a significant main effect for the selected response (all $Fs_{(1,84)} \leq 3.57$, $ps \geq 0.062$).

Significant interaction effects were observed for all parameters. For fixation percentage ($F_{(1,84)} = 26.45$, $p < 0.001$, $\eta^2 p =$

Table 5: Descriptive statistics for the eye-tracking variables of the L and K scales.

Variable	Instruction	AOI	L scale		K scale	
			M	SD	M	SD
Fixation percentage	SDR	Selected	5.78	4.58	2.70	1.77
		Unselected	9.78	4.98	3.88	2.37
	Honest	Selected	9.99	7.29	3.53	2.15
		Unselected	6.55	4.04	3.28	2.04
Fixation average duration	SDR	Selected	185.81	145.83	195.50	123.90
		Unselected	296.27	141.83	258.99	130.15
	Honest	Selected	287.57	173.26	251.37	123.14
		Unselected	176.54	117.22	203.27	108.20
Saccade percentage	SDR	Selected	2.51	2.82	1.09	1.02
		Unselected	3.95	2.94	1.59	1.40
	Honest	Selected	4.64	5.08	1.60	1.44
		Unselected	2.73	2.65	1.34	1.20
Saccade average duration	SDR	Selected	36.62	17.73	36.95	19.07
		Unselected	33.18	13.26	39.74	13.04
	Honest	Selected	33.70	12.20	42.45	12.69
		Unselected	36.91	14.60	39.32	15.51
Dwell time	SDR	Selected	376.16	328.21	355.93	249.17
		Unselected	624.51	330.08	512.54	293.09
	Honest	Selected	536.07	397.78	466.29	281.51
		Unselected	328.92	228.40	415.92	246.63
Blink frequency	SDR	Selected	0.280	0.194	0.273	0.152
		Unselected	0.483	0.210	0.390	0.181
	Honest	Selected	0.366	0.180	0.333	0.154
		Unselected	0.267	0.161	0.319	0.157

Table 6: Pairwise comparisons of the interaction effects between instruction and selected response AOI on eye-tracking variables in the L and K scales.

Variable	Selected response AOI			Unselected response AOI		
	SDR	Honest	*p*	SDR	Honest	*p*
L scale						
Fixation percentage	−	+	< 0.001	+	−	< 0.001
Fixation average duration	−	+	< 0.001	+	−	< 0.001
Saccade percentage	−	+	< 0.001	+	−	0.001
Saccade average duration	=	=	0.187	—	+	0.051
Dwell time	−	+	< 0.001	+	−	< 0.001
Blink frequency	−	+	< 0.001	+	−	< 0.001
K scale						
Fixation percentage	−	+	< 0.001	+	−	0.011
Fixation average duration	−	+	< 0.001	+	−	< 0.001
Saccade percentage	−	+	0.002	=	=	0.085
Saccade average duration	−	+	0.015	=	=	0.783
Dwell time	−	+	< 0.001	+	−	< 0.001
Blink frequency	−	+	< 0.001	+	−	< 0.001

Note: The symbol "+" ("−") indicates higher (lower) mean scores for a given variable (row) in a given instruction condition (column), compared to the opposite condition.

Table 7: Results of different machine learning algorithms in 10-fold cross-validation and the test set.

Algorithm		Accuracy	Precision	Recall	*F*-measure
Logistic	10-fold cross-validation	78.33%	0.78	0.78	0.78
	Test set	70%	0.70	0.70	0.70
SVM	10-fold cross-validation	76.67%	0.77	0.77	0.77
	Test set	74%	0.75	0.74	0.74
Random forest	10-fold cross-validation	80.83%	0.81	0.81	0.81
	Test set	72%	0.72	0.72	0.72
Naïve Bayes	10-fold cross-validation	80%	0.80	0.80	0.80
	Test set	78%	0.78	0.78	0.78
J48	10-fold cross-validation	83.33%	0.84	0.83	0.83
	Test set	72%	0.72	0.72	0.72

0.24), fixation average duration ($F_{(1,84)} = 37.84$, $p < 0.001$, $\eta^2 p = .31$), dwell time ($F_{(1,84)} = 33.04$, $p < 0.001$, $\eta^2 p = 0.28$), and blink frequency ($F_{(1,84)} = 34.21$, $p < 0.001$, $\eta^2 p = 0.29$), pairwise comparisons showed lower values for the selected response AOI and higher values for the unselected response AOI in the SDR condition. For saccade percentage ($F_{(1,84)} = 20.32$, $p < 0.001$, $\eta^2 p = 0.20$), pairwise comparisons showed that the SDR condition was associated with a lower saccade percentage in the selected response AOI, while no difference between conditions emerged for unselected response AOI. A similar pattern was detected for saccade average duration ($F_{(1,\ 84)}$=5.00, $p = 0.028$, $\eta^2 p = 0.06$), with pairwise comparisons highlighting lower average duration in the selected response AOI in the SDR condition, while no difference between conditions emerged for the unselected response AOI.

Taken together, these results suggest that SDR participants devoted less attention to the selected response and more attention to the unselected response (see Table 6).

4.4. Predictive Models. To identify in which condition participants had responded to the task (SDR vs. Honest), 32 variables considered in the statistical analysis were treated as possible predictors and included in feature selection. CFS identified the following six as the best set of predictors:

- L scale *T*-score;
- K scale *T*-score;
- Blink frequency on the unselected response AOI of the L scale;
- Dwell time on the unselected response AOI of the L scale;
- Fixation average duration on the unselected response AOI of the L scale;
- Fixation average duration on the unselected response AOI of the L scale.

ML algorithms were trained, validated, and tested on these six variables, according to the procedure described above. Table 7 reports the results of the 10-fold validation procedure and the model performance in the test set.

Classification accuracy was stable between the different classifiers, ranging from 70%–78% in the test set. The decision tree (i.e., J48) obtained the best performance (83.33%) in the 10-fold cross-validation, but it also generalized the least, as its accuracy dropped to 72% in the test set. The best classifier was naïve Bayes, which achieved good accuracy in the training set (80%) and maintained a similar performance in the test set (78%).

The rule used by the decision tree algorithm to classify a response as honest or SDR was as follows:

L scale T − score ≤ 58.

|L scale unselected AOI blink frequency ≤ 0.33: Honest.

|L scale unselected AOI blink frequency > 0.33: SDR.

L scale T − score > 58: SDR.

Notably, the rule was very simple, and its classification accuracy was 72%, considering only two variables (i.e., L scale *T*-score and L scale unselected AOI blink frequency).

Finally, looking at the confusion matrix of the test set classification, all algorithms made a number of errors, resulting in a slightly higher number of false negatives (i.e., SDR responses classified as Honest) than false positives (i.e., Honest responses classified as SDR): logistic FP = 6/25, FN = 9/25; SVM FP = 4/25, FN = 9/25; Random forest FP = 6/25, FN = 8/25; naïve Bayes FP = 4/25, FN = 7/25; J48 FP = 8/25, FN = 6/25.

5. Discussion

The first aim of the present research was to replicate and confirm the findings of previous studies regarding the role of *T*-scores and temporal and kinematic indicators in detecting SDR on a personality questionnaire. The results supported the first hypothesis (H1), according to which T-scores on the MMPI-2 underreporting L and K scales were expected to be higher in the SDR condition. These findings are aligned with the results of previous studies [20, 23, 25, 43] indicating that SDR respondents tend to obtain higher scores on the MMPI underreporting scales. In this sense, the result reflects that the study instructions were correctly understood by participants, as participants who were

instructed to engage in SDR presented themselves in a more positive way by selecting socially desirable alternatives. Additionally, on average, participants in the SDR condition scored 1.5 SD above the mean on the L scale and approximately 16 T-points differently from participants in the honest condition. In contrast, participants in the SDR condition scored much closer to the average on the K scale, and fewer than 10 T-points differently from participants in the honest condition. These findings could be considered a proxy confirmation of the scales' construct validity. Indeed, the 15 items on the L scale refer to relatively common behaviors, minor infractions, and faults/weaknesses that most people would admit to. By responding negatively to these items, SDR participants evidenced a tendency to provide a socially virtuous and well-adjusted self-image. L scale items measure social desirability more accurately than K scale items, which instead measure emotional control and denial, across several thematic areas (e.g., hostility, mistrust, family conflict, excessive worry). Furthermore, L scale items are more transparent and easier to feign than the less obvious items of the K scale.

The latter consideration is also useful in explaining why SDR participants registered significantly longer RTs on the L scale, but not the K scale. Thus, the results supported the second hypothesis (H2) only in relation to the L scale. K scale items, being less obvious in the construct they are designed to measure, likely determined that honest participants needed more time to respond. This finding is consistent with previous studies showing that, compared to honest respondents, SDR respondents take more time to respond to stimuli (see, for a meta-analysis [89]). In this vein, the self-schema model [90] suggests that fakers take longer to answer a self-report questionnaire than honest respondents. Indeed, research suggests that faking requires more time, either because it is more cognitively demanding [18, 20, 29] or because it heightens arousal due to a fear of detection [22].

Recently, mouse dynamics have been found to be useful for the identification of deception. Studies by Monaro et al. [19, 27] have shown that, when half of a sample answer an autobiographical questionnaire honestly and the other half answer according to fake profiles learned just prior to testing, honest respondents follow the more direct trajectory to the desired answer, whereas those answering according to a fake profile show trajectories that initially converge towards the actual autobiographical information and then switch towards a relevant alternative. In line with this, in the present study, the mouse trajectories of participants in the SDR condition were wider (in terms of AUC and MD) than those in the honest condition. Thus, the third hypothesis (H3) was supported, consistent with a previous application of mouse-tracking to identify SDR [25], which found wider mouse trajectories only for the L scale.

The second goal of the present study was to explore whether eye movements could improve the detection of SDR on personality inventories. In line with the fourth hypothesis (H4), SDR and honest responding were associated with different visual patterns. On the L scale, participants in the SDR condition fixated more often and frequently on the unselected response AOI, with more (but shorter) saccades, higher dwell time, and more frequent blinks. Similar results were found for the K scale, except for saccade percentage and average number, for which no significant differences were found between testing conditions in the unselected response AOI. In other words, when answering honestly, participants focused more on the option they eventually chose, in line with cognitive models of decision-making. In contrast, SDR participants attended more to the unselected response. A possible explanation for this is that, when faking, respondents attempt to avoid the "correct" answer, yet it nevertheless catches their attention. Supporting this explanation, participants in the SDR condition registered lower blink frequency in the selected response AOI. Albeit in the context of a feigning experimental paradigm, a similar result on eye movements has been found by Ales et al. [37]. Specifically, the authors studied the ocular movements of healthy participants asked to feign schizophrenia while responding to an SVT (i.e., IOP-29) compared with control participants instructed to respond honestly. Findings showed that feigners paid more attention than controls to those response options identified as more indicative of feigning, even if they eventually decided not to endorse them.

A direct comparison of the current results with those of previous studies is not possible, as AOIs were examined on the basis of the response, rather than item content. Furthermore, to the best of our knowledge, this was the first study to analyze items with two response alternatives, in line with the MMPI-2. Contrary to both Hooft and Born [43] and Fang et al. [41], little evidence was found for overall smaller or greater fixations and saccades in the SDR condition. Differences between conditions emerged only when the selected/unselected responses were separated, thus accounting for participants' decisions.

The third and main aim of the present research was to develop an ML model integrating *T*-scores and mouse and eye movements to accurately predict SDR on the MMPI-2 underreporting scales. The main advantage of ML models is their ability to draw inferences at the individual level, with utility for (e.g.) clinical, recruiting, and forensic settings. The present study validated models to predict SDR on the MMPI-2 with 70%–78% accuracy. In particular, the model trained with the naïve Bayes algorithm obtained satisfactory accuracy in both the validation and the test set, and good generalization on previously unseen subjects. SDR seems slightly more difficult to identify than honest behavior, as all the classifiers produced a greater number of false negatives than false positives. Interestingly, classification was based on only L and K scale *T*-scores, and L scale eye-tracking variables (especially for the unselected response AOI). Moreover, a simple decision tree model demonstrated that two variables (i.e., L scale *T*-score, blink frequency in the unselected response AOI of the L scale) could obtain 72% accuracy. Interestingly, Hooft and Born [43] attributed predictive importance to fixation number, using hierarchical logistic regression. In contrast, in the ML algorithms employed in the present study (which were not explicitly programmed), the average duration of fixations and number of blinks played a central role. Compared with the model proposed by Fang et al. [41], in

which fixation count and pupil size were decisive (though the model was built on only a small number of participants), the present models replicated the 74% accuracy with the SVM algorithm and improved on this (78% accuracy) with the naïve Bayes classifier.

As previously mentioned, SDR is a prevalent source of bias that undermines accurate assessments in high-stakes evaluative settings. The importance of research in SDR detection is underscored by its substantial social and economic costs. For instance, hiring personnel with undesirable traits, particularly in managerial roles, can lead to significant economic, managerial, and organizational losses. Even more severe consequences can arise from appointing unsuitable individuals to public positions of high responsibility, such as pilots, military personnel, law enforcement agents, and teachers. In forensic contexts, granting a driver's license to an ineligible driver poses a considerable risk to road safety. Similarly, SDR during a parenting skills assessment in child custody cases can result in incorrect judgments, endangering the child's physical and psychological well-being. Mental health professionals in forensic psychology, including those involved in family law and child custody hearings, face challenges in providing reliable support to decision-makers when family outcomes are at stake.

This research not only advances scientific understanding of SDR but also has a tangible social impact. In institutional settings, new SDR assessment strategies could be particularly beneficial for experts in personnel selection, enhancing their ability to accurately differentiate between candidates. The proposed SDR detection strategies can be utilized by professionals in private practice, as well as organizations involved in evaluation and personnel selection across both civilian and military sectors. In forensic contexts, these findings could assist mental health professionals by indicating whether assessment data is genuine or fabricated to influence the court, potentially leading to more informed sentencing decisions.

6. Conclusions

Overall, the current findings highlight that individuals who respond to personality questionnaires in a socially desirable manner tend to have longer RTs and wider mouse trajectories and often avoid looking at the answers they intend to fake, compared to honest respondents. Building on the pioneering work of Hooft and Born [43], these results contribute significantly to the development of innovative methods for detecting SDR by integrating new technologies, such as mouse tracking and eye tracking, with ML algorithms.

When interpreting the present results, it is important to consider some limitations. First, as noted by Hooft and Born [43], the SDR experimental procedure instructed participants to present themselves as the best candidate for a job while avoiding detection. This instruction may have affected participants' reporting accuracy, as personnel selection involves a significantly different psychological experience than the actual study task. Second, the study employed a within-participants design, which is advantageous for controlling individual and dispositional factors but limits the generalizability of the results to more ecological and operational settings. Additionally, this design introduces another limitation: Participants performed the same experimental task twice, meaning they had already read the items once when completing the task after the second set of instructions. Although the order of instructions was balanced, this could have influenced the results. The repeated exposure to the items might have led to a familiarity or learning effect that future research could explore using a between-subjects design and a congruent participant group. Another limitation is that participants were aware that their eye movements were being recorded due to the calibration process with eye-tracking technology prior to data collection. As a result, participants may have consciously directed their gaze. However, it is unlikely that they could self-regulate their gaze for the entire duration of the experiment.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Disclosure

Preliminary results of this work were published in the form of abstract: Mazza C., Ceccato I., Cannito L., Monaro M., Ricci E., Bartolini E., Colasanti M., Di Crosta A., Cardaioli M., Cardinale A., La Malva P., Palumbo R., Giromini L., Palumbo R., Di Domenico A., Roma P. (2023), An Integrated Machine Learning Model Considering T-Scores, Response Time, Kinematic Indicators, and Eye Movements in Identifying Socially Desirable Responding. *Suvremena psihologija. Contemporary Psychology. Journal for psychodiagnostics theory, practice and other fields of psychology*, Vol. 26, Supp. 1, p.25.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contributions

Cristina Mazza and Irene Ceccato contributed equally.

Funding

Dr. Irene Ceccato was supported by the Programma Operativo Nazionale PON-AIM (code: AIM1811283) to the University "G. d'Annunzio" of Chieti-Pescara.

Acknowledgments

Dr. Irene Ceccato was supported by the Programma Operativo Nazionale PON-AIM (code: AIM1811283) to the University "G. d'Annunzio" of Chieti-Pescara.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. *(Supporting Information)* SM1: mathematical representation of the machine learning models trained and validated using a 10-fold cross-validation procedure. SM2: impact of instruction on visual exploration of response options. SM3: list of abbreviations.

References

[1] D. L. Paulhus, "Socially desirable responding: the evolution of a construct," in *The role of constructs in psychological and educational measurement*, H. I. Braun, D. N. Jackson, and D. E. Wiley, Eds., pp. 49–69, Erlbaum, 2002.

[2] J. J. Donovan, S. A. Dwight, and G. M. Hurtz, "An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique," *Human Performance*, vol. 16, no. 1, pp. 81–106, 2003.

[3] R. L. Griffith and P. D. Converse, "The rules of evidence and the prevalence of applicant faking," in *New perspectives on faking in personality assessment*, M. Ziegler, C. Mac Cann, and R. D. Roberts, Eds., pp. 34–52, Oxford University Press, 2012.

[4] R. A. Baer and J. Miller, "Underreporting of psychopathology on the MMPI-2: a meta-analytic review," *Psychological Assessment*, vol. 14, pp. 16–26, 2002.

[5] D. P. Crowne and D. Marlowe, *The approval motive: Studies in evaluative dependence*, Wiley, 1964.

[6] D. L. Paulhus, *Paulhus Deception Scales (PDS): The Balanced Inventory of Desirable Responding-7: User's manual*, Multi-Health Systems, 1999.

[7] D. P. Crowne and D. Marlowe, "A new scale of social desirability independent of psychopathology," *Journal of Consulting Psychology*, vol. 24, pp. 349–354, 1960.

[8] A. L. Edwards, *The social desirability variable in personality assessment and research*, Dryden Press, 1957.

[9] F. Hoeth, R. Büttel, and H. Feyerabend, "Experimentelle Untersuchungen zur Validit t von Pers nlichkeitsfragebogen. [experimental investigation on the validity of personality questionnaires]," *Psychologische Rundschau*, vol. 18, pp. 169–184, 1967.

[10] D. L. Paulhus, "Measurement and control of response bias," in *Measures of social psychological attitudes: Vol. 1. Measures of personality and social psychological attitudes*, J. P. Robinson, P. R. Shaver, and L. S. Wrightsman, Eds., pp. 17–59, Academic Press, 1991.

[11] M. Schneider-Düker and J. F. Schneider, "Studies on the answering process in psychodiagnostic questionnaires," *Zeitschrift für Experimentelle und Angewandte Psychologie*, vol. 24, no. 2, pp. 282–302, 1977.

[12] J. N. Butcher, J. R. Graham, Y. S. Ben-Porath, A. Tellegen, W. G. Dahstrom, and B. Kaemmer, *MMPI-2. Manual for administration and scoring (rev. ed.)*, University of Minnesota Press, 2001.

[13] S. O. Lilienfeld and M. R. Widows, *Psychopathic personality inventory–revised: professional manual*, Psychological Assessment Resources, 2005.

[14] L. C. Morey, *Personality assessment inventory*, Psychological Assessment Resources, 1991.

[15] L. Khorramdel and K. D. Kubinger, "The effect of speediness on personality questionnaires: an experiment on applicants within a job recruiting procedure," *Psychology Science*, vol. 48, no. 3, pp. 378–397, 2006.

[16] E. Helmes and R. R. Holden, "Response styles and faking on the basic personality inventory," *Journal of Consulting and Clinical Psychology*, vol. 54, no. 6, pp. 853–859, 1986.

[17] R. R. Holden, L. L. Wood, and L. Tomashewski, "Do response time limitations counteract the effect of faking on personality inventory validity?," *Journal of Personality and Social Psychology*, vol. 81, no. 1, pp. 160–169, 2001.

[18] A. Foerster, R. Pfister, C. Schmidts, D. Dignath, and W. Kunde, "Honesty saves time (and justifications)," *Frontiers in Psychology*, vol. 4, p. 473, 2013.

[19] M. Monaro, L. Gamberini, and G. Sartori, "Spotting faked identities via mouse dynamics using complex questions," in *In Proceedings of the 32nd International BCS Human Computer Interaction Conference (p. 8)*, BCS Learning & Development, 2018.

[20] P. Roma, M. C. Verrocchio, C. Mazza et al., "Could time detect a faking-good attitude? A study with the MMPI-2-RF," *Frontiers in Psychology*, vol. 9, p. 1064, 2018.

[21] G. Sartori, A. Zangrossi, and M. Monaro, "Deception detection with behavioral methods: the autobiographical implicit association test, concealed information test–reaction time, mouse dynamics, and keystroke dynamics," in *Detecting concealed information and deception*, pp. 215–241, Academic Press, 2018.

[22] N. L. Vasilopoulos, R. R. Reilly, and J. A. Leaman, "The influence of job familiarity and impression management on self-report measure scale scores and response latencies," *Journal of Applied Psychology*, vol. 85, no. 1, pp. 50–64, 2000.

[23] P. Roma, C. Mazza, S. Mammarella, B. Mantovani, G. Mandarelli, and S. Ferracuti, "Faking-good behavior in self-favorable scales of the MMPI-2. A study with time pressure," *European Journal of Psychological Assessment*, vol. 36, no. 2, 2019.

[24] S. Shalvi, O. Eldar, and Y. Bereby-Meyer, "Honesty requires time-a reply to Foerster et al. (2013)," *Frontiers in Psychology*, vol. 4, p. 634, 2013.

[25] C. Mazza, M. Monaro, F. Burla et al., "Use of mouse-tracking software to detect faking-good behavior on personality questionnaires: an explorative study," *Scientific Reports*, vol. 10, no. 1, p. 4835, 2020.

[26] C. McKinstry, R. Dale, and M. J. Spivey, "Action dynamics reveal parallel competition in decision making," *Psychological Science*, vol. 19, no. 1, pp. 22–24, 2008.

[27] M. Monaro, F. I. Fugazza, L. Gamberini, and G. Sartori, "How human–mouse interaction can accurately detect faked responses about identity," *Lecture Notes in Computer Science*, vol. 9961, pp. 115–124, 2017.

[28] M. Monaro, L. Gamberini, and G. Sartori, "Identity verification using a kinematic memory detection technique," in *Advances in neuroergonomics and cognitive engineering*, pp. 123–132, Springer, 2017.

[29] M. Monaro, A. Toncini, S. Ferracuti et al., "The detection of malingering: a new tool to identify made-up depression," *Frontiers in Psychiatry*, vol. 9, p. 249, 2018.

[30] M. Monaro, C. Mazza, M. Colasanti et al., "Detecting faking-good response style in personality questionnaires with four choice alternatives," *Psychological Research*, vol. 85, no. 8, pp. 3094–3107, 2021.

[31] J. B. Freeman and N. Ambady, "MouseTracker: software for studying real-time mental processing using a computer mouse-tracking method," *Behavior Research Methods*, vol. 42, no. 1, pp. 226–241, 2010.

[32] R. Dale, C. Kehoe, and M. J. Spivey, "Graded motor responses in the time course of categorizing atypical exemplars," *Memory & Cognition*, vol. 35, no. 1, pp. 15–28, 2007.

[33] J. B. Freeman, N. Ambady, N. O. Rule, and K. L. Johnson, "Will a category cue attract you? Motor output reveals dynamic competition across person construal," *Journal of Experimental Psychology: General*, vol. 137, no. 4, pp. 673–690, 2008.

[34] J. Freeman, R. Dale, and T. Farmer, "Hand in motion reveals mind in motion," *Frontiers in Psychology*, vol. 2, p. 59, 2011.

[35] J. H. Song and K. Nakayama, "Target selection in visual search as revealed by movement trajectories," *Vision Research*, vol. 48, no. 7, pp. 853–861, 2008.

[36] M. J. Spivey, M. Grosjean, and G. Knoblich, "Continuous attraction toward phonological competitors," *Proceedings of the National Academy of Sciences*, vol. 102, no. 29, pp. 10393–10398, 2005.

[37] F. Ales, L. Giromini, L. Warmelink et al., "On the use of eye movements in symptom validity assessment of feigned schizophrenia," *Psychological Injury and Law*, vol. 16, no. 1, pp. 83–97, 2023.

[38] F. Ales, L. Giromini, L. Warmelink et al., "An eye tracking study on feigned schizophrenia," *Psychological Injury and Law*, vol. 14, no. 3, pp. 213–226, 2021.

[39] S. Berkovsky, R. Taib, I. Koprinska et al., "Detecting personality traits using eye-tracking data," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, Scotland Uk, 2019.

[40] F. Chen, N. Ruiz, E. Choi et al., "Multimodal behavior and interaction as indicators of cognitive load," *ACM Transactions on Interactive Intelligent Systems (Tii S)*, vol. 2, no. 4, pp. 1–36, 2012.

[41] X. Fang, Y. Sun, X. Zheng, X. Wang, X. Deng, and M. Wang, "Assessing deception in questionnaire surveys with eye-tracking," *Frontiers in Psychology*, vol. 12, 2021.

[42] A. Mirsadikov and J. George, "Can you see me lying? Investigating the role of deception on gaze behavior," *International Journal of Human-Computer Studies*, vol. 174, article 103010, 2023.

[43] E. A. Van Hooft and M. P. Born, "Intentional response distortion on personality tests: using eye-tracking to understand response processes when faking," *Journal of Applied Psychology*, vol. 97, no. 2, pp. 301–316, 2012.

[44] K. Rayner, "The 35th Sir Frederick Bartlett lecture: eye movements and attention in reading, scene perception, and visual search," *Quarterly Journal of Experimental Psychology*, vol. 62, no. 8, pp. 1457–1506, 2009.

[45] R. A. Abrams, D. E. Meyer, and S. Kornblum, "Speed and accuracy of saccadic eye movements: characteristics of impulse variability in the oculomotor system," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 3, pp. 529–543, 1989.

[46] K. Rayner, "Eye movements in reading and information processing," *Psychological Bulletin*, vol. 85, no. 3, pp. 618–660, 1978.

[47] I. Bacivarov, M. Ionita, and P. Corcoran, "Statistical models of appearance for eye tracking and eye-blink detection and measurement," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 3, pp. 1312–1320, 2008.

[48] J. Lee and J. H. Ahn, "Attention to banner ads and their effectiveness: an eye-tracking approach," *International Journal of Electronic Commerce*, vol. 17, no. 1, pp. 119–137, 2012.

[49] M. J. Tsai, H. T. Hou, M. L. Lai, W. Y. Liu, and F. Y. Yang, "Visual attention for solving multiple-choice science problem: an eye-tracking analysis," *Computers & Education*, vol. 58, no. 1, pp. 375–385, 2012.

[50] J. Zagermann, U. Pfeil, and H. Reiterer, "Measuring cognitive load using eye tracking technology in visual computing," in *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 78–85, Baltimore, MD, USA, 2016.

[51] J. Z. Lim, J. Mountstephens, and J. Teo, "Emotion recognition using eye-tracking: taxonomy, review and current challenges," *Sensors*, vol. 20, no. 8, p. 2384, 2020.

[52] L. Perkhofer and O. Lehner, "Using gaze behavior to measure cognitive load," in *Information systems and neuroscience*, pp. 73–83, Springer, 2019.

[53] W. L. Zheng, B. N. Dong, and B. L. Lu, "Multimodal emotion recognition using EEG and eye tracking data," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 5040–5043, IEEE, Chicago, IL, USA, 2014.

[54] M. Keskin, K. Ooms, A. O. Dogru, and P. De Maeyer, "EEG & eye tracking user experiments for spatial memory task on maps," *ISPRS International Journal of Geo-Information*, vol. 8, no. 12, p. 546, 2019.

[55] M. Keskin, K. Ooms, A. O. Dogru, and P. De Maeyer, "Exploring the cognitive load of expert and novice map users using EEG and eye tracking," *ISPRS International Journal of Geo-Information*, vol. 9, no. 7, p. 429, 2020.

[56] P. van der Wel and H. van Steenbergen, "Pupil dilation as an index of effort in cognitive control tasks: a review," *Psychonomic Bulletin & Review*, vol. 25, no. 6, pp. 2005–2015, 2018.

[57] A. Maffei and A. Angrilli, "Spontaneous eye blink rate: an index of dopaminergic component of sustained attention and fatigue," *International Journal of Psychophysiology*, vol. 123, pp. 58–63, 2018.

[58] A. Vrij, J. Oliveira, A. Hammond, and H. Ehrlichman, "Saccadic eye movement rate as a cue to deceit," *Journal of Applied Research in Memory and Cognition*, vol. 4, no. 1, pp. 15–19, 2015.

[59] S. Leal and A. Vrij, "The occurrence of eye blinks during a guilty knowledge test," *Psychology, Crime & Law*, vol. 16, no. 4, pp. 349–357, 2010.

[60] F. M. Marchak, "Detecting false intent using eye blink measures," *Frontiers in Psychology*, vol. 4, 2013.

[61] C. Mazza, F. Burla, M. C. Verrocchio et al., "MMPI-2-RF profiles in child custody litigants," *Frontiers in Psychiatry*, vol. 10, p. 725, 2019.

[62] R. K. Otto, "Use of the MMPI-2 in forensic settings," *Journal of Forensic Psychology Practice*, vol. 2, no. 3, pp. 71–91, 2002.

[63] P. Roma, F. Ricci, G. D. Kotzalidis et al., "MMPI-2 in child custody litigation," *European Journal of Psychological Assessment*, vol. 30, no. 2, pp. 110–116, 2014.

[64] P. Roma, E. Piccinni, and S. Ferracuti, "Applicazioni forensi del MMPI-2. [using MMPI-2 in forensic assessments]," *Rassegna Italiana di Criminologia*, vol. 10, no. 2, pp. 116–122, 2016.

[65] P. Pancheri and S. Sirigatti, *MMPI-2 – Minnesota Multiphasic Personality Inventory –2*, Manuale. Giunti O.S Organizzazioni Speciali, 1995.

[66] S. Sirigatti and C. Stefanile, *MMPI-2: Aggiornamento all'adattamento Italiano*, Giunti OS Organizzazioni Speciali, Florence, Italy, 2011.

[67] A. B. Tobii Pro, *Tobii Pro Lab (Version 1.145) [Computer Software]*, Tobii Pro AB, Danderyd, Sweden, 2014.

[68] J. Cohen, *Statistical power analysis for the behavioral sciences*, Routledge, 2013.

[69] J. T. Richardson, "Eta squared and partial eta squared as measures of effect size in educational research," *Educational Research Review*, vol. 6, no. 2, pp. 135–147, 2011.

[70] IBM Corp, *IBM SPSS Statistics for Windows, Version 25.0*, IBM Corp, Armonk, NY, 2017.

[71] Y. R. Muhsen, S. L. Zubaidi, N. A. Husin, A. Alnoor, D. Božanić, and K. S. Hashim, "The weight fuzzy judgment method for the benchmarking sustainability of oil companies," *Applied Soft Computing*, vol. 161, article 111765, 2024.

[72] M. T. Ali, Y. R. Muhsen, R. F. Chisab, and S. N. Abed, "Evaluation study of radio frequency radiation effects from cell phone towers on human health," *Radioelectronics and Communications Systems*, vol. 64, no. 3, pp. 155–164, 2021.

[73] C. Mazza, M. Monaro, G. Orrù et al., "Introducing machine learning to detect personality faking-good in a male sample: a new model based on Minnesota Multiphasic Personality Inventory-2 Restructured Form scales and reaction times," *Frontiers in Psychology*, vol. 10, 2019.

[74] S. Zago, E. Piacquadio, M. Monaro et al., "The detection of malingered amnesia: an approach involving multiple strategies in a mock crime," *Frontiers in Psychiatry*, vol. 10, p. 424, 2019.

[75] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[76] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer-Verlag, 2009.

[77] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015.

[78] D. B. Dwyer, P. Efallkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual Review of Clinical Psychology*, vol. 14, no. 1, pp. 91–118, 2018.

[79] G. Orrù, M. Monaro, C. Conversano, A. Gemignani, and G. Sartori, "Machine learning in psychometrics and psychological research," *Frontiers in Psychological Research*, vol. 10, 2019.

[80] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou et al., "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Scientific Reports*, vol. 5, no. 1, pp. 1–12, 2015.

[81] M. A. Hall, *Correlation-based feature selection for machine learning*, Doctoral dissertation, The University of Waikato, 1999.

[82] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 1137–1143, Morgan Kaufman Publishing, 1995.

[83] S. le Cessie and J. C. van Houwelingen, "Ridge estimators in logistic regression," *Applied Statistics*, vol. 41, no. 1, pp. 191–201, 1992.

[84] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to Platt's SMO algorithm for SVM classifier design," *Neural Computation*, vol. 13, no. 3, pp. 637–649, 2001.

[85] L. Breiman, "Random forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[86] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in *In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pp. 338–345, Morgan Kaufmann Publishers Inc., San Francisco, California, 1995.

[87] J. S. Quinlan, *C4.5: programs for machine learning*, Morgan Kaufmann, 1993.

[88] T. Mitchell, "Decision tree learning," in *Machine learning*, T. Mitchell, Ed., McGraw Hill, 1997.

[89] B. Verschuere, N. C. Köbis, Y. Bereby-Meyer, D. Rand, and S. Shalvi, "Taxing the brain to uncover lying? Meta-analyzing the effect of imposing cognitive load on the reaction-time costs of lying," *Journal of Applied Research in Memory and Cognition*, vol. 7, no. 3, pp. 462–469, 2018.

[90] R. R. Holden and D. G. Kroner, "Relative efficacy of differential response latencies for detecting faking on a self-report measure of psychopathology," *Psychological Assessment*, vol. 4, no. 2, pp. 170–173, 1992.