Head Office: Università degli Studi di Padova

Department of General Psychology

Ph.D. COURSE IN: Brain, Mind and Computer Science

CURRICULUM: Neuroscience, Technology and Society
SERIES: XXXVI

UNMASKING LIES:

ADVANCEMENTS IN DECEPTION DETECTION TECHNIQUES

Coordinator: Ch.ma Prof.ssa Anna Spagnolli

Supervisor: Ch.mo Prof. Massimo Grassi

Co-Supervisor: Ch.mo Prof. Giuseppe Sartori
Co-Supervisor: Ch.mo Prof. Lamberto Ballan

Ph.D. student: Dott.ssa Giulia Melis

Table of Contents

Table of Contents.....3

Synopsis.....7

Chapter 1. Introducing Deception Detection.....11

1.1 Definition and Prevalence.....11

1.2 Development of Techniques to Detect Deception.....14

1.2.1 Historical Evolution of the Methods.....15

1.2.2 Emotion-Based Lie Detection Techniques.....16

1.2.3 Cognitive Lie Detection Techniques.....19

1.2.4 Artificial Intelligence in Detecting Deception.....24

1.2.5 Neuropsychological and Psychological Malingering Detection.....26

Concluding Note.....33

Chapter 2. Detecting Deception at Single Item Level in Questionnaire.....35

2.1 Abstract.....35

2.2 Introduction.....36

2.3 Materials and Methods.....38

2.3.1 Proposed Approach.....38

2.3.2 Participants.....41

2.3.3 Procedure.....41

2.3.4 Faking Instructions.....42

2.4 Results and Discussion.....44

2.4.1 TF-IDF for Item Analysis.....46

2.4.2 Performance Evaluation.....48

2.5 Conclusion.....53

Chapter 3. Drugs Don't Work.....57

3.1 Abstract.....57

3.2 Introduction.....58

3.2.1 Drug Use in the Workplace 58
3.2.2 The Current Study 59
Experiment 1 60
3.3 Materials and Methods 60
3.3.1 Experimental Procedure 60
3.3.2 Hypotheses 65
3.3.3 Collected Measures 65
3.3.4 Exclusion Criteria 66
3.3.5 Participants 67
3.3.6 Data Pre-Processing 68
3.3.7 Features Computation 69
3.4 Data Analysis 69
3.4.1 Visual Comparison of Trajectories 69
3.4.2 Features Selection 70
3.4.3 Statistical Analysis 71
3.4.4 Classification Task 72
3.5 Discussion 74
3.6 Conclusion 75
Experiment 2 76
3.7 Materials and Methods 76
3.7.1 Experimental Procedure 76
3.7.2 Hypotheses 79
3.7.3 Collected Measures 79
3.7.4 Exclusion Criteria 79
3.7.5 Participants 79
3.7.6 Data Pre-Processing and Features Computation 80
3.8 Data Analysis 80

3.8.1 Visual Comparison of Trajectories 80
3.8.2 Features Selection 81
3.8.3 Statistical Analysis 82
3.8.4 Classification Task 85
3.9 Discussion 87
3.10 Conclusion 90
Chapter 4. Uncovering Lies During an Interrogation 91
4.1 Abstract 91
4.2 Introduction 92
4.2.1 Clues of Lies 92
4.2.2 Identity Deception 94
4.3 Materials and Methods 96
4.3.1 Participants 96
4.3.2 Experimental Procedure 97
4.4 Data Analysis 98
4.5 Results 100
4.5.1 Latencies Analysis 100
4.5.2 Errors Analysis 101
4.5.3 Questionnaires Analysis 102
4.5.4 Detection of Deception at Single Subject Level 103
4.5.5 Detection of Deception in Single-Faked Answers 105
4.5.6 Relative Measures 106
4.6 Discussion 108
4.7 Conclusion 113
Conclusion 115
Supplementary Materials 119
Bibliography 121

Synopsis

Pursuing truth through the unmasking of deception has consistently engaged the scientific community and the general public. This interdisciplinary interest, encompassing fields such as sociology, psychology, linguistics, and artificial intelligence, presents a compelling array of challenges and discoveries. Given their complex social, psychological, and communicative implications, analyzing lies is more than an intellectual challenge; it's a key to decoding the subtle mechanisms that govern the intricate dynamics of human interactions.

This work aims to explore and integrate diverse methodologies and theories in lie detection. It provides a comprehensive overview that underscores not only the challenges and potential but also some ethical considerations in studying deceit within our contemporary social framework.

Structured into four pivotal chapters, this dissertation introduces three approaches to lie detection, each uniquely designed and tested in distinct contexts: self-report assessment questionnaires, drug testing in the workplace, and face-to-face investigative interviews. Each methodology confronts the challenges of detecting lies in these specific settings, striving to surpass the limitations of existing techniques. The objective is to propose tailored solutions for each domain, thereby enhancing the effectiveness of lie-detection practices.

The first chapter establishes the theoretical and methodological foundation for understanding deception, particularly in forensic contexts. It underscores the prevalent challenge of identifying falsehoods, thereby setting the stage for the methodologies introduced later. The chapter begins with exploring the definition and prevalence of deception, laying the groundwork for the historical evolution of lie detection techniques in psycho-forensic assessments. This exploration starts with Vittorio Benussi's pioneering experiments and the employment of the polygraph in the early twentieth century, advancing to later developments focusing on emotion analysis and cognitive approaches. It then discusses the application of machine learning techniques, a pivotal advancement in enhancing traditional methods, and concludes by focusing on specific procedures for identifying feigned neuropsychological or psychological disorders, emphasizing some of the most established techniques and tools in this domain.

This multidisciplinary approach offers a historical and current perspective on deception detection methodologies, focusing on the techniques and theories applied in the three subsequent research projects. These projects represent three independent and somewhat innovative approaches, each introducing either a novel lie detection technique or using existing

methodologies in previously unexplored contexts. A key aspect of each study is machine learning analysis for classification tasks aimed at distinguishing between truth and lies. Depending on the study's objectives and practical applications, the developed models are designed to detect individual liars or identify specific details about which subjects choose to lie. The strategic application of machine learning demonstrates its efficacy in addressing the complexities of deceit in various contexts.

The second chapter introduces the first research project, which explores the use of the Term Frequency-Inverse Document Frequency (TF-IDF) index to detect lies within personality questionnaires. Departing from traditional methods that construct comprehensive simulation profiles through control scales, this approach is dedicated to the detailed analysis of specific responses to identify instances of deceit at the individual item level. The project tests the efficacy of TF-IDF in discerning between authentic and fabricated responses, offering a novel analytical tool for forensic psychology. To validate this methodology, three experiments were conducted in "faking good" scenarios, where participants are incentivized to present themselves in an overly positive manner. The results were notably promising: the TF-IDF model effectively distinguished between genuine and counterfeit responses. These findings open new perspectives for research in forensic psychology, suggesting that TF-IDF could be a valuable tool for addressing complex challenges like detecting deception in questionnaires at the item level.

The third chapter provides a preliminary analysis of drug use in the workplace and techniques for detecting this behavior. A discussion on the consumption of psychoactive substances in the workplace and the related challenges highlights the impact on individual employee well-being and economic costs for employers and society, exploring standard drug testing methods, their legislative variations, and the connected moral, ethical, and legal concerns. The chapter assesses the effectiveness and limitations of questionnaire-based tests in predicting or confirming substance abuse and their limits due to transparency and ease of providing deceptive answers. This section acknowledges the limitations and challenges of current methodologies, highlighting the need for more innovative and accurate approaches to address a growing relevant issue in workplace safety. In response to this need, the research presented in this project focuses on introducing and validating a methodology already known in forensic research: kinematic analysis of mouse movements. This method is based on the hypothesis that mouse movements during responses to double-choice questions can reveal behavioral patterns associated with lying. This methodology is applied here in a context that has never been inserted before, namely, to detect deceptive responses in workplace drug tests.

The research, conducted through two computer-generated online experiments, compares behaviors of drug-using and non-drug-using employees, exploring the potential of this analysis to understand the mental processes involved in generating deceitful responses.

The fourth chapter delves into identity deception, a specific deceit form where individuals intentionally conceal their identity, impersonate others or use counterfeit identity documents. This type of deceit is prevalent and seriously threatens national security. Furthermore, the digital age has exacerbated this challenge, with frequent alterations observed in attributes like images, names, ages, and genders, making it easier for terrorists and criminals to evade security protocols, as demonstrated by using false identities in terrorist attacks. To address this challenge, this research project employed techniques such as the method of the unexpected questions, utilizing the difficulty liars face in responding to unforeseen queries, and reaction time analyses in face-to-face investigative interviews. Machine learning analyses in this study offer insights into detecting individuals who lie and understanding the specific questions they deceive about. The dissertation ends with a general conclusion on the findings of the individual projects.

In summary, these chapters highlight the challenges of deception and outline advancements in computational tools to enhance the accuracy of deception detection in various contexts.

Chapter 1.
Introducing Deception Detection

Based on Sartori, G., & Melis, G. (2022). Deception in Court. *Medicina nei secoli: Journal of history of medicine and medical humanities*, 34(2), 71-102. Published under a CC BY SA license.

1.1 Definition and Prevalence

"Everybody lies. Cops lie. Lawyers lie. Witnesses lie. The victims lie. A trial is a contest of lies. And everybody in the courtroom knows this. The judge knows this. Even the jury knows this. They come into the building knowing they will be lied to. They take their seats in the box and agree to be lied to."

(Michael Connelly, 2008)

"Everybody lies," to a greater or lesser degree, is the assumption many people come to, so much so that at the foundation of what we are with others, there is often deception. Hence, any moral, ethical, or legal system should consider this intrinsic aspect of human nature (Ganis & Keenan, 2009). Drawing from evolutionary and psychosocial studies, lying has been deeply embedded in all societies since ancient times. This behavior is observed even in animal species. Consider, for instance, strategies they employ to survive in environments with limited resources, such as imitation or camouflage, which animals use to avoid becoming prey to predators (Ganis & Keenan, 2009). By the age of two, we can tell lies to deny a behavior we know to be incorrect; however, we must wait 4 or 5 years to make these lies more credible and consistent (Evans et al., 2011). Once we become adults, we let ourselves go, even going so far as to say three lies every ten minutes and in one of three social situations (Anolli, 2003). Based on some statistics, 80% of these lies are partly about the person telling them, and two-thirds benefit the liar more than others (DePaulo et al., 1996).

Believing deception is a modern-era invention is incorrect. We can reconstruct how simulation was one of the most adopted modes in the 2nd century CE or the Greco-Roman and Renaissance periods. Odysseus himself feigned mental illness to avoid the Trojan War,

pretending to be a farmer and not recognizing his friends. Also, the physician Galeno recorded a second example of malingering in Roman times. In a short pamphlet entitled *Quomodo morbum simulantes sint deprehendendi* (Kühn, 1830), Galeno reported the case of two patients who faked to be ill. In the first case, the patient faked colic to avoid a public meeting, while in the latter, the patient faked an injured knee to avoid accompanying his master on a long journey and staying at home with his partner.

It is easy to see how, on many occasions, given the appropriate reasoning and social skills, the rewards of deception can outweigh the costs associated with its discovery: producing a lie requires more mental resources than simply stating the truth, but if it enables the liar to achieve something highly sought then their efforts will be compensated. Confirming this, there have been many physical or mental faked illnesses throughout history, ranging from simulated psychopathology to amnesia. Famous for the latter reference is the Italian Case of the Collegno amnesic who malingered retrograde amnesia and spent nearly a year in Collegno without a name (Zago et al., 2004) or that of Rudolph Hess, a Nazi hierarch who claimed not to remember his role in the Third Reich period during the Nuremberg trial. A group of psychiatrists examined Hess and concluded that his amnesia was genuine. Subsequently, he declared: *"My memory is again in order. The reasons why I simulated loss of memory were tactical."* Whether to avoid a call to arms, escape the death penalty, or avoid prison, simulation is an attractive solution to social, economic, and personal problems (Lezak et al., 2004). Undoubtedly, the legal setting is a context where malingering behaviors creep in, reaching 25-45% of faking attitudes (Stracciari et al., 2010). Mittenberg and colleagues' work from 2002 (Mittenberg et al., 2002) provided a survey of 388 members of the American Board of Clinical Neuropsychologists concerning the frequency of malingering in personal injury claims (30%), disability or worker's compensation (32%), criminal cases (22%), medical or psychiatric cases (22%). These results suggest that rates of probable malingering varied depending on the presumptive diagnosis; moreover, it differed for civil and criminal cases. However, estimating the prevalence of simulation is challenging for several reasons; for instance, it will always be an underestimation of the phenomenon since the percentage of successful fakers does not emerge, which, by definition, remains excluded from the statistics.

In 2009, Abe (Abe, 2009) defined deception as a psychological process by which one individual deliberately attempts to convince another person to accept as true what the liar knows to be false to gain some type of benefit or to avoid loss. Although, as is often the case when trying to define a phenomenon, there is no universal definition, many consider lying an on-off process, where there is lying or not. On the contrary, lying is a continuous variable constantly

influenced by the social situation, the strategies adopted, the degree of awareness, the motivations of the deceiver, and more (Anolli, 2003; Ellingson & McFarland, 2011). Based on the specific scenario of legal presentation and the form in which advantages and gains come, deception may take on various facets: a subject willing to deceive can report symptoms that, in reality, do not exist or can accentuate signs living in the past, which have faded. Generally, it can be of two types: the individual can exaggerate generic symptoms belonging to various psychopathological areas, such as anxiety or depression (generalized malingering), or symptoms related to a particular disease (specific malingering) (Stracciari et al., 2010). One of the most helpful discriminations to mention is between "fake good" and "fake bad." On the one hand, a subject may be interested in faking an organic/mental disorder or cognitive impairment in a civil or criminal setting to gain compensation or a reduction in legal penalty; this attitude is known as faking bad or malingering. On the other hand, the deceiver may also attempt to present a more desirable version of himself to hide inconvenient features that could disfavor him (e.g., in litigation for child custody in favor of the other parent), so-called faking good or dissimulation.

Regarding faking bad, DSM-5 (American Psychiatric Association, A. P. A., 2013) describes malingering as *"the intentional production of false or grossly exaggerated physical or psychological symptoms, motivated by external incentives. Under some circumstances, malingering may represent adaptive behavior, for example, feigning illness while a captive of the enemy during wartime."*

When trying to spot deception, the expert must first ask himself whether the subject might be motivated to fake from the context, taking note of two key features:

a) the intentionality of the subject in the production of symptoms;
b) external incentives associated with the behavior (Stracciari et al., 2010).

Dissimulation, or faking good, is not mentioned in DMS-5 but can be described as the mirror of malingering: *"the tendency to give positive self-description," "claiming improbable virtues and denying common human frailties"* (Braun et al., 2001; Crowne et al., 1964). As malingering, fake good might be present in several forms, such as hiding or minimizing undesired characteristics and exaggerating or making up positive qualities to pursue a personal goal. In criminal law, faking good is frequent when evaluating social dangerousness. In contrast, it is common in civil law when assessing driving skills in the renewal-license setting or parental skills in the child custody context. For example, in everyday life, faking good is frequent during a job interview, where the candidate wants to show a more desirable version of himself to

increase the probability of obtaining the job. Detecting these efforts has become a priority in the field, feeding growing activity to develop incrementally objective and efficacious tools. Despite its relevance, scientific literature, and instruments to identify dissimulation, are unfortunately limited and insufficient. Malingering behaviors probably attracted more research due to their welfare and social costs (e.g., insurance compensation).

Despite all the progress made in pursuing this goal, it must be emphasized that no technique or tool can be considered entirely objective. Each of these strategies and techniques measures the behavior of an individual who is deliberately lying and is intent on doing so based on several factors that influence his behavior (cooperation, motivation, secondary gains, personality) (Stracciari et al., 2010). Consequently, any results should be considered in probabilistic terms since malingering may only be sure when a) the malingerer is discovered while performing the psychic or physical function claimed to be impaired and b) when he confesses (Hurst, 1940). In addition, malingering should be strongly suspected when any possible combination of these factors is presented: 1) symptoms occur in a medico-legal setting; 2) a marked discrepancy between the individual's claimed impairment and the objective findings; 3) there is a lack in collaboration during the assessment and low compliance with the prescribed treatment regimen; 4) presence of antisocial personality disorder (*ibidem*).

In the forensic context, particularly in forensic neuropsychology, it is good practice to identify the conscious intentionality of the subject and the existence of external incentives. If, in a clinical setting, it is usual to go along with the symptoms that the patient reports, instead, in the forensic context, it is crucial first to take into account the two points listed above. Considering these parameters help the clinician to become a wise expert. Depending on the form of simulation, intentionality may be absent or accentuated. The presence of consciousness about the planning of the disorder (challenging to ascertain) should be a criterion for identifying the symptoms of deception. Moreover, many times, the external advantage may not be immediately recognizable. When doubt exists, it is essential to look in the case history for economic benefits that are not evident at first glance (*ibidem*).

1.2 Development of Techniques to Detect Deception

The ambition to identify deception through observing human behavior is a fascinating aspect that captivates many. There is a widespread belief in the existence of distinguishable differences in the behaviors of truthful individuals and deceivers. Typical assumptions include the notion that deceivers tend to avert their gaze more frequently than their truthful counterparts

(Vrij, Akehurst, et al., 2006) or exhibit signs of nervousness, leading to a range of behaviors considered indicative of deception, such as increased body movements (Hartwig et al., 2015). This has prompted numerous researchers to investigate, from an empirical standpoint, the existence of reliable indicators of deception, with the ultimate goal of developing strategies for deception detection. However, this remains a challenging issue for examiners (Sartori et al., 2016); researchers themselves seem to agree only minimally on the reliable cues to deception (Luke et al., 2023).

Despite many might believe that the performances of experts, such as police officers, investigators, judges, psychologists, or psychiatrists, in detecting deception are remarkably accurate, scientific research shows that, unfortunately, they are only slightly better than chance (Kassin & Gudjonsson, 2004), with accuracy rates between 45% and 60%, averaging 54% (Vrij, 2000, 2004). A meta-analysis highlights that the accuracy of deception judgment without any aid or training is approximately 50% (Bond Jr & DePaulo, 2006). This level of accuracy has the same odds to provide a judgment by chance. This evidence suggests that it is tough to assess malingering by relying on personal judgment and is often inadequate when merely based on an interview.

Given the significant challenge and importance of detecting lies, the scientific community has been motivated to develop and refine various methods to improve lie detection. Starting with the following paragraph, this section explores the evolution of these methods within the forensic context. It traces the journey from the first lie detector devised by Vittorio Benussi in 1914 to contemporary advancements incorporating Artificial Intelligence. This exploration introduces the tools and techniques employed to enhance lie detection accuracy.

1.2.1 Historical Evolution of the Methods

Over time, the first methods for detecting lies developed based on the disorder that was being simulated, such as psychiatric or cognitive disorder. In the case, for example, of faked psychiatric disorders, the literature of the late 1800s and early 1900s reported applicable criteria for deception. An interesting work is that written by Roncoroni in 1987, who, through the study of some inmates in prisons, identifies four criteria that together would diagnose the simulation of mental illness:

1) anamnesis,
2) general anatomical and physiological criteria,
3) general psychological criteria, and

4) specific criteria for individual mental illnesses (Roncoroni, 1987).

One of the first attempt to develop instruments capable of detecting faked cognitive disorder was born between the 1800s and 1900s with the Ziehen's Test, used to identify a simulator in a short-term memory task. This test is based on the principle that a simulation attempt should be considered if a person cannot complete a task that an actual patient with, for example, dementia can do. Other techniques born in this period that will be later described are the 15 Rey Items Test and the Floor Effect Strategy (Rey, 1958; Rogers et al., 1993). Alfredo Coppola, one of the clinicians who evaluated the Collegno Amnesic case and the only one who diagnosed *"malingered retrograde amnesia,"* also used these methods (Zago et al., 2004). Over time, the *"cunning means,"* introduced initially by Ingegnieros (1904), was defined by the same author as an adjunct to the examination of the subject and *"the best psychological reagent to the simulation."* It involved, for example, mentioning to the simulator a false symptom not manifested by him and seeing if it would be referred to within a short time. In 1905, with *"La Perizia Psichiatrico-Legale,"* Cesare Lombroso argued that the general characteristic of simulation is the exaggeration of the symptoms that the simulator compares with those of a real subject harboring pathology (Lombroso and Leggiardi-Laura, 1905). It is precisely these symptoms that Lombroso focused on exposing the lie, making himself directly available to the judge to detect significant psychophysiological changes during interrogation (Zago, S., Difonzo, 2018). It would be about ten years later that Vittorio Benussi (1914) would introduce the first polygraph.

1.2.2 Emotion-Based Lie Detection Techniques

Some researchers suggest a correlation between deception and increased signs of nervousness and agitation (DePaulo et al., 2003) and posit that the most straightforward and potentially effective method for lie detection might be the analysis of behavior, encompassing both verbal and nonverbal cues (Ganis and Keenan, 2009). This hypothesis stems from two primary factors:

1. the fear of being discovered, which has also been described as responsible for cues to deception (Ekman et al., 1985), producing higher physiological activation;
2. The substantial cognitive effort that would require a lie, so much so that initially, the research aimed to discover techniques to identify the lie focused precisely on

observing verbal and non-verbal behavior that accompanied the deceptive statements.

Consequently, studies have increasingly concentrated on detecting and monitoring psychophysiological variables. If emotions typically associated with deceptive contexts, such as anxiety, tension, and the fear of discovery, manifest in observable behavioral changes, then quantifying and tracking these changes could potentially reveal evidence of deception. This approach assumes that the physical expression of these emotions, particularly in the intensity of behavioral responses, could serve as indicators or "traces" left by lying, thereby aiding in accurate lie detection.

In 1914, Vittorio Benussi pioneered the development of the polygraph, an emotion-based device designed to detect deception by simultaneously recording various physiological responses such as blood pressure, heart rate, and breathing patterns, particularly during a subject's engagement in specific tasks (Benussi & Antonelli, 2002). The polygraph's significant advantage lies in its ability to measure these reference indices in real time as the participant undergoes a test.

Subsequent developments in lie detection technologies, inspired by the rationale behind the polygraph, have evolved over the years, focusing on the analysis of non-verbal behavior. Among the leading experts in this field is Paul Ekman, who posits that lie detection should concentrate more on non-verbal cues than verbal statements. According to Ekman, words can be manipulated easily; therefore, relying solely on verbal communication can lead to misinterpretation. Conversely, non-verbal behavior, such as body movements and facial expressions, is more difficult to control. This challenge arises because, when lying, an individual focuses primarily on fabricating the lie, often neglecting the involuntary non-verbal cues that may betray them.

Ekman's research suggests that the behavioral manifestation of emotions, mainly through facial expressions, is innate and universal, not culturally learned. In the 1970s, he conducted studies with an isolated tribe in New Guinea and identified six basic universal expressions: joy, sadness, fear, disgust, anger, and surprise, later expanding this list. Together with Maureen O'Sullivan, Ekman initiated studies on lying that focused on these expressions, leading to the discovery of facial micro-expressions—brief involuntary facial movements lasting between 1/25 to 1/2 of a second. In 1974, Ekman and his colleague Wally Friesen developed the Facial Action Coding System (FACS), a comprehensive tool for identifying and classifying facial micro-expressions by breaking them down into individual Action Units (AUs)

(Ekman & Friesen, 1978). The authors assert that concealing a genuine emotion is challenging, often resulting in its inadvertent expression (Ekman & Friesen, 1972). Unlike truth-tellers, individuals who lie tend to exhibit fewer micro-expressions and gestures due to the cognitive load involved in deception, often leading to neglect of their gestural communication. However, recognizing and accurately interpreting these subtle facial expressions and micro-movements are not straightforward tasks. Bond Jr and DePaulo analyzed subjects' abilities to discern truth from lies, finding that 47% of lies and 61% of non-deceptive responses were correctly identified (Bond Jr & DePaulo, 2006). This finding underscores the complexities of lie detection and the need for expertise and precision in interpreting non-verbal cues.

In addition to facial expression analysis, other emotion-based techniques that rely on similar premises and warrant mention include heart rate analysis (Kugelmass & Lieblich, 1966), thermography (Warmelink et al., 2011), voice stress analysis (Horvath, 1982). Ongoing research continues to evaluate these techniques, and in some instances, they are utilized to record physiological responses during interrogation. However, like the polygraph, these methods possess some limitations.

Although the polygraph can detect physiological activation patterns, such patterns should not be directly equated with the act of lying (Saxe, 1991). For instance, the psychophysiological arousal triggered by increased heart rate cannot be conclusively linked to deception. A subject undergoing a polygraph test may exhibit an elevated heart rate due to the stress of the test itself or other unrelated factors rather than as a direct response to lying. Furthermore, not all individuals demonstrate physiological arousal in response to producing deceptive statements (Ganis & Keenan, 2009).

Techniques like the polygraph, which are based on physiological activation, do not measure a direct correlation to lying but rather assess a pattern of physiological response that any stressful event could elicit. The generalizability of the polygraph is problematic, as various subjective factors can contaminate the data, rendering it only partially reliable. The polygraph has been shown to have a high risk of false positives—erroneously categorizing honest individuals as liars—and a lower incidence of false negatives (Brett et al., 1986).

Given the limited scientific evidence supporting the accuracy of the polygraph and ethical concerns related to privacy invasion, its use has been significantly restricted. In fact, in some countries, the use of the polygraph has been legally prohibited and viewed as a violation of individual rights (Stracciari et al., 2010).

Since the 1980s, the advent of event-related potentials (ERPs) and the subsequent integration of functional magnetic resonance imaging (fMRI) in recent years have significantly

advanced research in various fields, particularly in neuroscience. This evolution has facilitated the development of sophisticated cognitive and neural investigation techniques, especially in understanding the neural mechanisms underlying deceptive responses (Davatzikos et al., 2005; Meixner & Rosenfeld, 2011). These methodologies represent a shift from traditional physiological measures to more focused analyses of specific brain components.

ERPs are brain potentials measured using scalp electrodes and record electrical brain activity in response to physical stimuli and psychological processes, including the preparation for motor activity. These potentials are essentially changes within the spontaneous electroencephalogram (EEG) in response to an event, either internal or external to the individual, indicating a brain response closely timed with a stimulus or event (Pennisi, 1998). The rationale of ERP techniques is that the recognition of infrequent, well-known events (e.g., crime scene details) can modulate brain potentials such as the P300 or that response conflict (e.g., inhibiting an honest response while producing a deceptive one) can affect the amplitude of medial frontal negativities. This approach offers significant advantages, as it is less susceptible to manipulation by the subject, making it a more reliable method than the polygraph, which is more prone to manipulation (Rosenfeld & Bessinger, 1990).

The second notable technique is fMRI, which enables the visualization of hemodynamic responses in the brain related to specific activities. This method is based on the principle that active brain areas experience an increase in blood flow (Ladavas, E., & Berti, 2002), allowing for the identification of regions receiving increased blood and oxygen supply during particular tasks. fMRI has been frequently utilized to investigate brain areas involved in deception. For instance, a study employed fMRI on participants instructed to lie during a task, revealing activation in the ventromedial prefrontal cortex, a response not observed in the control group (Fullam et al., 2009). However, despite the promising insights provided by these techniques, their application in legal contexts remains controversial. The potential for premature or imprudent use of these tools in judicial settings raises concerns, especially given that the results from such technologies still need to be fully validated and could lead to misinterpretation or misuse in legal proceedings.

1.2.3 Cognitive Lie Detection Techniques

The scientific community widely recognizes that deception demands greater cognitive resources compared to truth-telling, as evidenced in the literature (Christ et al., 2009; Vrij, Fisher, et al., 2006). Two robust and straightforward indicators of the cognitive load associated with deception are reaction times (RTs) and error rates (Verschuere, Spruyt, et al., 2011). This

concept stems from the hypothesis that more complex tasks necessitate longer completion durations than simpler ones. Consequently, deception, being more cognitively demanding than truth-telling, is likely to manifest in prolonged RTs and increased errors. Supporting this, a meta-analysis examining the relationship between reaction time and deception found that honest respondents typically respond more quickly (Maricuțoiu & Sârbescu, 2016), suggesting that the time taken to construct a lie results in an escalation in both errors and RTs (Walczyk et al., 2003). Based on this assumption, some researchers have incorporated response times into their experimental designs, for instance, in examining deceptive responses to the MMPI-2 (Mazza et al., 2020) or in conjunction with unexpected questions (Monaro et al., 2021). Noteworthy methodologies leveraging these metrics include the Guilty Knowledge Test (GKT) (Lykken, 1959) and the Autobiographical-IAT (aIAT), (Sartori et al., 2008).

In light of these considerations, several strategies have been implemented within various experimental paradigms with the aim of augmenting cognitive load. This endeavor seeks to enhance the salience of measures associated with deception. In other words, the underlying premise posits that increased task complexity corresponds to prolonged reaction times. If, for example, we asked the participant to switch between a task that requires answering some questions lying and one in which they must be honest, switching between tasks with opposite instructions produces a cost in terms of RTs and errors. The traditional example is the Stroop effect, where individuals are asked to identify the ink color in which a word is written. It becomes particularly challenging when the word itself is the name of a different color (e.g., the word “BLUE” in red ink). This difficulty arises because the automatic process of reading a word is more deeply ingrained than identifying the ink color (MacLeod, 1991). In 1927, Jersild published a study on the ability to shift attention between different tasks (Jersild, 1927). In this research, students had to work through a list of items repeating or alternating between tasks (adding, multiplying, form-naming, and so on). Jersild observed that when the tasks were similar yet contradictory (e.g., adding versus subtracting three from a two-digit number), the reaction times (RTs) were significantly longer compared to performing a single, repetitive task. In contrast, RTs remained stable or improved when the tasks were markedly different (e.g., subtracting and writing). Spector and Biederman later elucidated this phenomenon (Spector & Biederman, 1976), proposing that when tasks share no common elements, the items themselves can signal the appropriate task to execute: subtraction applies only to numbers, whereas writing involves words. However, a significant effect emerges when alternating tasks necessitates tracking the previously performed operation (addition or subtraction). This leads to a “Switch cost”—a delay in RTs and an increase in error rates immediately after a task switch. This effect

has shown potential in the realm of lie detection. Debey and colleagues conducted a study investigating the impact of switching between questions that required honest responses and those demanding deception related to activities performed in the laboratory (experiment 1) or autobiographical facts (experiment 2) (Debey et al., 2015). The findings revealed an increase in RTs and a fluctuation in accuracy when switching from lying to truth-telling and vice versa, with no significant difference between the two scenarios (from lies to truth or truth to lies).

Recent Approaches in Malingering Detection

Most previously mentioned techniques fall under the traditional approach to lie detection, which aims to identify simulation by analyzing generalized deceptive behaviors rather than specific ones. However, recent methodologies include tools to detect detailed fabricated information: the aIAT (Sartori et al., 2008) and TARA (Gregg, 2007).

Autobiographical IAT: The autobiographical Implicit Association Test (aIAT), (Sartori et al., 2008), as a variant of the well-known Implicit Association Test (IAT) (Greenwald et al., 1998), is a technique that enables the determination of the existence of a specific autobiographical memory trace within an individual's mind/brain. This represents a significant advancement, as it facilitates the assessment of which of two competing autobiographical events is factual.

The method involves a computerized categorization task incorporating stimuli from four categories. Two categories are logical, consisting of sentences that are either true or false at the moment of testing (e.g., *"I am in front of a computer"* or *"I am climbing a mountain"*). The other two categories include alternative versions of the autobiographical event under investigation (e.g., *"I went to Paris for Christmas"* or *"I went to London for Christmas"*), where only one is accurate. The aIAT comprises five classification blocks. In each trial, participants are presented with a stimulus at the center of a computer monitor and asked to classify it quickly and accurately by pressing one of two labeled keys. Blocks 1, 2, and 4 are single categorization blocks, while Blocks 3 and 5 are combined. In Blocks 1, 2, and 4, each response button is associated with only one category. In contrast, in the double blocks (3 and 5), each button categorizes sentences belonging to two different categories.

The aIAT operates on the principle that pairing a true autobiographical event with certainly true sentences should result in faster responses. Thus, the specific pattern of RTs in the double categorization blocks indicates which autobiographical event is true or false. The actual event is identified by faster RTs in the block where it shares the same motor response

with true sentences. In other words, when congruent sentences are paired (e.g., a truthful autobiographical event with an undoubtedly true sentence), responses are expedited. The critical comparison is between the average RTs in Blocks 3 and 5.

This computerized task has been validated through experiments ranging from laboratory to real-life situations, achieving an average classification accuracy of 92% (Stracciari et al., 2010). However, the technique has limitations, as informed individuals can manipulate the measure by controlling their response latency (Sartori et al., 2017).

Timed Antagonistic Response Alethiometer: Another computer-based task is the Timed Antagonistic Response Alethiometern, also known as TARA (Gregg, 2007). TARA is a timed, multi-block classification task that requires participants to quickly and accurately classify a series of statements as true or false by pressing one of two keys. It involves truth-tellers engaging in two alternating compatible classification tasks, while liars must employ contradictory strategies. Since incompatible classification is more cognitively demanding than compatible classification, dishonest respondents are expected to have longer reaction times if accuracy remains constant. In essence, longer average RTs indicate dishonesty, while shorter ones suggest honesty. Experimental studies using TARA have achieved an accuracy rate of approximately 85% in detecting liars.

As time progresses, a new generation of instruments has been developed, aiming to measure deception-related data while minimizing the awareness of the deceiver. These methods, known as “implicit” techniques, are distinguished by their ability to offer a more dynamic assessment of the cognitive processes undertaken by the subject (Koop & Johnson, 2011). This represents a practical and economical approach, which is certainly not achievable by solely collecting and analyzing reaction times. Recent advancements in this area have introduced innovative techniques that utilize mouse movement, hand-motor tracking, keyboard dynamics, or eye movements to study the continuous evolution of mental processes underlying behavior (Eckstein et al., 2017; Freeman et al., 2011; Rheem et al., 2018; Sirois & Brisson, 2014), including deception (Colasanti et al., 2023; Duran et al., 2010; Grimes et al., 2013; Walczyk et al., 2012). In a recent study, the authors present and demonstrate a novel approach for distinguishing between truthful and deceptive responses in computer-administered questionnaires. This method offers a combined approach of mouse tracking and eye tracking (Maleck & Gross, 2023)

Hand motor responses, particularly during computerized multiple-choice tasks, offer insight into various psychological dynamics. Unlike reaction time, a static index resulting from cognitive operations, hand movement provides a real-time, high-fidelity motor trace of ongoing

mental processes. For instance, to accurately capture hand movements during a subject's response to computerized tasks, it is feasible to track mouse dynamics (Freeman & Ambady, 2010). Mouse dynamics encompass all aspects of mouse movement, including spatial and temporal features. This technique has been increasingly applied across various fields, proving valuable in elucidating the cognitive complexity of subjects' responses.

These cutting-edge methods enhance the effectiveness of detecting deceit and open new avenues for research in forensic psychology and cybersecurity. For instance, Hibbeln investigated mouse movements in the context of online insurance fraud (Hibbeln et al., 2014). They characterized deception by increased normalized movement distance, reduced movement speed, elongated reaction times, and a higher frequency of left clicks. In 2010, Duran and colleagues conducted a study using a Nintendo Wii remote control to compare motor trajectories while participants engaged in tasks requiring truthful and deceptive responses. The analysis of arm movement trajectories revealed that deceptive responses could be differentiated from truthful ones based on motor start time, total reaction time, movement trajectory, and kinematic parameters such as velocity and acceleration (Duran et al., 2010). Monaro, Sartori, and Gamberini applied this paradigm to detect feigned identities using unexpected questions and mouse dynamics (Monaro, Gamberini, et al., 2017). In their experiments, participants were divided into "liars" and "truth-tellers" and responded to questions about their real or fake identities. The study found that RTs, mouse trajectories, and errors in response to unexpected questions effectively distinguish between liars and truth-tellers, achieving a high accuracy rate of approximately 95% in detecting the veracity of self-declared identities. Accordingly, kinematic analysis of the subject's response is a promising technique to use in forensics, both in cases where the truthfulness of the testimony needs to be evaluated and in insurance cases where forgeries are not easily detected.

Compared to other techniques (e.g., RT or psychophysiological technique), kinematic analysis of mouse movements has different advantages (Sartori et al., 2017):

1. It captures the cognitive complexity of stimulus processing through various indicators beyond reaction time.
2. It functions as an implicit lie detection technology.
3. The extensive range of motion features it analyzes are theoretically challenging to counteract through deception.
4. It may achieve high accuracy for specific symptoms rather than a broad spectrum of symptoms characteristic of a particular pathology.

5. The technology is cost-effective and does not require equipment beyond what is already used for computer interaction.
6. It is particularly suited for online lie detection.
7. The technique does not necessitate specialized expertise for its application, even at the screening stage.

1.2.4 Artificial Intelligence in Detecting Deception

A burgeoning area of research within this field is characterized by integrating advanced computer science technologies to address unresolved issues and enhance existing methodologies. An up-and-coming line of inquiry emerges from the synergistic collaboration between neuroscience, computer science, and bioengineering. The overarching goal is to develop a computerized system that labels patients with specific diagnoses based on objective criteria. In detecting somatic and psychiatric disorders, using objectively collected data, an ideal algorithm would determine whether a patient's reported symptoms are feigned or genuine (Sartori et al., 2017). While this objective has not yet been fully realized, specific techniques and methods show promising results in forensic applications. Notably, within the field of Artificial Intelligence (AI), Machine Learning (ML) stands out (Mitchell, 1997).

Machine Learning, a subset of AI, employs statistical methods to enhance the capability of an algorithm in identifying patterns within datasets. This approach may use mathematical algorithms to classify objects into distinct groups, predict outcomes for new data, or generate novel insights from empirical observations, such as patient data. It focuses on isolating and recognizing key features characterized by objective indices of deception and analyzing the interplay between observed variables to develop automated classification rules for the data.

Breiman et al. (2001) delineate the distinctions between statistics and ML. They explain that the traditional statistical approach presupposes a specific stochastic data model generating data, focusing evaluations on the fit of the data to this model. Since the 1940s, statistical inference based on data modeling has been a cornerstone in analyzing scientific experiments. In contrast, ML regards the data mechanism as unknown and prioritizes predictive accuracy over model adherence.

Statistics involves selecting a model that encapsulates our understanding of a system, a choice justified by the researcher's belief in its efficacy in capturing the system's essence. Conversely, ML necessitates selecting a predictive algorithm based on empirical performance, emphasizing learning patterns based on past effectiveness in analogous scenarios. Both

strategies are valuable in deriving biologically meaningful insights (Orrù et al., 2020). Orrù and colleagues (2020) claim that the integration of ML with conventional inferential statistics to achieve two key objectives:

a) Developing models that generalize and replicate results with new data, addressing concerns about the reliability of research findings in behavioral research (Pashler & Wagenmakers, 2012). They argue that incorporating ML-based analysis into the analytical workflow of psychological experiments can enhance accuracy and reduce reproducibility issues.
b) Creating models that emphasize prediction, not just at the group level but also at the individual level.

In recent years, ML has emerged as a powerful tool in sophisticated analytical techniques, demonstrating its efficacy in identifying deception (Yang et al., 2022). Pioneering research showed that ML could be applied to kinematic analysis of motor responses in dual-choice tasks to automatically differentiate between truthful and deceptive responses (Monaro, Gamberini, et al., 2017), expanding the application of ML to various domains. For instance, keyboard dynamics have been used to identify fake identities in both laboratory settings (Monaro, Galante, et al., 2018) and online environments (Monaro et al., 2019), while mouse tracking analysis has been applied in the study of depression (Monaro, Toncini, et al., 2018) and in detecting false identities (Monaro, Fugazza, et al., 2017) . A particularly noteworthy application of ML is in the realm of psychiatric symptoms, which are inherently complex to assess. In this context, ML techniques can distinguish malingerers from honest respondents with notable accuracy (Mazza et al., 2020; Monaro, Toncini, et al., 2018).

The field of deception detection research has also expanded its focus to include the identification of verbal cues as indicators of deceit (for a comprehensive overview of verbal lie detection research, refer to Vrij et al., 2022). Equally intriguing is the recent exploration into the application of ML models, specifically engineered to discern between truthful and deceptive narratives with remarkable precision (Otasowie, 2020). The latest breakthrough in this arena is epitomized by the employment of Large Language Models (LLMs), which have been effectively applied to this complex challenge (Loconte et al., 2023). This multifaceted methodology accentuates the adaptability and efficacy of machine learning in a broad spectrum of psychological and behavioral analyses.

1.2.5 Neuropsychological and Psychological Malingering Detection

In clinical settings, the concept of malingering, along with methods for its detection, is comprehensively delineated, particularly about neuropsychological or psychological disorders. Notably, neurocognitive disorders such as amnesia, dementia, and deficits in cognitive and executive functions are prominent. Neurocognitive impairment diagnosis primarily relies on the performance levels observed during cognitive assessments. During these evaluations, there is a potential risk for examinees to feign their impairments, either by exaggerating or fabricating symptoms. However, it is crucial to recognize that simulating an improved performance, or "fake good," is not typically feasible since individuals can only perform within their true capabilities. The only exception to this rule occurs in scenarios involving "coaching," where an examinee might receive advice, possibly from a legal advisor or psychologist, on accurately responding during a test to attain an artificially enhanced performance level.

Traditional Techniques to Detect Malingering

The following are some methods and logic that the professional may draw to detect malingering:

- *Anatomic-Clinical Correlation:* This approach aims to identify malingering by corroborating neuropsychological test results with neuroimaging findings. In instances where discrepancies arise, there is a suspicion of malingering behavior. These discrepancies may manifest in the disorder's qualitative and quantitative aspects, including the lesion's location and the severity of symptoms.
- *Floor Effect Strategy* (Rey, 1958; Rogers et al., 1993): This strategy assumes that individuals feigning symptoms, or "simulators," will likely attempt to deceive across all tests during an assessment. To counter this, psychologists embed simple tasks within more complex ones. The key feature of this approach is that even patients with severe dementia are typically capable of completing these elementary tasks. In fact, the average score achieved by a validation sample is set as the benchmark. Conversely, a simulator may intentionally fail these simple tasks. Therefore, any errors in these elementary tasks are indicative of potential simulation. Several tests employ this principle, including the Rey 15-Item Memory Test (Rey, 1958), the Rey Word Recognition Test (Rey, 1941), and the Rey Dot Counting Test (Rey, 1941).
- *Symptom Validity Testing***:** This strategy developed by Pankratz (Pankratz, 1983) involves a forced-choice test in which random answering would theoretically yield a

50% accuracy rate. The underlying assumption is that genuine patients with a specific cognitive deficit, such as memory impairment, will perform a memory test at a chance level. However, performance below this chance level is considered implausible. It suggests that the individual is likely able to recognize the correct answers but chooses not to select them, deliberately portraying a falsified poor profile. This strategy is incorporated in numerous tests, including the Test Of Memory Malingering (TOMM) (Tombaugh, 1996), Letter Memory Test (Inman et al., 1998), Digit Memory Test (DMT) (Hiscock & Hiscock, 1989), Portland Digit Recognition Test (PDRT) (Binder & Willis, 1991), Victoria Symptoms Validity Test (VSVT) (Slick et al., 1996), Computerized Assessment of Response Bias (CARB) (Allen et al., 1997), 21-Item Test (Iverson et al., 1991), Coin In The Hand Test (Kapur, 1994), and others.

- *Violation of a Scientific Law*: This criterion posits that if an examinee contravenes a well-established scientific principle during an examination, they may be considered a malingerer. For example, if an individual cannot verbally repeat a sequence of three items in a short-term memory task but succeeds in recalling a sequence of five items, this inconsistency suggests malingering. Such a pattern violates the generally understood principles of cognitive functioning, where the ability to recall a more extended sequence implies the capacity to recall a shorter one.

When addressing psychological/psychiatric disorders, reference is made to the conditions outlined in the Diagnostic and Statistical Manual of Mental Disorders - 5th Edition (DSM-5). The primary challenges in psychiatric diagnosis include a) the lack of observable neural alterations or biomarkers, b) diagnoses based on criteria that can sometimes be ambiguous, and c) reliance primarily on self-reported symptoms for diagnosis. Given these factors, it might seem relatively straightforward to either feign or mask a psychiatric disorder. In response to this challenge, various authors have proposed a range of methods to detect both underreporting (faking good) and overreporting (faking bad) of psychiatric symptoms:

- *Incoherent Psychopathology:* This principle is foundational in classical psychiatry and the control scales of personality inventories. Experts compare the examinee's reported symptoms against the established phenomenology of the specific mental disorder. Discrepancies between the examinee's symptoms and the disorder's known characteristics can indicate potential malingering or dissimulation.
- *Scenarios Method:* This approach involves administering a questionnaire that situates psychiatric symptoms within realistic scenarios. Individuals feigning disorders

(simulators) tend to respond pathologically, describing their symptoms as both fabricated and severe.

- *Qualitative Analysis of Typical Malingering Characteristics:* This analysis identifies some typical attributes of malingering (Conroy & Kwartner, 2006):
    - Simulators often appear increasingly normal over time. Lengthy interviews may reveal this pattern, suggesting malingering.
    - Endorsement of positive rather than negative symptoms. Simulators frequently fabricate positive symptoms like delusions and hallucinations while rarely feigning catatonic behaviors or inappropriate affect.
    - Aberrations in content rather than form of thought. A truly mentally ill person will tend to produce disorganized speech, loose associations, and flights of ideas that characterize thought disorder and are, instead, almost impossible to fake during a lengthy interview.
    - Over-playing and reminding. Malingerers are more likely to emphasize their delusions excessively.
    - The simulators tend to give approximate answers.
    - Positive response to suggested symptoms. Malingerers often confirm symptoms, believing this will enhance the perceived severity of their psychopathology.
    - A conglomerate of symptoms not consistent with any mental illness. High reporting of symptoms that do not align or are inconsistent with recognized mental disorders, aiming to present a more severe or complex condition.
- *Discrepancy Method:* This method assesses the congruence of symptom patterns with known psychiatric or neurobehavioral syndromes. It involves a clinical evaluation of the differences between the examinee's reported or exhibited symptoms and the typical manifestations of the claimed disorder. Larrabee identified five critical forms of discrepancies to consider before diagnosing genuine conditions (Larrabee, 1990):
    - Internal Consistency of Neurobehavioral Domains: This involves examining gross divergences in performance on tests that are expected to be highly correlated, indicating a lack of internal consistency in the examinee's responses.
    - Disease-Deficit Compatibility: This discrepancy checks whether the reported or demonstrated symptoms align with primary manifestations of the claimed disorder, identifying instances where symptoms do not correspond with the disorder's known characteristics.

- Inconsistency with Severity of Injury: This aspect evaluates the alignment between the reported severity of the disease and the impairment with related symptoms, looking for divergences that suggest exaggeration or fabrication.
- Ecological Discrepancy: This involves comparing scores in specific domains on clinical tests with the examinee's observed behavior in the same domains, identifying inconsistencies that may indicate feigned symptoms.
- Violation of Performance Curves: Honest responders typically perform better on easier items and struggle with more difficult ones in tests. In contrast, malingerers may exhibit an unusual pattern, such as performing well on difficult items and poorly on easier ones, which defies the expected performance curve.

- *Rare Symptoms and Unrelated Symptoms* (Sartori et al., 2016). This approach focuses on reporting symptoms that are either infrequently observed in clinical populations or appear as a non-specific collection of numerous symptoms without a coherent diagnostic pattern. If an examinee reports exceptionally rare or improbable symptoms characterized by extraordinary severity or fantastical qualities, this may indicate malingering. Malingerers often fail to gauge the typical severity of symptoms accurately. Malingerers might inaccurately portray a disorder based on common misconceptions or stereotypes. For example, describing a schizophrenic condition as "having two personalities" (Conroy & Kwartner, 2006) reflects a misunderstanding of the disorder and may signal feigned symptoms. Reporting symptoms inconsistent with the examinee's observed behavior is another red flag for malingering. This includes symptoms that do not align with the individual's overall presentation or known disorder patterns.
- *Structured Interview of Reported Symptoms* (SIRS) (Rogers, R., Bagby, R. M., & Dickens, 1992). This interview comprises 172 items to evaluate a broad spectrum of psychopathology and symptomatology. It is structured into eight primary scales and five supplementary scales, specifically designed to identify 13 distinct response styles in respondents. The outcomes of the primary scales are categorized into four classifications: honest responding, indeterminate, probable feigning, and definite feigning. Renowned for its satisfactory reliability and validity, this tool boasts a 97% accuracy rate in classifying malingerers (Rogers, 1997). Its primary scales are: RS (Rare Symptoms), which targets seemingly valid yet infrequently encountered symptoms, akin to the F scale from MMPI-2; SC (Symptom Combinations), which is similar to RS but refers to pairs of symptoms; IA (Improbable or Absurd Symptoms) is for

intrinsically improbable and absurd symptoms, BL (Blatant Symptoms), SU (Subtle Symptoms), SEV (Severity of Symptoms), SEL (Selectivity of symptoms), RO (Reported vs. Observed Symptoms). The supplemental scales include DA (Direct Appraisal of Honesty, asking explicitly to indicate their honesty during evaluation), DS (Defensive Symptoms; common symptoms that, if denied, are a sign of defensiveness), SO (Symptom Onset; atypical onset), OS (Overly Specified Symptoms), INC (Inconsistency of Symptoms; 32 repeated items).

- *Structured Inventory of Malingered Symptomatology* (SIMS) (Smith & Burger, 1997). This self-administered questionnaire, comprising 75 dichotomous (true/false) items, is specifically designed to identify feigned psychiatric and cognitive symptoms. The test effectively discerns bizarre experiences and highly atypical symptoms across various domains, including low intelligence (LI), affective disorders (AF), neurological impairment (N), psychosis (P), and amnestic disorders (AM). It yields a total score that indicates the likelihood of malingering in psychological disorders, demonstrating a sensitivity of 97.06% (Smith & Burger, 1997). A Systematic Review and Meta-Analysis (van Impelen et al., 2014), which included 41 studies, concluded that this instrument effectively differentiates between instructed feigners and honest responders. It generates higher scores in groups known to have an increased prevalence of feigning; may potentially overestimate feigning in patients with schizophrenia, intellectual disability, or psychogenic non-epileptic seizures; and shows reasonable resistance against coaching techniques. Recent research by Orrù and colleagues has explored the integration of machine learning techniques to develop a condensed questionnaire version. This endeavor aims to reduce the length of the questionnaire while preserving its accuracy in discriminating between genuine and feigned presentations (Orrù et al., 2021).
- *Inventory of Problems-29* (IOP-29) (Viglione et al., 2017): This concise, self-administered, paper-and-pencil measure is designed to differentiate between genuine and feigned mental and cognitive disorders. It is particularly effective in distinguishing bona fide psychiatric and cognitive complaints from feigned ones.
- *M-Test* (Beaber et al., 1985): Explicitly developed to identify malingering of schizophrenic symptoms, the M-Test evaluates the likelihood of feigning or exaggerating psychiatric symptoms.

- *Control Scales:* These can be categorized into three distinct types: a) Specific Psychometric Tools, explicitly developed to detect deceitful response styles; b) Methods in Standard Psychodiagnostic Tests, integrated into standard psychodiagnostic tests to detect simulation and dissimulation; c) Personality Inventories in Clinical Settings, created for use in clinical settings rather than forensic ones and aid in identifying deceptive responses (Stracciari et al., 2010):
    - *Minnesota Multiphasic Personality Inventory* (MMPI-3) (Ben-Porath & Tellegen, 2020). This test is the most widely used and researched multi-scale measure of psychopathology. It comprises 567 dichotomous (true/false) items and offers insights into an individual's personality traits and alignment with specific nosographic categories. The test includes 10 clinical scales, 15 content scales, and various supplemental scales. A critical feature of MMPI-3 is its array of validity scales, which are instrumental in hypothesizing about dissimulation and simulation by checking the reliability of the profile. These validity scales include Combined Response Inconsistency (CRIN), Variable Response Inconsistency (VRIN), which assesses answering similar or opposite question pairs inconsistently, True Response Inconsistency (TRIN), assessing answering questions all true or all false, Infrequent Responses (F), Infrequent Psychopathology Responses (Fp), Infrequent Somatic Responses (Fs), Symptom Validity Scale (FBS), Response Bias Scale (RBS), Lie (L), Defensiveness (K), F-K index (Dissimulation Index). Additionally, the MMPI-3 employs a method of analyzing both evident and subtle items. The F-K index, in particular, is effective in accurately classifying malingerers, with an accuracy rate reaching approximately 90%(Stracciari et al., 2010).
    - *Personality Assessment Inventory* (PAI) (Morey & Staff, 1991): The PAI is a well-established, multi-scale inventory recognized for its psychometric reliability and utility in clinical and forensic contexts. A notable feature of this inventory is the Negative Impression Management scale (NIM), specifically designed to identify exaggerated reports of psychopathological symptoms and potential malingering.
- *Cognitive test:* Many tests for detecting malingered cognitive symptoms mainly focus on detecting feigned memory impairment (one of the most frequent). Here, we present those most frequently used in clinical and forensic practice as they address often feigned symptoms.

- *Test of Memory Malingering* (TOMM) (Tombaugh, 1996): The TOMM is a forced-choice recognition test explicitly designed for adults. It comprises up to three trials, each containing 50 items, and is utilized to distinguish between genuine memory-impaired patients and individuals who may be feigning memory impairment. This tool is particularly effective in identifying exaggerated or fabricated memory deficits. Tombaugh's investigations (Rees et al., 1998; Tombaugh, 1996, 1997) demonstrated that the TOMM can accurately classify 95% of non-demented patients as not malingering, underscoring its efficacy and reliability in memory assessment contexts.
- *Rey's 15-Item Visual Memory Test* (MFIT) (Rey, 1964): This test assesses malingered memory impairment. It involves displaying 15 different symbols arranged in a table with three rows and five columns. The respondent views the table for 10 seconds and is then asked to draw as many symbols as they can recall. The traditional scoring method involves counting the total number of items correctly recalled, with scores lower than nine arousing suspicion of malingering.
- *Word Memory Test* (WMT) (Green et al., 1996): This is a computer-administered memory task. Participants are initially shown a list of 20-word pairs. Each pair is displayed briefly - the first word followed by the second word one second later, with a two-second interval before the next pair appears. After the list is shown twice, the subject is asked to recall as many word pairs as possible. This test is designed to evaluate memory recall and consistency in responses.
- *Dot Counting Test* (DCT) (Lezak et al., 2004): The DCT is a screening measure for malingering, consisting of 12 cards, each printed with a series of dots. Respondents are shown these cards in a fixed, non-sequential order and are instructed to count the dots as quickly as possible. For honest patients, it is expected that the time taken to count the dots will increase proportionally when the dots are ungrouped, as opposed to when they are grouped, reflecting the natural difficulty of the task.

All the instruments discussed previously have been subject to rigorous validation studies, the findings of which have been published. These studies are crucial in ensuring that

these tools comply with the Daubert criteria[1], a framework used to determine the admissibility of scientific evidence in legal proceedings. The Daubert criteria encompass the following key elements:

1. Empirical Testing: The underlying theory or technique must be capable of being tested, falsifiable, and refutable, ensuring its empirical foundation.
2. Peer Review and Publication: This criterion ensures that the instrument has undergone rigorous scrutiny by other experts in the field, validating its credibility and scientific soundness.
3. Known or Potential Error Rate: It is essential to quantify the reliability of the technique, which is achieved by establishing a known or potential error rate.
4. Standards and Controls: The development and consistent application of standards and controls in the operation of the tool are vital to maintaining its integrity and reliability.
5. General Acceptance in the Scientific Community: The theory and technique should be widely accepted within the relevant scientific community, indicating its validity and usefulness.

These criteria collectively ensure that the psychological instruments are scientifically sound and legally admissible, enhancing their applicability in clinical and forensic contexts.

Concluding Note

Readers are advised to anticipate some degree of redundancy in the content between this Introduction and the introductory sections of each chapter. This overlap is due to the nature of the chapters, which originated from standalone papers. These papers have either been previously published or are currently under review.

[1] Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993).

Chapter 2.
Detecting Deception at Single Item Level in Questionnaire

2.1 Abstract

In forensic settings, the psychometric analysis of pathological personality is especially vital, as a substantial proportion of behaviors — ranging from 25% to 45% — are deceptive. However, detecting simulation poses a significant technical challenge in forensic practice. Although techniques are available to identify a faker, they cannot identify the specific questions distorted by faking. Moreover, not all psychological questionnaires have a control scale to detect simulatory or dissimulatory behaviors. This work evaluates the effectiveness of the Term Frequency-Inverse Document Frequency (TF-IDF) index as an innovative and effective technique for detecting deception at the single response level.

We validated the technique on three datasets containing responses to the 10-item Big Five questionnaire (total of 694 participants, respectively 221, 243, and 230) in three "faking good" situations. Each participant responded twice, once faking to achieve an objective in one of three contexts (one to obtain child custody and two to land a job) and once honestly. The proposed TF-IDF model has proven very effective in separating honest and dishonest responses and identifying which answers to the questionnaire were distorted in the dishonest condition.

Based on Purpura, A., Giorgianni, D., Orrù, G., Melis, G., & Sartori, G. (2022). Identifying single-item faked responses in personality tests: A new TF-IDF-based method. Plos one, 17(8), e0272970. Published under a CC BY SA license.

2.2 Introduction

Deception in forensic contexts poses a significant challenge. This study explores the complexities of analyzing data from questionnaires, especially when the integrity of responses is compromised by deliberate distortion. Notably, respondents can easily manipulate answers to straightforward questions, especially when they discern a tangible benefit. Such distortions typically manifest in two forms: "faking bad" and "faking good." The former involves respondents exaggerating or fabricating psychological ailments, often observed in criminal litigations and insurance claims (Sartori et al., 2017). In contrast, "faking good" occurs in environments where respondents are expected to provide exceptionally favorable responses.

While estimates on the frequency of simulation vary, they range from 25-45% in forensic settings (Mittenberg et al., 2002; Young, 2015). Over the past few decades, a substantial body of literature has emphasized the high prevalence of maladaptive personality traits in the criminal population (Coid et al., 2002; Fazel & Danesh, 2002) or personality disorders, which is approximately three times higher than in non-forensic settings (Rotter et al., 2002). Personality assessment outcomes are critical in criminal proceedings, legal decisions, probation assignments, and incarceration (Edens et al., 2001), creating an incentive for individuals undergoing assessment to exaggerate or feign symptoms. Thus, evaluating pathological and maladaptive personality traits is crucial in legal and criminal justice environments (Niemeyer et al., 2022).

Faking in psychological questionnaires is typically controlled using control scales, such as the L, F, and K scales of the Minnesota Multiphasic Personality Inventory-3 (Ben-Porath & Tellegen, 2020) and the X, Y, and Z scales of the Millon Clinical Multiaxial Inventory-III (Theodore Millon, 2015). Abnormal scores on these scales indicate an overall tendency of the subjects to modulate their responses in the direction of social desirability (faking good) or hyperbolic psychopathology (faking bad). In general, faking is a continuous variable, influenced by the stakes and the respondent's conscious or unconscious strategy. For this reason, efforts have been made to develop tests to identify fakers. Still, no procedure has been proposed to evaluate the intensity of specific symptoms or identify a simulation focused on a particular item. At best, these scales determine a general propensity for simulation but lack precision in pinpointing the source of feigned responses. Furthermore, not all questionnaires include control scales to monitor such behaviors.

In response to these challenges, this study aims to develop an efficient technique for spotting faked responses to individual items rather than a general faking response style. We

explored the potential of the Term Frequency – Inverse Document Frequency (TF-IDF) model (Baeza-Yates & Ribeiro-Neto, 1999) as a tool for detecting simulation in self-report instruments. This approach identifies the infrequency of a particular response and the degree of deviation from other subjects' responses, highlighting deception at the single-item level.

In classical test theory and item response theory (IRT), an item's value is evaluated based on its performance. In classical test theory, a participant's overall responding style is not modeled – an example of a responding style is the tendency to use the full range of response alternatives or just the extreme values. Recently, a variant of IRT, the IRT tree model, has been proposed (Böckenholt & Meiser, 2017), adding a further cognitive component representing the response style. Such response style is intended to represent the propensity of the responder to use extreme values in a Likert scale. Such cognitive component could be modified to represent the propensity of the responder to fake, as proposed by Sun et al., (2022). The basic idea behind using an IRT tree model for faking detection in Sun et al., (2022) is to fit a probabilistic model to represent the decision process of a respondent to a questionnaire. The model is then employed to compare honest and fake respondents' styles and to distinguish between them.

Unlike IRT-tree, the TF-IDF approach offers a purely data-driven and model-free solution for faking detection that only requires the collection of honest responses. This aspect provides some significant advantages for a widespread application of the proposed approach compared to IRT. A TF-IDF-based approach does not require a domain expert to design the experiments and can be used following the same process regardless of the domain or type of questionnaire. Additionally, a TF-IDF-based approach allows comparing the respondents' style across different questionnaire items instead of being limited to each item. Our TF-IDF approach combines a response's relative positioning with respect to the validation sample and a specific subject's response style (Zhang et al., 2011). From this perspective, it also differs from the methods that detect faking by analyzing only the extreme responses.

We validated the proposed model using the 10-item Big Five Test (Rammstedt & John, 2007), a widely used short version of the Big Five personality test suited to quickly collect a psychological profile over the web. We administered this questionnaire to a group of volunteers instructed to answer twice: the first time dishonestly according to specific instructions and the second time honestly. As faking changes according to the respondent's goal, we provided the following three faking contexts: (a) a child custody case (CC), (b) a job interview for a sales manager position (JIS), and (c) a job interview for a position in a humanitarian organization (JIHO).

This research makes a unique contribution by proposing and applying the Term Frequency-Inverse Document Frequency (TF-IDF) technique to detect falsification patterns in personality tests at the item level. This groundbreaking approach enhances the reliability and validity of test outcomes by distinguishing between authentic responses and those that have been artificially manipulated. Furthermore, we also evaluate the proposed approach based on TF-IDF compared to similar techniques based on the distribution of raw response values.

Additionally, this technique allows the creation of three datasets to develop new item-level faking-detection approaches. This technique allows for its extension across various testing paradigms, facilitating the accumulation of an adequate corpus of genuine answers and addressing a fundamental limitation of existing lie detection methods that rely on obtaining data from honest and deceptive groups to classify new subjects. In contrast, our approach only requires a representative sample of honest responses for comparison. The extensive databases, encompassing data from 694, sharpen our understanding and establish a solid groundwork for applying this method to new cases. Expanding TF-IDF implementation across questionnaires and testing scenarios will further enrich our understanding of its applicability for deception detection.

2.3 Materials and Methods

2.3.1 Proposed Approach

TF-IDF is a standard term-weighting method used in information retrieval to compute a numeric representation of textual data based on the distribution of the terms occurring in it. It is one of the mathematical tools search engines use to retrieve web pages relevant to a user query and rank them according to their relevance. It is also used as a text representation strategy in natural language processing (Zhang et al., 2011). The TF-IDF representation strategy is based on the intuition that a specific term in a document, *d*, is representative of its general meaning proportional to (a) its term frequency (TF) within *d* and (b) its inverse frequency in the collection (IDF; i.e., a term that frequently appears in *d* and at the same time is only rarely used in others) is more representative for the content of *d* than different ones that are either not frequently used in *d* or are very frequently used also in other documents. Formally, given a collection, *C*, of documents, the TF-IDF value for each term, *t*, in a document $d \in C$ is calculated as:

$$\text{TF-IDF}(t, d, C) = TF(d, t) * IDF(t, C), \quad (1)$$

where TF (d, t) indicates the number of times a target term, t, appears in a document, d. IDF = log(N/n), where N indicates the number of documents in C and n is the number of documents in which t is used.

The application of TF-IDF is not only limited to the field of computer science but has also been used, for example, in cognitive neuropsychology to model semantic memory and its disorders (Sartori & Lombardi, 2004) as well as brain responses to relevance in picture recognition (Mechelli et al., 2006). When applied to item analysis, we compute each response's TF-IDF in a questionnaire, q, considering as TF the number of times the subject provides the current response. For example, if a subject responded with a rating of 5 only to one item out of the 10 items, then TF (5, q) = 1. Indeed, for each item, the subjects could choose their degree of agreement on a 5-point Likert scale.

The IDF is computed as log(N/n), where N is the number of participants who responded to the test (for instance, 221) and n is the number of times the participant's response was given also by other participants to the same item. The final score associated with each test item is the product of the TF and IDF values. This adaptation of TF-IDF has high values for responses to single items that are highly atypical. From this perspective, TF-IDF behaves as an anomaly detector, indexing the very atypical responses at the single-item level (not at the subject level). The following example illustrates how the TF-IDF could be used to spot a faker. A 10-item personality test is administered to 1000 participants, who can express their degree of agreement on a 5-point Likert scale. Subject 1 replies "5" to Item 3 and, in total, to seven items out of 10. Regarding the whole sample's replies to Item 3, 300 subjects reply "2", 650 reply "3", and 50 reply "5". Therefore, the TF-IDF score associated with Item 3 for Subject 1 is computed as $7 * \log(1000/50) = 9.11$. This value will be high for people who provide the same—atypical—response many times in a questionnaire. From this perspective, dishonest responses are expected to have an abnormally high TF-IDF, compared to the TF-IDF distribution of honest respondents. Figure 2.1 summarizes the proposed approach based on TF-IDF in the pseudo-code.

```
def tfidf(subject responses, per item idfs,
    per item thresholds):
faked responses indices = [];
For response index in range(len(subject responses)):
current response = subject responses[response index];
 TF = count(response, subject responses);
IDF = per item idfs(response index, current response);
If TF*IDF >per item thresholds[response index]: faked responses
    indices.append(response index);
Return faked responses indices;
```

Figure 2.1 Code Snippet 1. TF-IDF for single-item faking detection.

For each response a subject gives on a questionnaire, we compute its corresponding TF and IDF scores as described above. Then, if their product is higher than a certain threshold, we flag the current response as faked.

Given a training set, the threshold is selected as the percentile score in the TF-IDF values distribution (for each item) which maximizes the precision metric for faking detection on a separate validation set. It is not calculated on the raw TF-IDF scores but on the corresponding percentile of the distribution of the TF-IDF scores on a certain item. We then evaluated the threshold value following a k-fold cross-validation procedure (with k = 10). This process allowed for evaluating the proposed approach's performance in terms of accuracy, precision, recall, and F1 score.

In our study, we first calculated the IDF scores associated with each item in the honest responses. Afterward, we considered a small subset of dishonest responses—as a validation set—and fine-tuned the faking detection threshold for each item based on the distribution of these scores. For example, should we observe that the value associated with the 75th percentile in the distribution of TF-IDF values in our validation set was a reliable threshold to distinguish between honest and dishonest responses for the questionnaire items, we would use said value to detect fakers for each item. We identified the 75th percentile empirically based on our data distribution. More specifically, we first computed the distribution of TF-IDF scores associated with each questionnaire item; then, we evaluated values at different percentiles of each distribution as potential threshold candidates to distinguish dishonest from honest responses. Finally, we selected the percentile value that maximizes the precision performance metric for detecting faked response items.

When a new questionnaire item is to be evaluated for dishonesty, we rely on the previously computed set of IDF scores for the respective honest responses and on the TF scores computed for the current responses to assign a TF-IDF score to each response. Then, if the obtained TF-IDF value is higher than a certain threshold—which we validated beforehand, as explained above—we report the item as faked. In other words, we rely on the TF-IDF score associated with a subject's responses to a questionnaire as an indicator of their attitude and as a proxy to detect faking behaviors. For this purpose, TF-IDF, as proposed here, acts as an anomaly detector, identifying suspicious differences between the to-be-evaluated case and the distribution of honest respondents' answers. Furthermore, in validating this approach, considering various faking contexts helps us exclude potential TF-IDF approach biases to specific data distributions or situation-specific behaviors (Birkeland et al., 2006; Merckelbach et al., 2009).

To validate the technique, we conducted three experiments in which we asked participants to respond to the Italian version of the 10-item Big Five questionnaire (Guido et al., 2015) twice, the first time dishonestly according to specific instructions and the second time honestly. We gave the instructions in three particular contexts: (a) a child custody case, (b) a job interview for a sales manager position, and (c) a job interview for a position in a humanitarian organization. We used three faking contexts because faking changes according to the respondent's objective. Having the same subject respond dishonestly and honestly gives us a ground truth helpful in evaluating the TF-IDF method's accuracy level in identifying faked responses. Indeed, in most practical applications, only one set of responses per subject will be available, and the proposed approach's goal is to spot faking patterns in the subject's responses.

2.3.2 Participants

A total of 817 native Italian speakers were initially recruited online via social networks. Following the exclusion criteria outlined in Section 2.3.4, "Faking Instructions," the final sample comprised 694 participants. Information regarding the participant's age, gender, and education level was collected and is presented in Table 2.1

Table 2.1 Number of Participants, Age, and Education in the Three Groups

Group	Number (Female)	Average Age (SD)	Average Education (SD)
CC	221 (176)	41.74 (14.29)	15.40 (2.9)
JIS	243 (185)	43.07 (12.04)	15.48 (3)
JIHO	230 (188)	40.83 (11.53)	15.81 (2.8)

Note. CC = child custody, JIS = job interview salesperson, JIHO = job interview humanitarian organization

2.3.3 Procedure

The experimental procedure was approved by the Ethics Committee at the Department of General Psychology at the University of Padua (Protocol n. 4545). Participants were informed of the nature of the study, provided written consent, and completed the demographic questionnaire (age, sex, and education) and the main questionnaire. In the three conditions, we asked participants to respond to the same set of 10 items twice, first altering their answers to obtain personal gain in the three contexts described above and then honestly. When a test is taken in a real faking-prone setting, the participant has not responded to the same test honestly

before. The order faking first - the honest second was adopted to have the faked condition close to the real administration condition.

All available responses were in multiple-choice format, and participants could choose only one option for each item. We asked the participants to respond by estimating their agreement using a 5-point Likert scale (1= *"strong disagreement"*; 5 = *"strong agreement"*) on the Italian version (Guido et al., 2015) of the following 10 items from the Big Five Inventory (Rammstedt & John, 2007). The English version of the test is reported below:

"I see myself as someone who . . ."

1. is outgoing, sociable (extroversion);
2. is reserved (extroversion reversed);
3. tends to find fault with others (agreeableness, reversed);
4. is generally trusting (agreeableness);
5. tends to be lazy (conscientiousness, reversed);
6. does a thorough job (conscientiousness);
7. gets nervous easily (emotional stability, reversed);
8. is relaxed, handles stress well (emotional stability);
9. has an active imagination (openness);
10. has few artistic interests (openness, reversed).

Before analyzing the collected data, we reversed participants' responses for reversed items (signaled above) so that all the final scores were directly comparable.

2.3.4 Faking Instructions

Group 1-CC: *"Pretend you and your spouse are separating and fighting in court for the custody of your child (imagine having a child if you don't have one). A court-appointed psychologist will evaluate your parental skills and require you to respond to the following personality questionnaire. Answer the test in a way that emphasizes your parental skills, hiding behaviors or thoughts that are generally considered negative for a parent. Your goal is to get custody of your child."*

Group 2-JIS: *"Pretend you are attending a job interview for the position of sales manager for a very prestigious company. It is your dream job, and you want to give the best possible image of yourself. Your goal is to obtain the job and to be considered better than the other candidates. Answer in a way that makes you look good, hiding behaviors or thoughts that*

are generally considered negative for the position you aspire to. Try to give a positive image of yourself, even if it means lying."

Group 3-JIHO: Participants received instructions similar to those we provided to Group 2, with the position being the only difference (a position in a humanitarian organization).

The procedure included a question to verify that the participants read and fully comprehended the instructions in all three conditions before completing the questionnaire (Liu & Wronski, 2018). We did not include participants who did not answer this final question correctly in the analysis. The additional final question for Group 1 (Child custody litigation) was *"What instructions have you received when filling in the questionnaire?"* with the following response choices:

- I always had to tell the truth.
- I received no specific instructions.
- First I had to tell the truth, and then I had to lie to get custody of the children.
- I always had to lie to get custody of the children.
- First I had to lie to get custody of the children, and then I had to tell the truth.

Of the 258 participants, 37 did not satisfy this quality check. Therefore, we included only 221 participants in our analysis, discarding the responses of all participants who answered incorrectly. The additional final question for Groups 2 and 3 was "*What instructions have you received when filling in the questionnaire?"* with the following response choices:

- I always had to tell the truth.
- I received no specific instructions.
- First I had to tell the truth, and then I had to lie to look good in a job interview.
- I always had to lie to make a good impression.
- First I had to lie to look good in a job interview, and then I had to tell the truth,

This quality check for Group 2 (JIS) excluded 44 of 287 participants, with 243 included in the final analysis, who were selected based on their correct responses to the final questions. From Group 3 (JIHO), we excluded 42 of 272 participants, including 230 participants in the following evaluation. The proposed approach to compute TF-IDF for questionnaire responses and the data employed for the evaluation are available in our public repository: https://github.com/albpurpura/TFIDF-Faking.

2.4 Results and Discussion

This section presents the results of our experiments to validate the proposed approach based on TF-IDF. We begin by giving a few characteristics of the data we collected and providing the analysis we performed with the raw data. Table 2.2 shows the mean and standard deviation of the honest (H) and dishonest (D) responses in the three faking contexts. Data indicate that faking in different contexts resulted in changes to different items. For example, in the JIS dataset, faking did not significantly affect one of the two items related to a person's agreeableness (Q4). In addition, we reported the probability of fakers to alter their response to each item in the different contexts we considered.

Table 2.2 Mean, SD, and Faking Probability of the Questions in Each Honest and Dishonest Questionnaire in Different Datasets.

Dataset	Measure	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
CC	Mean H	3.87	2.43	2.96	3.42	3.17	4.48	2.94	3.01	3.54	3.32
	Mean D	4.47*	2.20*	3.90*	3.69*	4.41*	4.58	4.35*	4.23*	3.40	3.98*
	SD H	0.90	0.96	0.99	0.97	1.22	0.68	1.22	1.09	1.10	1.25
	SD D	0.71*	0.93*	0.96*	0.90*	0.85*	0.81	0.80*	0.97*	1.24	0.93*
	Faking prob.	0.60	0.48	0.72	0.52	0.76	0.39	0.74	0.77	0.66	0.59
JIS	Mean H	3.67	2.44	2.95	3.42	3.21	4.47	2.74	2.92	3.63	3.45
	Mean D	4.55*	2.75*	3.46*	3.49	4.52*	4.67*	4.18*	4.40*	4.12*	3.84*
	SD H	0.97	1.09	1.04	1.06	1.20	0.77	1.18	1.14	1.04	1.26
	SD D	0.74*	1.25*	1.18*	1.04	0.82*	0.74*	1.05*	0.86*	1.06*	1.14*
	Faking prob.	0.67	0.61	0.77	0.62	0.74	0.39	0.79	0.80	0.70	0.62
JIHO	Mean H	3.77	2.38	3.00	3.55	3.21	4.49	2.81	3.00	3.56	3.42
	Mean D	4.53*	2.69*	3.94*	3.96*	4.47*	4.57	4.37*	4.35*	4.10*	4.08*
	SD H	0.95	1.04	1.11	0.96	1.25	0.71	1.16	1.18	1.10	1.26
	SD D	0.86*	1.25*	1.11*	0.94*	0.91*	0.86	0.84*	0.88*	1.02*	0.99*
	Faking prob.	0.64	0.63	0.71	0.65	0.73	0.38	0.77	0.76	0.66	0.54

Note. Values in the faked questionnaire's statistics marked with a * indicate a statistically significant difference ($p < 0.05$).

As Table 2.2 shows, different contexts affect different dishonest responses. For example, Q2 was changed in one direction in faking in the child custody (CC) context and in the opposite direction in a job interview. Moreover, the 10 items' average scores did not differ significantly in the three honest-respondent groups. The faking probability per item shows that the likelihood of faking an item is strongly related to the faking context.

Although the probability of faking changes depends on the context and the item, we generally expected that the answers given in the faking conditions would score higher than those provided in the honest condition. To verify this, we calculated the proportion of responses that increased from the honest to the dishonest condition (e.g., from 3 to 5) or remained unchanged (Table 2.3). The overall pattern is similar in all three faking conditions. Most interestingly, in 13–15 percent of the cases, dishonest responses were changed in an unexpected direction, with a lower score in the dishonest condition compared to the honest one.

Table 2.3 Unchanged and Increased Responses by Group.

Group	Total responses	% Unchanged	% Increased
CC	2210	38 %	48 %
JIS	2430	33 %	52 %
JHO	2300	35 %	52 %

Note. The table includes percentages of raw score responses that remained unchanged in the honest and dishonest conditions and changed in the expected direction according to the faking instructions. With the instructions we gave, we expected an increase in the raw scores (faking good).

This result could have occurred due to different faking strategies adopted by other participants as well as test-retest issues. Previous studies have shown that the test-retest reliability of the 10-item Big Five Questionnaire is 0.72 (Rammstedt & John, 2007). Since the reliability is 0.72, it could have changed the answers between the dishonest and honest conditions. In other terms, the test-retest vagaries of non-faked responses could be the origin of the decrease in response scores. Another possible explanation for the 15% decrease in faked responses may be related to the order of the presentation of the two versions, which was Dishonest-Honest. Honest responses are given after each participant has already responded faking to the same test, which could have impacted the reliability of his/her honest responses.

In addition, we investigated the correlation between honest and dishonest answers. Figure 2.2 shows the correlation matrices between honest and dishonest responses in the three datasets. What emerges is that the correlations between honest and dishonest responses are

around 0. This result indicates that, in all three groups, no linearity exists between honest and dishonest means; therefore, there is no straightforward procedure for predicting honest responses from dishonest ones.

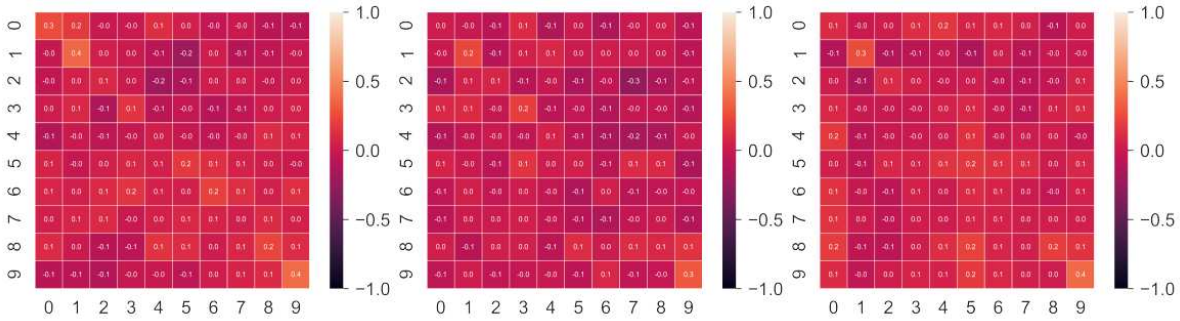


Figure 2.2 Correlation Matrices Between Honest and Dishonest Responses in the CC, JIS, and JIHO Datasets, respectively, from left to right. Correlation between responses in the honest and dishonest condition is virtually nonexistent, rendering the identification of the honest responses from the dishonest ones a difficult task.

2.4.1 TF-IDF for Item Analysis

We changed the raw data into TF-IDF indices as described above to verify whether the TF-IDF approach can easily detect faked responses. As a first step, we verified whether the TF-IDF discriminates between honest and dishonest responses more efficiently than raw scores; as seen in the correlation matrices, no linearity exists between the two distributions. For this purpose, we used the KL divergence, an entropy-based measure, to compare probability distributions. The values of this measure range from 0 to infinity, where 0 indicates that the two considered distributions (e.g., honest and dishonest) completely overlap. Higher values of the KL-divergence indicate a higher degree of separation between honest- and dishonest-response distributions. The results reported in Tables 2.4, 2.5, and 2.6 show much larger KL-divergence scores between the distributions of honest and dishonest responses for TF-IDF if compared with raw responses. This analysis highlighted that raw honest and dishonest distributions mostly overlap, while TF-IDF distributions show sensible differences with relatively higher KL-D values.

Table 2.4. KL Divergence for CC Dataset

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Raw scores	0.37	0.04	0.54	0.07	1.00	0.08	1.05	0.96	0.03	0.27
TF-IDF	2.93	3.66	3.32	4.51	7.21	0.44	8.21	5.46	4.76	7.30

Note. KL divergence between the response value distributions in the CC dataset of honest and dishonest responses for raw scores and TF-IDF. The closer to 0 the KL divergence is, the less distinguishable the two distributions are (honest and dishonest).

Table 2.5. KL Divergence for JIS Dataset

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Raw scores	0.64	0.05	0.21	0.03	0.88	0.12	0.85	1.02	0.24	0.10
TF-IDF	1.29	2.23	1.08	4.38	2.86	0.44	2.65	3.38	2.95	3.28

Note. KL divergence between the response value distributions in the JIS dataset of honest and dishonest responses for raw scores and TF-IDF. The closer to 0 the KL divergence is, the less distinguishable the two distributions are (honest and dishonest).

Table 2.6. KL Divergence for JIHO Dataset

	Q1	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**
Raw scores	0.58	0.06	0.44	0.11	0.71	0.13	1.09	0.84	0.16	0.23
TF-IDF	7.18	1.40	2.45	3.99	6.39	0.76	2.20	2.41	1.67	3.25

Note. KL divergence between the response value distributions in the JIHO dataset of honest and dishonest responses for raw scores and TF-IDF. The closer to 0 the KL divergence is, the less distinguishable the two distributions are (honest and dishonest).

Subsequently, to determine how the TF-IDF representation of honest responses differs from that of the dishonest ones, we calculated the average for each participant over all 10 items (see Figure 2.3). These results indicate that the average TF-IDF for fakers is significantly higher than for honest respondents. For example, in the JIS dataset (central panel), the distribution of the average TF-IDF is below 2, and we mainly observed higher values (>3) for dishonest responses. A TF-IDF above 3 indicates that responses are faked, with odds ranging from 3.8 to 6.7, depending on the context.

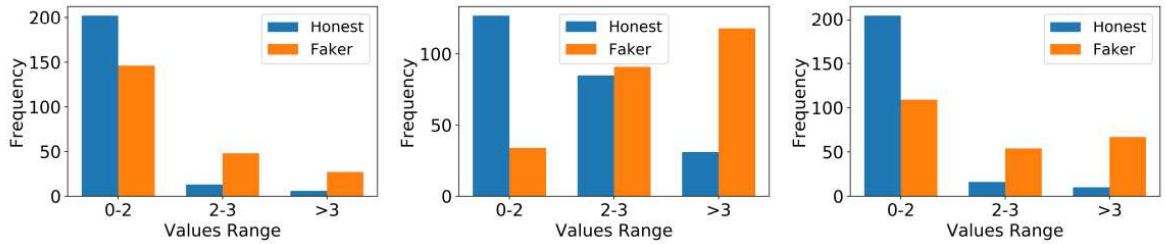


Figure 2.3 Average of TF-IDF Response Values Over All 10 Items for Honest and Dishonest Response. Datasets from left to right: CC, JIS, JIHO. The odds for TF-IDF above 3 are the following: 1/3.8 (for every honest response with a TFIDF >3, there are 3.8 Fakers for CC), 1/6.7 (for JIS), and 1/4.5 (for JIHO). Transforming odds to probabilities yields the following probabilities of dishonesty for TF-IDF >3: 79.2% for CC, 87% for JIS, and 82% for JIHO.

These results show that fakers can be distinguished from honest respondents based on the TF-IDF average over all the questionnaire items' responses.

2.4.2 Performance Evaluation

TF-IDF acts as a novelty detector, with higher values indicating a response that a subject used frequently but all other participants used infrequently. The index is participant-specific and item-specific. However, TF-IDF can be used not only to spot fakers—as the results in Figure 2.3 indicate—but also to identify which items have been faked. As aforementioned, this feature is essential, as no procedure is available to address this problem. As we describe in more detail in Section 2.3, "Material and Methods," the proposed process for item-level faking detection can be summarized as follows:

1. Calculate the TF-IDF score for each response of a target participant.
2. Compare the obtained TF-IDF scores to a threshold specific to each questionnaire item.
3. If the TF-IDF values are outside a specific range, the response is categorized as an anomaly (faked).

We applied this procedure to the three datasets (i.e., CC, JIS, JIHO), yielding the results summarized in Figure 2.4. Given that each participant responded to the 10-item test twice, once in the dishonest condition and once in the honest one, we could identify the items with distorted responses for each participant. As indexed by the probability density function, fakers faked between 7 and 9 items more frequently, with few participants faking 1, 2, or 10 items (see Figure 2.3). We evaluated the performance of the proposed approach first in terms of accuracy (Figure 2.4) and then in terms of precision, recall, and F1 Score (Table 2.7). We define the

accuracy measure as the number of items on a questionnaire correctly identified as faked or nonfaked. Figure 2.4 shows the average accuracy of our TF-IDF approach in various scenarios.

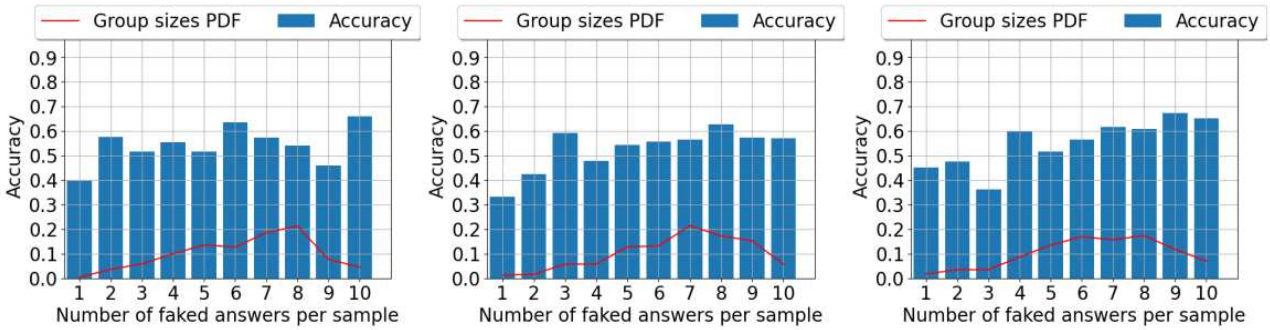


Figure 2.4 Probability Density Function (PDF) of the number of dishonest responses per questionnaire (red line) and per-item accuracy for the faking detection task using the TF-IDF model on various datasets: CC (left panel), JIS (central panel), and JIHO (right panel). The most frequent number of dishonest responses is 7/10. We report the average accuracy in identifying faked responses in each histogram's bars.

We observed that the accuracy remained relatively stable in the 0.4−0.6 range in all three conditions, indicating that faked items can be identified with the same level of accuracy independently of the specific faking strategy the subject used.

Table 2.7 shows the average performance in the three datasets. The results are compared to a simpler distribution-based model (DM), where dishonest responses are identified following the same procedure, but using raw scores. Results indicate a substantial increase in all the evaluative parameters for TF-IDF compared to raw scores. Table 2.7 also shows the precision, recall, and F1 score metrics. They are defined as follows:

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1\ Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{4}$$

Where TP, FP, and FN indicate, respectively, the true positives (i.e., the number of dishonest responses correctly classified), false positives (i.e., the number of responses wrongly classified as dishonest), and false negatives (i.e., the number of dishonest responses our model did not catch), in our classification task (Powers, 2020).

Table 2.7 Performance Evaluation

	Group	Precision	Recall	F1 Score	Accuracy
TF-IDF	child custody (CC)	0.6700	0.4597	0.5046	0.5566
	sales manager (JIS)	0.7178	0.5279	0.5664	0.5642
	humanitarian organization (JIHO)	0.6808	0.5705	0.5877	0.5848
DM (raw scores)	child custody (CC)	0.4735	0.4176	0.4057	0.3516
	sales manager (JIS)	0.4692	0.3237	0.3413	0.3123
	humanitarian organization (JIHO)	0.4966	0.4295	0.4142	0.3422

Note. Performance evaluation of the TF-IDF compared to the distribution-based model based on raw response values. The TF-IDF outperforms the use of raw scores in all the evaluation metrics.

In practice, the TF-IDF can be used to analyze single-subject responses when no honest responses for a suspected faker are available. Figures 2.5, 2.6, and 2.7 show the TF-IDF scores associated with the responses of three of our subjects (one for each dataset). In particular, we report the sequence of responses of three faking subjects (blue line) to each of the questionnaire's 10 items. We compare each response to the distribution of honest responses participants gave for the same questionnaire item. The distribution of honest responses for each item is indicated in the figures with a boxplot showing the mean of each distribution (orange line), the first quartiles (extremes of the boxes), and the 75th and 25th percentile values (upper and lower whiskers). Finally, highlighted in red are the responses where the participants changed their responses when asked to fake. For example, for the participant reported in Figure 2.5 (Participant 5 from the CC group), only Q6 had the same response in the honest and dishonest conditions. In this case, the algorithm accurately detected seven of the nine dishonest responses when selecting the 75th percentile threshold for each item (i.e., when flagging as dishonest the responses that fall outside the 75th percentile of the distribution of honest responses to the same item). In this case, the total accuracy is 8/10 because the algorithm correctly identified all non-faked responses (Q6) as honest but missed two faked ones (Q2 and Q9) incorrectly identified as honest.

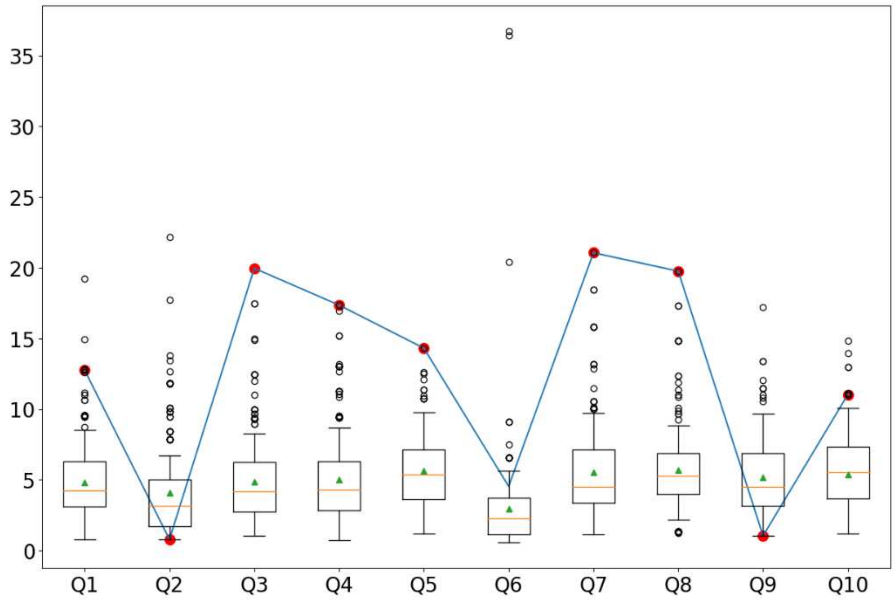


Figure 2.5 TF-IDF Response Values for Participant 5 (Dishonest Responses) compared to the distributions of TF-IDF values of honest respondents in the CC dataset. The red dot indicates the items that the participant faked. In this case, only Q6 received the same response in the honest and dishonest conditions. In this specific case, the algorithm accurately signaled seven of the nine dishonest responses. The total accuracy is 8/10. Two dishonest responses were missed.

Figure 2.6 illustrates the analysis of a single participant for the job interview for a sales manager position. In this case, the subject faked their responses in all but two items, Q4 and Q6, correctly identified as honest, and faked their responses in the remaining ones. Of the dishonest responses, the TF-IDF accurately detected six, incorrectly flagging two of them, Q2 and Q10, as honest, with an overall accuracy of 8/10.

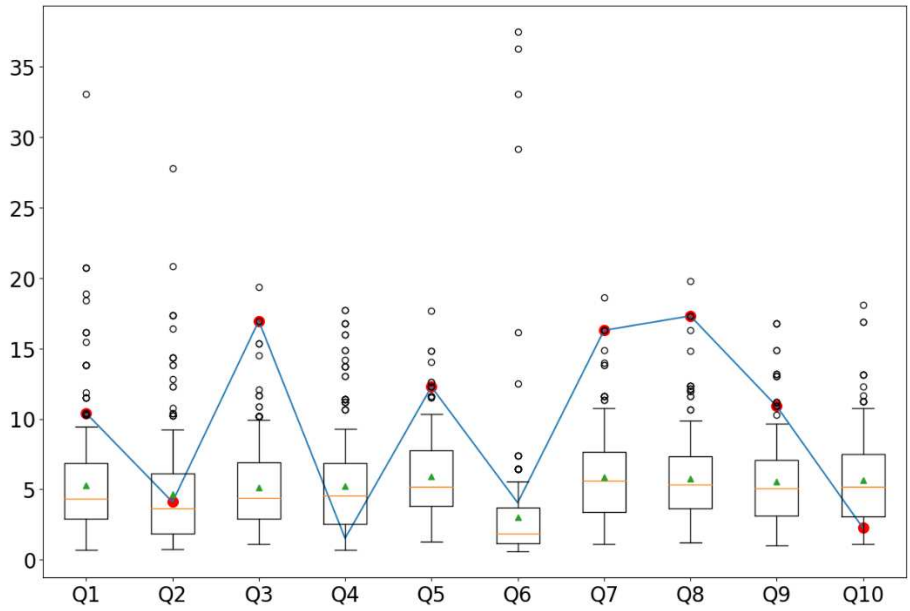


Figure 2.6 TF-IDF Response Values for Participant 48 (Dishonest Responses) compared to the distributions of TF-IDF values of honest respondents in the JIS dataset. The red dots indicate the items that the participant faked. In this case, Q4 and Q6 received the same response in the honest and dishonest conditions (i.e., the subject did not fake). Among the eight dishonest responses, the TF-IDF algorithm accurately spotted six of them. The total accuracy is 8/10.

Figure 2.7 shows a single example of a faker for landing a job in a humanitarian organization. Here, the subject faked their responses in six of 10 items, with the TF-IDF correctly identifying two of them, for an overall accuracy of 6/10 (i.e., it correctly identified the four honest responses and two of the dishonest ones, and it incorrectly classified four of the 10 responses in the questionnaire). Within this evaluation framework, we consider dishonest all responses that are different from the respective honest ones the same participant provided when taking the test honestly. However, as mentioned above, the test's reliability is far from perfect (Rammstedt & John, 2007). Therefore, we believe that with a less strict identification of dishonest responses, the performance evaluation of our approach would yield higher values.

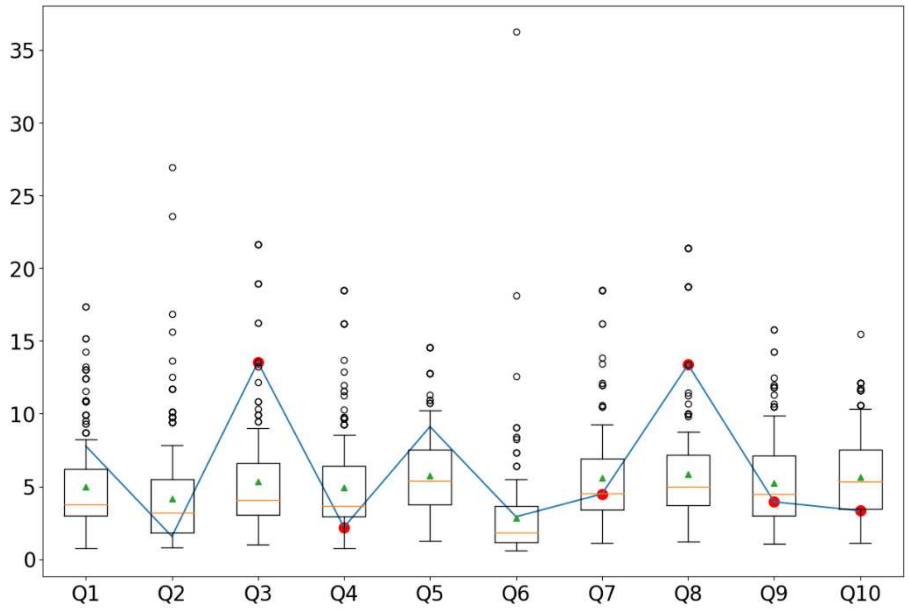


Figure 2.7 TF-IDF Response Values for Participant 4 (Dishonest Responses) compared to the distributions of TF-IDF values of honest respondents in the JIHO dataset. Here, we report a case where the precision is below the reported average. The red dots indicate the items that the participant faked. In this case, Q1, Q2, Q5, and Q6 received the same response in the honest and dishonest conditions (i.e., the subject did not fake). Of the six dishonest responses, two were correctly identified. Q4, Q7, Q9, and Q10 were not reported as faked. The total accuracy is 6/10.

2.5 Conclusion

Psychological tests usually require participants to endorse a statement describing a personality feature or behavior using a Likert scale (e.g., a 5-point scale). Respondents may easily fake responses on such tests by modifying their original, honest responses to achieve an objective (e.g., landing a job). To handle this problem, several subsidiary control scales have been developed, validated, and added to personality scales (e.g., desirability scales, such as the Balanced Inventory of Desirable Responding (BIRD; Hart et al., 2015). However, these procedures help—at best—spot fakers but do not help specifically identify faked responses. In short, identifying items that have been faked (e.g., to obtain a job) is currently a major unsolved issue.

This complexity is further exacerbated by individual differences and the nuanced nature of simulation, which may manifest subtly with various degrees of intensity. Furthermore, consideration must be given to the fact that the test's assumptions and objectives can directly influence the simulation itself. For instance, exaggerating extroversion may be strategic for a sales manager role, while overstating conscientiousness might be more relevant for a banking position. Currently, no technique is available to identify the specific faked responses. Existing control scales help identify only the participant's general propensity to fake and do not pinpoint where faking occurs.

For these reasons, we propose using a new technique inspired by original research in information retrieval. The technique is known as the term frequency-inverse document frequency (Baeza-Yates & Ribeiro-Neto, 1999) (TF-IDF) and is at the base of search engine technologies. When applied to search engines, this technique helps identify the most relevant web page concerning a query. This method has been used in cognitive neuropsychology to model semantic disorders (Sartori & Lombardi, 2004).

One of the main contributions of this approach is that, unlike traditional item analysis, where identical responses to the same item from two subjects are evaluated in the same way, the TF-IDF modeling strategy takes a different approach. In this method, two identical responses (from two participants) to the same item may yield different evaluations. This variation is based on other participants' responses to the same items and the subject's responses to the questionnaire's other items. In other words, TF-IDF aggregates a unique measure that considers the distribution of group responses to a target item and the subject's response style (e.g., how frequently they also select a specific response value in other items). The TF-IDF index, as presented, can be regarded as a novelty detector, where a high TF-IDF score for a

participant's response to an item would indicate that their response deviates significantly from those of honest respondents. Additionally, unlike other lie detection methods that require data from both honest and deceptive respondents to classify a new subject, this technique only necessitates honest responses from a representative sample (see Figures 2.5, 2.6, and 2.7).

TF-IDF, as presented here, can be considered a novelty detector. If the TF-IDF of a subject's response to an item is abnormally high, the participant's response style does not belong to the sample of honest respondents. To verify the effectiveness of TF-IDF in identifying faked responses, we asked 694 participants to respond twice to a short Italian version of the Big Five test (10-item Big Five questionnaire; Guido et al., 2015). The first time, we asked the participant to fake to achieve an advantage in one of three scenarios: (a) to obtain child custody, (b) to obtain a sales manager position, or (c) to obtain a position in a humanitarian organization. The second time, we asked the participant to respond honestly.

The exploratory analyses, performed on raw scores, indicated that:

- faking (good) resulted in an increase in average response values on all but one of the 10 items and
- faking in different contexts, as expected, resulted in distortions of responses to different test items.

In evaluating the proposed TF-IDF approach, we observed the following:

- a TF-IDF based representation of item responses leads to a better separation between honest and dishonest responses than raw scores;
- the average (of the 10 items of the considered questionnaire) TF-IDF response values of the honest respondents were lower than that of the dishonest ones;
- lower TF-IDF values characterize honest responses to single items, and larger values are characteristic of dishonest ones;
- the proportion of dishonest responses associated with large TF-IDF representations (e.g., above the 75th percentile of the honest-value distributions) was consistently larger than that of their honest counterparts;
- when applying the TF-IDF for item-level faking-detection, precision (the percentage of faked items correctly identified of all the faked items) was between 67% and 71%, indicating that TF-IDF correctly identifies most of the faked single-item responses. The same faking detection procedure, based on raw response values, instead yielded much lower precision values (see Table 2.7).

All the results indicate that TF-IDF can help catch deceptive responses at the single-item level with unprecedented accuracy. The identification of fakers and faking at the single-item level is achieved without using any control scale, only by capitalizing on one distinguishing feature of TF-IDF. It is worth noting that TF-IDF, as a novelty detector, allows for spotting a faker if a group of honest respondents is available. This feature has interesting practical implications given that faking is context-specific and depends primarily on the faker's strategy, which depends on the objective that faking is expected to achieve (e.g., the custody of a child or landing a job position).

Finally, we have given examples of how the TF-IDF can be used to analyze a single participant's responses (see Figures 2.5, 2.6, and 2.7). When a suspected faker's responses are analyzed, we do not have the ground truth of their honest responses. When the TF-IDF falls outside a given threshold (e.g., the 75th percentile of the TF-IDF of the honest responses), it may help determine whether a response is faked.

Although the present study was validated on the 10-item Big-five test questionnaire, its validity should also be evaluated for other tests. For this reason, although it is not encompassed within this dissertation, a substantial portion of this project has been devoted to the collection of data and the implementation of this methodology to other tests such as: Dirty Dozen Questionnaire (N=761) (Jonason & Webster, 2010); PTSD Checklist for DSM-5 (N=201) (Blevins et al., 2015); Parental Reflective Functioning Questionnaire (N=339) (Luyten et al., 2017); Personality Inventory for DSM-5 (N=512) (Krueger et al., 2013); Prospective and Retrospective Memory Questionnaire (N=702) (Crawford et al., 2003); Generalized Anxiety Disorder (N=529) (Stein & Sareen, 2015); Impact of Event Scale-Revised (N=179) (Weiss, 2007); short Personality Inventory for DSM-5 (N=519) (American Psychiatric Association, 2013); Short Dark Triad (N=241) (Jones & Paulhus, 2014); International Adjustment Disorder Questionnaire (N=225) (Shevlin et al., 2020); Patient Health Questionnaire-9 (N=529) (Löwe et al., 2004); Negative Acts Questionnaire-Revised (N= 532) (Einarsen et al., 2009). Moreover, future research can build on this study to explore comparisons with real clinical subjects, broaden validation across diverse questionnaires and demographics, and refine TF-IDF techniques for discerning various forms of deception. Such efforts will ensure an extensive scope of application, bolstering the method's overall efficacy across real-world domains. The versatility and standardized nature of the tool underscore its groundbreaking impact potential. Since the role of the presentation order of the honest and dishonest condition has not been investigated here, future studies should balance the sample and provide further information on the possible priming effect. Moreover, in this study, we asked participants to respond to three

scenarios and fake their test responses accordingly. As reported in this research, the method's efficacy should be confirmed with high-stake test-takers in natural settings rather than in low-stake instructed faking conditions.

Chapter 3.
Drugs Don't Work

3.1 Abstract

This research explores deception detection concerning illicit drug consumption in the workplace by analyzing mouse movement dynamics during an online computerized double-choice task. A total of 122 workers were assigned to either the "honest group," consisting of individuals who have never worked under the influence of drugs, or the "deceptive group," comprising those who have. Participants were required to categorize sentences related to drug use at work using a computer mouse. The data collected included errors, response times, and various temporal and spatial mouse parameters. These data were employed to investigate significant differences between honest and deceptive conditions. Across two independent studies, the results indicate that during the task, participants in the deceptive condition trace broader trajectories with their mouse compared to the honest group, deviating more significantly from the ideal trajectory, and commit a greater number of errors. The logistic regression classification task achieved an accuracy level of 72%, while 64% accuracy was observed in the machine learning models. The limitations and potential benefits of this approach are discussed.

Author note.

This research was developed and conducted during the research period abroad at the Department of Marketing, Innovation and Organisation, Faculty of Economics and Business Administration from Ghent University, Belgium. This research was conducted in collaboration with and under the supervision of Prof. Adriaan Spruyt. Financial sponsorship was provided by BOF.COF.2019.0005.01, BOFSTG2019006301, and BOF24Y2019000601.

3.2 Introduction

3.2.1 Drug Use in the Workplace

The consumption of psychoactive substances in the workplace poses a significant challenge for many countries, with a considerable proportion of individuals who engage in the use of illegal drugs being employed across various sectors (European Monitoring Centre for Drugs and Drug Addiction, 2022; Frone, 2013). Statistics suggest that a notable percentage of European workers are expected to encounter issues related to the consumption of alcohol or drugs. For example, it is estimated that approximately 5% to 20% of the working population in Europe experiences serious problems associated with alcohol use. The European Monitoring Centre for Drugs and Drug Addiction's recent analysis highlights that, beyond impacting the individual health of employees, the consumption of alcohol and drugs in the workplace presents a substantial problem. This is attributed to increased workplace accidents and injuries, higher rates of absenteeism, and inappropriate behaviors, culminating in a considerable economic burden for employers, governments, and society.

The most common workplace drug testing methods, subject to legislative variations, include blood testing, breath analysis, and urine analysis (Ones & Viswesvaran, 2001). While most European countries have implemented legislation or agreements to prohibit or regulate the consumption of alcohol and drugs in the workplace, these measures vary significantly in national culture, awareness, and prioritization of the issue. Additionally, the testing of bodily samples for alcohol and drugs in the employment context raises profound moral, ethical, and legal concerns. These issues necessitate careful deliberation to determine fairness and appropriateness in such situations. Workplace drug testing is a sensitive topic, primarily due to the challenge of striking a balance between safety and productivity requirements and the essential need to prevent invasions of privacy and discrimination. For instance, blood samples are infrequently used in testing procedures as they may be perceived as a more invasive breach of privacy. For example, the constitutions of Belgium and Finland assert the indivisibility of fundamental rights, such as the right to privacy, and uphold that individuals cannot consent to waive such rights (Shahandeh & Caborn, 2003). In contrast, in the Czech Republic, employee drug testing is mandatory, and refusal to comply constitutes a breach of legal duty, enabling employers to enforce workplace bans. Conversely, in Austria, employees who refuse drug testing do not face consequential actions such as dismissal (Corral et al., 2012). Consequently, in instances where drug testing is mandated, employees who engage in drug use at the workplace are likely to endeavor to conceal such behavior (El-Bassel et al., 1997). A common

strategy adopted by these individuals involves abstaining from drug use in the days leading up to the test to delete traces of the substances from their bodies. This tactical approach underscores the complex challenges inherent in effectively monitoring and addressing workplace drug use.

To address pertinent concerns, the employment sector is increasingly adopting paper-and-pencil tests to assess both current and prospective employees for their propensity towards drug abuse. A prime example of this methodology is the Drug Abuse Screening Test (DAST; Skinner, 1982). Contrasting with physiological tests like blood and urine analyses, which identify active substance use, these questionnaires are designed to assess the use or potential use for drug abuse among job applicants, with a focus on predicting or confirming substance abuse. These tests, through targeted inquiries, aim to glean insights into the candidates' attitudes, values, and perceptions regarding drug or alcohol abuse.

However, a critical limitation of most drug use assessment scales is their transparency regarding their purpose; the intent to measure potential drug use is readily apparent to the respondent. Consequently, this transparency makes the approach susceptible to deceitful responses, as the explicit nature of the questions facilitates simulation. This approach is deemed vulnerable to malingering due to the inherent design of these questionnaires, which lack mechanisms to resist faking. Consequently, they exhibit poor sensitivity in detecting deception (Ramachandran et al., 2019). Given the high stakes and the lack of effective methods to identify simulation attempts in such contexts, this research endeavors to validate a novel methodology for detecting deceptive responses in workplace drug use tests.

3.2.2 The Current Study

To achieve this goal, this study aims to implement and explore the kinematic analysis of mouse movements, a measure used to understand the dynamics of mental processes (Abney et al., 2015; Barca & Pezzulo, 2012; Freeman et al., 2016; Quétard et al., 2015), including processes involved in deception production (Duran et al., 2010; Hibbeln et al., 2014; Mazza et al., 2020; Monaro, Gamberini, et al., 2017). This tool records the motor dynamics of the hand, which are the -*x* and -*y* coordinates of the computer mouse, during an online task that involves answering double-choice questions about drug use at work. Please refer to Section 1.2.3, "Cognitive Lie Detection Technique," in Chapter 1 for a detailed exploration of the employed technique.

Two distinct participant groups were involved in the online study. The first group comprises individuals who have neither used drugs at work nor worked under the influence of drugs. Conversely, the second group includes participants who have, at least once, used drugs

at work or worked while under their influence. Participants in both groups were tasked with classifying sentences as "TRUE" or "FALSE," explicitly pertaining to drug use in the workplace (e.g., *"I used drugs at work"*). However, while the first group was instructed to declare their abstinence from workplace drug use truthfully, the second group was directed to claim they had never used drugs at work falsely. Sentences were displayed at the screen's center, triggered by the participant pressing a square button located at the screen's bottom center. After the sentence presentation, participants must select "TRUE" or "FALSE" by clicking one of the two options at the screen's top right or top left corners using their mouse. Mouse movements are tracked from the moment of pressing the square button until the selection of the response option.

The hypothesis posits that individuals who lie about their drug use at work will exhibit increased reaction times and a higher rate of errors in their responses compared to those who truthfully answer that they do not use drugs at work. Additionally, an analysis of mouse kinematics is expected to reveal broader trajectories in the movements of participants who lie in contrast to those who answer truthfully.

Experiment 1

3.3 Materials and Methods

The study was preregistered on Open Science Framework (https://osf.io/vuc7r/). A between-subjects design was utilized for this study, which was conducted entirely online and consisted of a pre-screening questionnaire and the mouse tracker task.

3.3.1 Experimental Procedure

Pre-screening Questionnaire. The initial phase of the research encompassed a pre-screening questionnaire to select appropriate participants for the mouse tracker test. Specifically, the aim was to recruit two distinct groups: individuals who had worked under the influence of drugs and those who had not. A brief pre-screening questionnaire using Prolific's Survey Builder (prolific.co) was administered to reach a wide participant pool. This allowed the publication of a 1-5 question survey on Prolific. Data was collected from 1000 participants, who were asked to respond to the following five questions:

1. Have you ever used drugs during working hours?
2. Have you ever been under the influence of drugs at work?
3. Have you ever violated safety regulations at work?

4. Have you ever forgotten to follow safety regulations at work?
5. Have you ever had an accident at work due to a violation of safety regulations?

Participants were provided with the following response options: *"Yes," "No,"* and *"I have never had any work experience."* Participants were instructed to answer *"Yes"* even if they only did it once. The latter option served as a control measure to exclude participants who lacked work experience. The final three questions were incorporated as screening measures for a parallel study on safety behavior in the workplace. Given the similarities between the two studies, these screening questions were included to prevent duplicate participation and to avoid the possibility of the same participant engaging in two experiments. Those participants who responded affirmatively to only questions 1 and 2 were classified as belonging to the group of liars. At the same time, those who had never used drugs at work (answered *"No"* to questions 1 and 2) were categorized as honest participants. Given that Prolific automatically records the Prolific ID of participants, this information was used to contact them through the Prolific platform and invite them to participate in the second part of the study, which involved the mouse tracker task.

Mouse Tracker Task. The experiment was constructed using the Lab.js platform (lab.js.org) and executed on a self-hosted JATOS server (jatos.org). Upon reading the general instructions and providing consent to participate, participants responded to ten practice sentences to familiarize themselves with the task. It is important to emphasize that each practice statement was presented only once, and the order of presentation was randomized for each participant. These sentences were deliberately crafted with known answers and included examples such as *"I have a name"* or *"At this moment, I am flying to Mars."* For a comprehensive list of these sentences, please refer to Table 3.1.

Table 3.1 Practice sentences.

False answer	**True answer**
I am an alien.	I have a name.
I have five eyes.	I have a nose.
At this very moment, I am driving a train.	At this very moment, my heart is beating.
At this very moment, I am sleeping.	At this very moment, I am awake.
I have wings.	I am alive.

Note. Ten statements related to the participant's current situation (5 requiring a TRUE response and 5 requiring a FALSE response).

A square button was displayed before each sentence in the center of the lower part of the computer screen (See Figure 3.1). The sentence appeared in the screen's center after pressing this button. Participants had to categorize sentences as true or false by pressing one of two buttons in the screen's upper right or left corner. The left button was labeled *"FALSE," and the right button was labeled "TRUE."* Participants were instructed to prioritize both accuracy and speed in their responses. However, in the event of an error or if their response exceeded 4000 milliseconds, an error message promptly notified them of the issue.

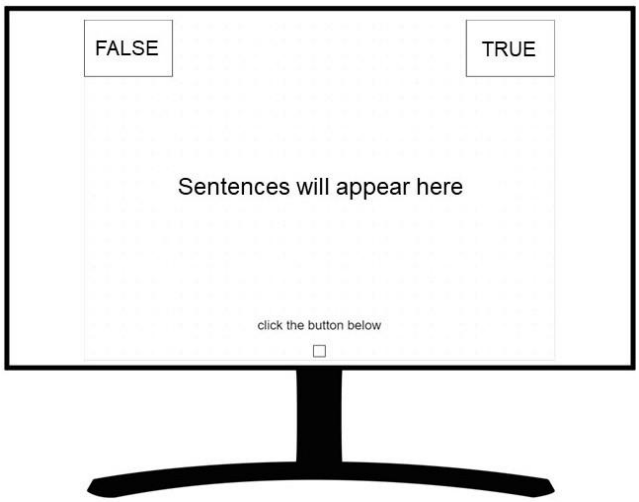


Figure 3.1 Example of the computer screen as it appeared to the subjects.

Following the practice task, participants proceeded to the main task, which involved imagining themselves in an online job interview scenario and responding to sentences related to drug use at work. The deceptive group received instructions to simulate never having used drugs, while the honest group answered truthfully because they had indeed never used drugs at work. The format of sentence presentation remained consistent with the practice task, except for the content of the sentences. The main trial block comprised 30 trials in total: 20 target statements, including 10 requiring a TRUE response and 10 necessitating a FALSE response (see Table 3.3), as well as 10 control statements, with 5 requiring a TRUE response and 5 requiring a FALSE response (refer to the Table 3.2). Furthermore, within each response category (TRUE vs. FALSE), precisely half of the target statements pertained to the use of drugs at work, while the other half related to being intoxicated at work. This differentiation was maintained (even in the pre-screening phase) to encompass participants who might have used drugs shortly before starting their work activities.

Table 3.2 Control sentences.

False answer	True answer
I am the President of the USA.	I am a human being.
At this very moment, I am surfing in the ocean.	At this very moment, I am working on a computer task.
I use a pencil to respond to this computer task.	I use the computer mouse to respond to this task.
Yesterday, I walked on the moon.	I am sitting in front of a computer screen.
At this very moment, I am climbing Mount Everest.	At this very moment, I am responding to some sentences.

Note. Ten statements related to the participant's current situation (5 requiring a TRUE response and 5 requiring a FALSE response).

Table 3.3 Trial sentences.

False answer	True answer
I have used drugs at work.	I have never used drugs at work.
While working, I have used drugs.	While working, I have never used drugs.
Consuming drugs at work sometimes happens.	Consuming drugs at work never happens.
During working hours, I sometimes use drugs.	During working hours, I never use drugs.
It has happened before that I used drugs at work.	It has never happened that I used drugs at work.
Working under the effect of drugs sometimes happens.	Working under the effect of drugs never happens.
I have worked under the influence of drugs.	I have never worked under the influence of drugs.
I have been under the influence of drugs at work.	I have never been under the influence of drugs at work.
Sometimes I do my job under the influence of drugs.	I never do my job under the influence of drugs.
I have been intoxicated at work.	I have never been intoxicated at work.

Note. Twenty statements related to drug use at work (10 requiring a TRUE response, while the other a FALSE response. The first five sentences ask participants if they have ever used drugs while at work, while the other five ask if they have ever been intoxicated while at work.

Liars were expected to provide truthful responses on the control trials, which consisted of factual statements about their current situation, and to provide deceptive answers on the target trials, which involved statements about drug use at work. In contrast, honest participants were instructed to respond honestly to both types of trials. As in the practice task, each statement appeared exactly once, and the order of presentation was randomized for each participant.

In both the honest and deceptive condition, participants were required to respond accurately and quickly. However, if they made an error or took too long to respond (more than 4000 milliseconds), a message would be displayed, informing them that they had given an incorrect answer or taken too long. For the honest and liar participants, error feedback was provided on both target and control trials. Deceptive participants received error feedback on target trials if they answered truthfully (e.g., sentence: *"I used drugs at work";* response: *"TRUE"*). Error feedback was also triggered for them on control trials if their response was wrong (e.g., sentence: *"I am a human being"*; response: *"FALSE"*). Honest participants received error feedback if they answered inaccurately on target and control trials. For example, if they answered *"TRUE"* to the target statement *"I used drugs at work"* or to the control sentence *"I am an alien."*

To ensure that the length of sentences did not introduce systematic noise into the data, sentences of similar length were carefully selected. For each short sentence in the target group, at least one sentence of the same length was included in the control group. Furthermore, a *t*-test confirmed that the difference in the number of letters between each set of statements was not statistically significant.

Once participants completed the task, they were asked to provide additional information such as their gender, age, and dominant hand. Additionally, a final question was included to assess honesty and ensure that only participants who had never used drugs were included in the honest group. This step excluded individuals who may have used drugs during work and provided deceptive responses during the pre-screening and main tests. The question was explicitly directed at the honest group and inquired whether they had lied about drug use at work. It was made clear that the job interview had concluded, and truthful responses were necessary. Participants had the option to reply with *"No, I was honest. I've never used drugs during work!"* or *"Yes, I lied to try to create a positive impression!"*

At the end of the study, five questions about the usability of the task were included to gather feedback from the participants. This feedback was valuable for improving the task in future experiments. Subsequently, participants were automatically redirected to the Prolific platform.

3.3.2 Hypotheses

Mouse tracking data can yield various parameters for analysis. Despite the explorative nature of the study, in contrast to the control group, we anticipate that the deceptive group will exhibit the following hypotheses (H1 to H4):

H1) Trajectories in the deceptive group are expected to deviate more from the ideal trajectory, represented by the straight line connecting the start and end points.

H2) The deceptive group is expected to demonstrate lower values for temporal variables such as reaction time, acceleration, or speed when compared to truth-tellers.

H3) The deceptive group is expected to make more errors during the task.

H4) The deceptive group may display a general difficulty in responding to control questions due to the continuous shift between providing deceptive responses (target questions) and truthful responses (control questions).

3.3.3 Collected Measures

In addition to the number of errors, distances, velocities, acceleration, and mouse-tracking measures for each trajectory were computed. Those were calculated starting from the -*x* and -*y* coordinates and the timestamps. Specifically:

1. xpos_max; xpos_min; ypos_max; ypos_min: Maximum and minimum x- and y-position
2. MAD: Signed Maximum absolute deviation from the direct path connecting the start and end point of the trajectory (straight line). If the MAD occurs above the direct path, this is denoted by a positive value; if it occurs below, by a negative value.
3. MAD_time: Time at which the maximum absolute deviation was reached first.
4. MD_above; MD_above_time: Maximum deviation above the direct path; Time at which the maximum deviation above was reached first.
5. MD_below; MD_below_time: Maximum deviation below the direct path; Time at which the maximum deviation below was reached first.
6. AD: Average deviation from the direct path.

7. AUC: Area under curve, the geometric area between the actual trajectory and the direct path where areas below the direct path have been subtracted.
8. xpos_flips; ypos_flips: Number of directional changes along x-axis and y-axis.
9. xpos_reversals; ypos_reversals: Number of crossings of the x-axis and y-axis.
10. RT: Response time, the time at which tracking stopped.
11. initiation_time: The time at which the first mouse movement was initiated.
12. traj_duration: Time from the first movement to the click on the answer
13. idle_time: Total time without mouse movement across the entirety of the trial.
14. Hovers; hover_time: Number of periods without movement in a trial; Total time of all periods without movement in a trial.
15. total_dist: Total distance covered by the trajectory.
16. vel_max; vel_min; vel_max_time; vel_min_time: Maximum and minimum velocity; Time at which maximum and minimum velocity occurred first.
17. acc_max; acc_min; acc_max_time; acc_min_time: Maximum and minimum acceleration; Time at which maximum and minimum acceleration occurred first.

3.3.4 Exclusion Criteria

Participants who fell into the following categories were excluded:

a. Those claiming in the pre-screening questionnaire to have no work experience.
b. Participants responding YES to questions 3, 4, and 5 in the pre-screening questionnaire.
c. Participants answering NO to questions 1 or 2 in the pre-screening questionnaire.

For participants assigned to the honest condition, if they responded affirmatively to the final lie check question (*"Yes, I lied to try to create a positive impression!"*), they were also excluded. Exclusions based solely on the target and control trials (excluding practice trials) were carried out for participants who met the following criteria: Made a mistake or failed to respond within the response deadline of 40%[2] or more incorrect responses out of 30 trials with

[2] This percentage is based on the probability of obtaining a given number of correct responses if a participant is answering at chance level. Given that, for each trial, the probability to provide a correct answer is 0.50, then one need to have 19 out of 30 correct responses to perform significantly better than chance ($p = 0.049$).

a response latency below 4000 milliseconds (i.e., a minimum of 19 correct responses out of 30 trials with a response latency below 4000msec). Additionally, participants were excluded if their mean response latency on trials with a proper response and a response latency below 4000 milliseconds exceeded the mean of their respective condition by more than 2.5 standard deviations. To maintain the study's anticipated sample size of 122 participants (See Section 3.3.5 "Participants"), those excluded based on the predetermined selection criteria were replaced with new participants. This replacement ensured the study's integrity and the attainment of the desired sample size.

3.3.5 Participants

The sample size for the study consisted of 1000 participants for the pre-screening survey and 122 participants for the mouse tracker task, similar to other lie detection research based on mouse tracker analysis (Mazza et al., 2020). An a priori G*Power analysis determined that a sample size of 122 was sufficiently large to achieve a statistical power of $(1-\beta) = 0.90$ in an independent *t*-test, with a significance level of 0.05 and a medium effect size ($d = 0.60$). Data collection was concluded upon obtaining 122 valid responses. Any necessary exclusions were compensated by collecting additional responses until 122 valid responses were received.

Recruitment of participants was conducted through the online platform Prolific, and all assessments were administered in English. Participation was limited to Desktop PCs or Laptops, with the use of a trackpad not permitted. The eligibility criteria for the pre-screening questionnaire were as follows: First Language: ENGLISH; Primary Language: ENGLISH; Minimum age: 18; Country of residence: UK or USA; and Safety procedures at work screening criteria provided by Prolific: *"Occasionally my job task requires the use of a safety procedure"* or *"Sometimes my job involves compliance with a safety procedure"* or *"Almost all my job tasks are guided by safety procedures."*

Participants in the study were kept unaware of its true purpose. All participants provided the informed consent before the experiment. The final study sample comprised 122 participants (53 female and 1 non-binary participant). All participants were over 18 ($M = 38.84$, $SD = 11.10$). Allocation to the honest or deceptive group was determined based on the responses provided in the Prolific Survey Builder. 61 participants were assigned to the honest group, while another 61 were allocated to the deceptive group. Chi-square analysis revealed significant differences in gender distribution between the groups ($\chi^2 = 8.13$, $p = 0.017$). More specifically, males were more prevalent in the deceptive group, $p < 0.001$, while females were significantly more

represented in the honest group, $p < 0.001$. No significant differences were observed in the distribution of non-binary individuals between the two groups.

The study's duration was approximately 1 minute for the pre-screening questionnaire and roughly 7 minutes for the mouse tracker task. Participants received compensation of £9.00 per hour for completing the pre-screening questionnaire and the mouse tracker task. Eligible participants who had completed the pre-screening questionnaire were contacted again through Prolific, using their Prolific ID, to participate in the mouse tracker task.

3.3.6 Data Pre-Processing

For each remaining participant, we retained only the data corresponding to correct trials with reaction times below 4000 milliseconds and within 2.5 standard deviations of the participant's mean reaction time. Subsequently, data processing involved utilizing the -*x* and -*y* coordinates and timestamps to calculate temporal and spatial variables, employing the Mousetrap R package. Notably, as the study was conducted entirely online, it is worth mentioning that four challenges needed to be addressed:

1. Labjs accumulated timestamps cumulatively throughout the task. For instance, if the initial trial started at 0 seconds and ended after 3 seconds, the subsequent trial did not reset, starting from the 3rd second and progressing accordingly.

To rectify this, the initial step involved reconfiguring the data so that the start of each task corresponded to zero seconds.

2. Participants completed the task on different computers equipped with monitors of varying sizes and, consequently, different pixel ratios. This divergence implies that the mouse trajectories generated cannot be directly compared across participants.
3. The square button that shows the responses has been designed to be 20*20 pixels in size, making it easy for all participants to press it. As a result, multiple points are present within this square that can be pushed by participants to activate the program's response, resulting in varying starting points with different coordinates among participants. The same holds for the coordinates collected when the participant presses a point within the TRUE or FALSE boxes to register their response.

For these reasons, the coordinates were spatially normalized through the *align_start_end* function of the Mousetrap R package to allow the starting and ending points to match among participants. This aligned all trajectories to share a standard initial and final coordinate. Moreover, trajectories were flipped to the right side using the *mt_remap_symmetric*.

4. For the same reason expressed in point 2, various computers recorded mouse trajectory timestamps with varying degrees of precision.

The function *mt_resample* of the Mousetrap R package was implemented to manage this point. This function resampled trajectories to ensure that samples occurred at constant intervals of a specified length. Furthermore, time-normalized trajectories were computed using a constant number of equally sized time steps through the *mt_time_normalize* function. By default, 101-time steps were used to maintain consistency in the temporal aspect of the data analysis.

3.3.7 Features Computation

Following data processing, a comprehensive set of mouse-tracking measures was computed for each trajectory, utilizing the Mousetrap R package. Initially, the *mt_derivatives* function was employed to calculate the distance traveled, velocity, and acceleration for each recorded position in the trajectory. Distance was determined as the Euclidean distance between successive coordinates, while velocity was computed as the distance covered per time interval. Acceleration represented the change in absolute velocity, normalized with respect to time. Subsequently, several additional mouse-tracking measures for each trajectory were computed manually or by utilizing the *mt_measures* function from the Mousetrap R package. For each variable, the responses were averaged for each participant. The final dataset included 32 variables.

3.4 Data Analysis

3.4.1 Visual Comparison of Trajectories

As an initial phase of the analysis, we graphically represented the trajectories of the two distinct groups. Figure 3.2, comprising four panels, delineates these trajectories. Panel A displays the average trajectories for both groups in response to control questions. Conversely, Panel B illustrates the trajectories for target questions. A modest disparity is observed between the two groups, especially in response to control questions; however, this variance is less

marked than initially hypothesized. Notably, both groups' trajectories— honest and deceptive— show a nearly congruent deviation from the ideal trajectory in the target questions. Subsequently, we depicted the average trajectories for control and target questions, segregated according to the group. Panel C shows the average trajectories for the honest group, while Panel D focuses on those for the deceptive group. The data reveals that both groups exhibit significant discrepancies between their responses to control and target questions. Notably, the trajectories for the target questions are markedly broader than those for control questions. Contrary to our initial assumption that this disparity would be more pronounced in the deceptive group, the results seem to indicate a conspicuous difference in both groups.

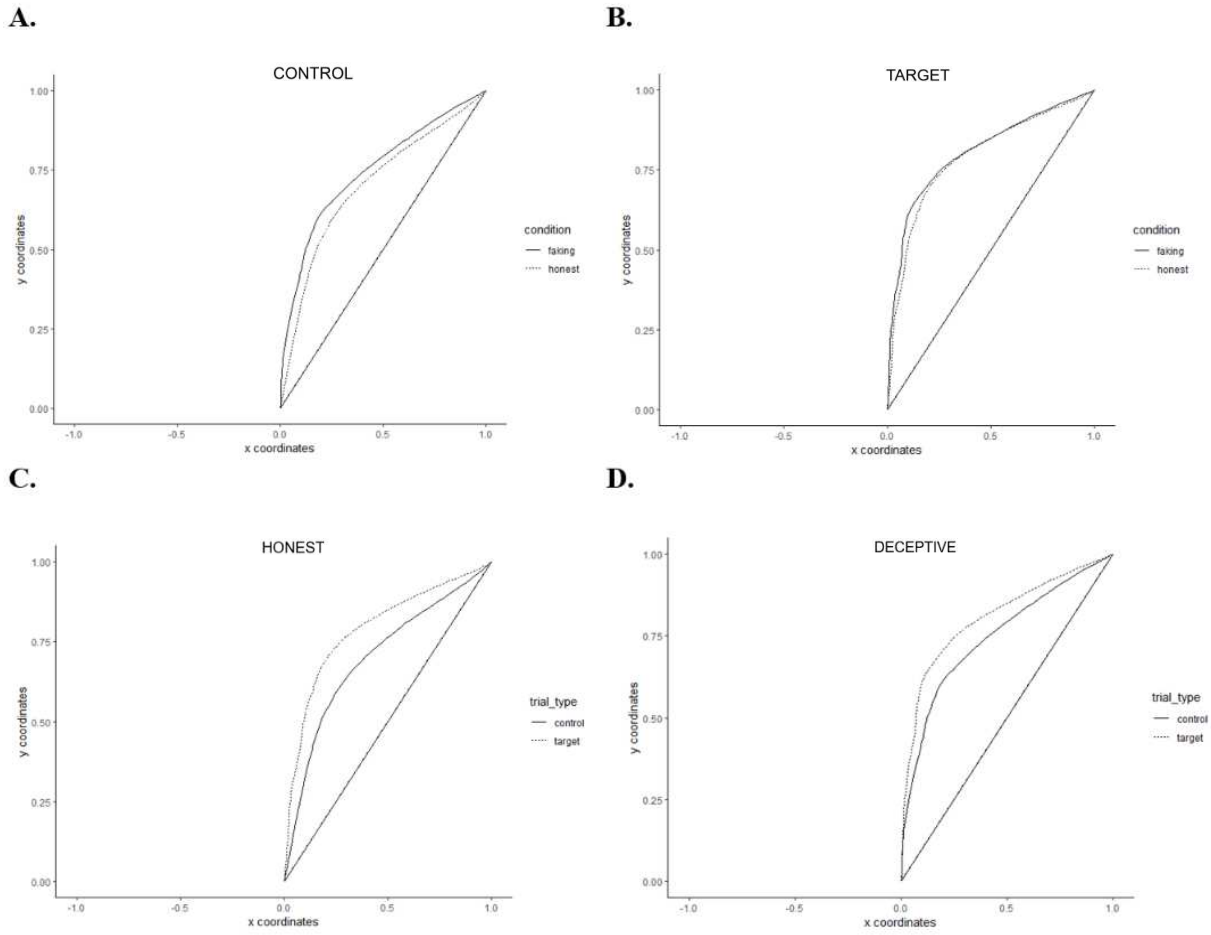


Figure 3.2 Average trajectories for control [Panel A], target [Panel B], honest [Panel C], and deceptive group [Panel D].

3.4.2 Features Selection

A feature selection process was implemented to augment the precision and generalizability of our future predictive models. This process aims to explore those independent variables that exhibit the most significant correlation with the dependent variable (condition: honest or deceptive) while ensuring minimal inter-correlation among other variables.

Initially, we employed the point-biserial correlation coefficient method to assess the relationships between the variables and the condition. This statistical approach was chosen for its ability to evaluate the strength and direction of association between a dichotomous variable and continuous variables. This analysis identified several key variables: reaction time (r_{pb} = -0.25), initiation time (r_{pb} = -0.23), and the number of directional changes on both the y-axis (r_{pb} = 0.19) and x-axis (r_{pb} = 0.17) in addition to ypos_reversal (r_{pb} = 0.18). These variables demonstrated the most substantial correlations with the condition variable.

The Correlation-based Feature Selector (CFS) was utilized to evaluate the attributes further. This method calculates each feature's merit by considering its predictive ability alongside the degree of redundancy among features. For the search strategy, the Greedy Stepwise algorithm, recognized for its efficiency in feature selection, was employed. This algorithm sequentially selects features based on their correlation strength with the model's output. Out of 32 independent variables analyzed, the features ultimately chosen by the algorithm were "initiation_time" and "ypos_flips."

3.4.3 Statistical Analysis

An independent *t*-test was conducted to investigate the underlying differences between the two conditions more thoroughly. This analysis revealed significant disparities between the two independent groups in reaction times (t = -2.83, p = 0.005, Cohen's d = -0.513). Notably, the average reaction time for the deceptive group was 1856.42 milliseconds (SD = 298.36), while that for the honest group was 2003.30 milliseconds (SD = 339.27). This suggests that individuals in the deceptive group reacted more quickly than those in the honest group, a disparity contrary to our initial hypotheses.

Of particular interest was also observing the comparison between the two experimental groups and the question type: control and target. Considering the normality distribution of the RT variable and the gender imbalance between groups, we investigated the impact of experimental condition and question types on reaction times, accounting for gender imbalance between groups through an Analysis of Covariance (ANCOVA). The assumption checks for the ANCOVA confirmed the appropriateness of the analysis. Homogeneity of variances, as assessed by Levene's test, showed no significant violation, F (3, 240) = 0.528, p = 0.663. The normality of residuals, tested by the Shapiro-Wilk test, was also satisfactory (p= 0.305), and the Q-Q plots supported this finding. The ANCOVA results revealed significant main effects of the condition, F (1, 239) = 11.051, p = 0.001, η^2_p = 0.044, and trial type, F (1, 239) = 24.44, p < 0.001, η^2_p = 0.093, on reaction times. Additionally, gender as a covariate showed a significant

effect, F (1, 239) = 4.797, p = 0.029, η^2_p = 0.020. However, the interaction between condition and trial type was not significant, F (1, 239) = 0.125, p = 0.724, η^2_p = 0.001. Post hoc comparisons using Tukey's test revealed significant mean differences in reaction times, particularly within the deceptive group between control and target trial type (ΔM = −178.3, SE = 54.9, t (239) = −3.25, p = 0.007, d = −0.588) and between the control and target trial type in the honest condition (ΔM = −205.8, SE = 54.9, t (239) = −3.75, p = 0.001, d = −0.678). Interestingly, a significant difference also emerged between groups in comparing target sentences (ΔM = −145, SE = 55.4, t (239) = −2.62, p = 0.046, d = −0.478), with the honest group reaching longer reaction times. These findings suggest that the type of trial and the experimental condition significantly influenced participants' reaction times.

3.4.4 Classification Task

All the variables were then utilized to develop a model to differentiate between honest and deceptive participants. A forward stepwise logistic regression analysis was conducted to identify significant predictors for the outcome variable "condition." The stepwise regression approach methodically added and sequentially removed predictors to refine the model.

Table 3.4 Model Summary

Model	Deviance	AIC	BIC	df	ΔX²	p	Nagelkerke R²
1	169.128	171.128	173.932	121			
2	161.255	165.255	170.864	120	7.872	0.005	0.083
3	156.909	162.909	171.321	119	4.346	0.037	0.127

Table 3.5 Coefficients

						Wald Test		
Model	**Parameter**	**Estimate**	**SE**	**OR**	**z**	**Wald Statistic**	**df**	**p**
1	(Intercept)	0.000	0.181	1.000	1.110×10^{-16}	1.233×10^{-32}	1	1.000
2	(Intercept)	3.436	1.293	0.032	2.658	7.063	1	0.008
	rt	-0.002	0.001	1.002	-2.682	7.194	1	0.007
3	(Intercept)	2.783	1.337	0.062	2.081	4.331	1	0.037
	rt	-0.002	0.001	1.002	-2.631	6.921	1	0.009
	ypos_flips	0.502	0.248	0.605	2.021	4.083	1	0.043

Note. Condition level "deceptive" coded as 1.

The logistic regression analysis results indicate that the model fits the data well, as suggested by the decrease in deviance and goodness-of-fit statistic predictors (Table 3.4). The third Model, which retained Reaction Time (rt) and y-flips as predictors, performed significantly better than the initial Model. The coefficient presented in Table 3.5 for reaction time was -0.002 (p = 0.009), indicating a negative relationship with the condition: for each unit increase in reaction time, the log-odds of the condition variable (deceptive) decrease by 0.002. Additionally, "ypos_flips" had a significant effect ($p = 0.0$43). The positive coefficient for "ypos_flips" indicates that higher values of "ypos_flips" increase the log odds of "deceptive." The diagnostic results regarding multicollinearity, measured using the Tolerance and Variance Inflation Factor (VIF), show that the variables in the model do not exhibit significant multicollinearity, as all values are close to 1. This suggests that the predictors included in the model are not highly correlated. The model's performance metrics are as follows: accuracy = 0.64, AUC = 0.67, sensitivity = 0.62, specificity = 0.66, and precision = 0.64.

Given these features' significance in the analyses conducted thus far, these two selected features were also used in a final classification task using a Machine Learning model. To ensure that the model's performance generalizes well to new data, the standard k-fold cross-validation technique (with k=10) was employed. This technique is beneficial when dealing with limited data samples and the need for out-of-sample accuracy. By implementing this approach, the dataset was divided into ten distinct subsets. The ML algorithm was trained on nine subsets to learn to distinguish between honest and deceptive individuals, and then it was evaluated on the remaining subset to assess its performance. This process was repeated ten times, each time using different training and validation subsets, with careful consideration to ensure that the subjects in the training and test sets differed in each fold. This method ensures that every data point corresponding to each of the 122 participants is used for training and validation, thoroughly assessing the model's performance across various subsets and reducing the risk of overfitting. It allows for a balanced and reliable evaluation of the model's generalization capabilities. After completing this procedure, various performance metrics, including accuracy, Area Under the Curve (AUC), recall, precision, and F1-score, were computed for each algorithm to gauge the overall effectiveness of the model. As displayed in Table 3.6, the results highlight the performance of the two most effective algorithms, with Quadratic Discriminant Analysis identified as the top-performing model.

Table 3.6 Performance Indices for the Top Five Algorithms.

Model	**Performance Indices**

	Accuracy	AUC	Recall	Precision	F1 Score
Light Gradient Boosting Machine	0.65	0.61	0.65	0.65	0.64
Quadratic Discriminant Analysis	0.61	0.68	0.61	0.64	0.60

Note. The table presents the values of the performance indices for the top two algorithms used in the study. The indices listed across the columns include Accuracy, Area Under the Curve (AUC), Recall, Precision, and F1 Score. The corresponding values for each algorithm are listed in the respective rows.

3.5 Discussion

The results obtained from the implemented analyses underscore distinct differences between the honest and deceptive groups. An initial exploration reveals that the variables most correlated with the experimental condition while showing minimal inter-correlations are the time participants started the first movement (initiation time) and the number of directional changes on the y-axis (ypos_flips). However, according to the independent *t*-test results, the reaction times are the only variable demonstrating a statistically significant difference between the two groups. Spatially, "ypos_flips" suggests that the deceptive group tends to exhibit more vertical directional changes than the honest group. This observation might indicate increased indecision among participants in the deceptive group. Temporally, individuals in the deceptive group initiate movement and respond significantly faster than those in the honest group, potentially reflecting impulsiveness. These results receive further support from the logistic regression analysis, which identifies reaction times and "ypos_flips" as significant predictors of deceptive behavior. The inverse relationship with reaction times implies that longer response durations are associated with a lower likelihood of deception. In contrast, the positive correlation with "ypos_flips" indicates that higher values of this variable are linked to an increased probability of deceptive behavior. The incremental enhancement in the model's fit and explanatory power with the inclusion of these predictors highlights their significance in predicting deceptive outcomes.

In totality, these findings still need to fully meet our initial expectations, which posited that deceptive participants would exhibit slower responses and trace broader trajectories than their honest counterparts. No statistically significant differences were observed among the spatial variables considered influential in differentiating the two groups according to existing literature (Mazza et al., 2020; Monaro, Gamberini, et al., 2017). This hypothesis was expected to be corroborated by an increase in directional changes (as partially observed in the "ypos_flips" variable) as well as by a more pronounced difference in variables such as the Area Under the Curve (AUC) or the Maximum Absolute Deviation (MAD). Moreover, no differences

were found except for reaction times, even when analyzing control and target sentences in detail. This challenge extended to classification efforts using machine learning techniques. Despite performance exceeding random chance, the models achieved only a marginally better accuracy than 64% in differentiating the two groups.

The results from the reaction time analysis suggest that individuals exhibiting deceptive behaviors may be more instinctive in initiating movement than their honest counterparts. The increased reactivity could be attributed to their perception of the task as a cognitively demanding challenge. Specifically, lying might have been perceived as a challenging and stimulating endeavor, thereby increasing the effort invested in the deception process and resulting in enhanced performance, ultimately reducing reaction times. Additionally, our participant selection criteria, which included individuals with any history of drug use in the workplace, might have inadvertently diminished the difficulty typically associated with deception. Including participants with a history of drug use, even if limited to a single instance, may have simplified the complexity of lying. This disparity between anticipated predictions and actual outcomes highlights the complexities inherent in this domain and opens up new avenues for future research. Such inquiries are essential for a deeper understanding of these behaviors and their possible applications in practical scenarios.

3.6 Conclusion

This study aimed to examine the impact of deception about drug use in the workplace on mouse movements during an online task involving double-choice questions. The findings revealed differences between groups, demonstrating that participants in the deceptive condition took significantly less time to respond than the honest group and made more vertical axis direction changes. A possible explanation for this only hinted at and unexpected result could be due to the low difficulty associated with deception. Specifically, the participants of the study, all of whom had a history of drug use in the workplace—even in instances limited to a single occurrence—may have found the task of lying to be considerably effortless.

To address this issue, a second experiment is proposed. In this subsequent study, we aim to refine our selection criteria for the deceptive group, focusing on individuals who have engaged more frequently in drug use at work. This approach is expected to intensify the cognitive load associated with lying, potentially leading to more pronounced variations in spatial trajectory. Additionally, we plan to enhance the cognitive load by introducing increased temporal pressure in the experimental design. This adjustment is intended to highlight further disparities associated with deception.

Experiment 2

3.7 Materials and Methods

The study was preregistered on Open Science Framework (https://osf.io/vuc7r/).

A between-subjects design was utilized for this study, which was conducted entirely online and consisted of a pre-screening questionnaire and a mouse tracker task. In light of the limitations identified in the first study, several modifications have been made to the experimental procedure, which will be detailed subsequently.

3.7.1 Experimental Procedure

Pre-screening Questionnaire. As for the first study, the initial phase of the research encompassed a pre-screening questionnaire to select appropriate participants for the mouse tracker test. Specifically, the objective was to recruit two distinct groups: individuals who had frequently worked under the influence of drugs and those who had not. A brief pre-screening questionnaire using Prolific's Survey Builder (prolific.co) was administered to reach a wide participant pool. Data was collected from 2990 participants, who were asked to respond to the following five questions:

1. How frequently do you use drugs during working hours?
2. How frequently are you under the influence of drugs during working hours?
3. How frequently do you violate safety regulations at work?
4. How frequently do you forget to follow safety regulations at work?
5. How frequently do you have an accident at work due to a violation of safety regulations?

Participants were asked to answer the questions by referring to their current work experience (approximately within the past 12 months). Answers provided to participants for the first 4 questions were: *"Every day," "3-4 times a week", "1-2 times a week", "2-3 times a month", "Once a month," "Less than once a month"* and *"Never."* For the last questions instead: *"At least once a month," "Almost every month," "At least once every 3 months", "2-3 times a year", "Once a year," and "Never."* Those participants who responded to questions 1 and 2 selecting *"Every day," "3-4 times a week,"* or *"1-2 times a week"* will be included in the group of liars. At the same time, those who never use drugs at work ("*Never"* answer for questions 1 and 2) will be part of the group of honest participants. The same procedure

employed in Study 1 was used to recruit appropriate participants for the second part of the study, which involved a mouse movement task.

Mouse Tracker Task. In this experiment, the task was constructed using the Lab.js platform (lab.js.org) and executed on a self-hosted JATOS server (jatos.org). Upon reading the general instructions and providing consent to participate, participants were given ten practice sentences to familiarize themselves with the task. The experimental procedure is identical to that used in Study 1, as are the practice sentences. Please refer to the procedure outlined in the "Mouse Tracker Task" Section of the first experiment and to the practice sentences in Table 3.1 for further details.

A significant difference introduced in this study is the time pressure deliberately imposed on participants. Specifically, unlike the first study, where feedback from an incorrect response lasted for 2000 milliseconds and error feedback lasted for 3000 milliseconds, in this second study, if a participant made an error or was too slow to respond (more than 4000 milliseconds), the feedback message would remain displayed for 5000 milliseconds. Additionally, accompanying the message for an error or excessive slowness was the text: *"You have to wait for 5 seconds before proceeding to the next sentence."* This modification was implemented to encourage participants to respond quickly without making errors. Additionally, to enhance the participants' motivation to achieve optimal performance in honest and deceptive conditions, an additional bonus of £10 was offered to the participant who was quickest to respond with the fewest errors—one from the honest group and one from the deceptive group. This incentive was clearly stated in the instructions. Inverse efficiency scores were used to identify the two participants deserving of the bonus. For each participant, this score was calculated by dividing the average reaction time (RT) by the proportion of correct responses.

Following the practice task, as in the first study, participants proceeded to the main task, which involved imagining themselves in an online job interview scenario and responding to sentences related to drug use at work. The deceptive group received instructions to simulate never having used drugs, while the honest group answered truthfully because they had indeed never used drugs at work. The format of sentence presentation remained consistent with the first experiment, except for the content of the sentences. Unlike the first study, where participants had used drugs at least once in their lifetime, participants in this second study habitually used drugs at work. For this reason, the content of the target questions was modified. Specifically, while the target sentences in the first study referred to past or one-time drug use at work (e.g., "*I have used drugs at work"*), in this study, the sentences refer to current usage (e.g., *"I use drugs at work"*). The complete list of target sentences used is provided in Table 3.7.

Table 3.7 Trial sentences.

False Answer	True Answer
I use drugs at work.	I never use drugs at work.
While working, I use drugs.	While working, I never use drugs.
Sometimes, I consume drugs at work.	Consuming drugs at work never happens to me.
During working hours, I sometimes use drugs.	During working hours, I never use drugs.
It happens that I use drugs at work.	It never happens that I use drugs at work.
I work under the influence of drugs.	I never work under the influence of drugs.
While working, I am under the influence of drugs.	While working, I'm never under the influence of drugs.
Sometimes, I do my job under the effect of drugs.	I never do my job under the effect of drugs.
Sometimes, I am intoxicated during working hours.	During working hours, I am never intoxicated.
It happens that I work under the effect of drugs.	It never happens that I work under the effect of drugs.

Note. Twenty statements related to drug use at work (10 requiring a TRUE response, while the other a FALSE response. The first five sentences ask participants if they use drugs while at work, while the other five ask if they work under the effect of drugs.

Also, in this study, sentences of similar length were carefully selected to ensure that the length of sentences did not introduce systematic noise into the data. For each short sentence in the target group, at least one sentence of the same length was included in the control group. Furthermore, a *t*-test confirmed that the difference in the number of letters between each set of statements was not statistically significant.

The study then concluded in the same manner as Study 1, with the sole difference being that participants who lied were also asked an additional question about whether they had difficulty lying. We used this feedback to improve the task in future experiments and assess the difficulty level of deception. Following this, participants were automatically redirected to the Prolific platform.

3.7.2 Hypotheses

Many different parameters can be derived based on mouse-tracking data. Despite the explorative nature of the study, in comparison to the control group, we expect that the deceptive group:

H1) will exhibit trajectories that deviate more from the ideal trajectory (i.e., the straight line connecting the start and end points);

H2) will exhibit lower values of temporal variables such as acceleration or speed compared to truth-tellers, demonstrating that it takes them longer to respond to sentences;

H3) will commit more errors;

H4) will demonstrate a general difficulty in responding to control questions due to the continuous shift between lying (target questions) and being honest (control questions).

3.7.3 Collected Measures

The same variables reported in Study 1 were calculated in this second experiment. Please refer to the "Collected Measures" Section of Study 1 for further details.

3.7.4 Exclusion Criteria

The same exclusion criteria were adopted in this second study. However, due to changes in the pre-screening questionnaire, participants who answered *"Less than once a month," "Once a month,"* or *"2-3 times a month"* to questions 1 or 2 were excluded from this second study.

3.7.5 Participants

The sample size for the study consisted of 2990 participants for the pre-screening survey and 122 participants for the mouse tracker task. Please refer to the "Participants" Section in Study 1 for power calculations and sample size determination.

Recruitment of participants was conducted through the online platform Prolific, and all assessments were administered in English. Participation was limited to Desktop PCs or Laptops, with the use of a trackpad not permitted. The eligibility criteria for the pre-screening questionnaire were as follows: First Language: ENGLISH; Primary Language: ENGLISH; Minimum age: 18; Maximum age: 65; Country of residence: UK or USA; Employment Status: Full-time; Safety procedures at work screening criteria provided by Prolific*: "Occasionally my job task requires the use of a safety procedure"* or *"Sometimes my job involves compliance with*

a safety procedure" or *"Almost all my job tasks are guided by safety procedures."* The study sample comprised 122 participants (47 female and 1 non-binary participant). All participants were over 18 (M = 40.22, SD = 10.59). Allocation to the honest or liar group was determined based on the responses provided in the Prolific Survey Builder. 61 participants were assigned to the honest group, while another 61 were allocated to the deceptive group. No statistically significant differences were observed between groups concerning age and gender.

The study's duration was approximately 1 minute for the pre-screening questionnaire and roughly 7 minutes for the mouse tracker task. Participants received compensation of £9.00 per hour for completing the pre-screening questionnaire and the mouse tracker task. Eligible participants who had completed the pre-screening questionnaire were contacted again through Prolific, using their Prolific ID, to participate in the mouse tracker task.

3.7.6 Data Pre-Processing and Features Computation

The same data pre-processing and feature computation procedure outlined in the first study was employed for this second experiment.

3.8 Data Analysis

3.8.1 Visual Comparison of Trajectories

We conducted a visual comparative analysis of the mean mouse trajectories for the honest and deceptive groups, focusing on their movement patterns in response to control and target questions. Figure 3.3 comprises multiple panels illustrating these trajectories: Panel A depicts the average paths for both groups in response to control questions, while Panel B presents those corresponding to target questions. Our observations reveal that, in both scenarios, the deceptive group tends to follow trajectories that diverge more substantially from the ideal path, conceptualized as the straight line from the starting point (0,0) to the response. Subsequently, we computed the average trajectories for the control and target questions, segmented according to the group. Panel C illustrates the average trajectories for the honest group, whereas Panel D exhibits those for the deceptive group. A notable finding is the increased deviation from the ideal trajectory in response to target questions compared to control questions, which is observable in both groups. This deviation is more pronounced in the trajectories associated with the deceptive group.

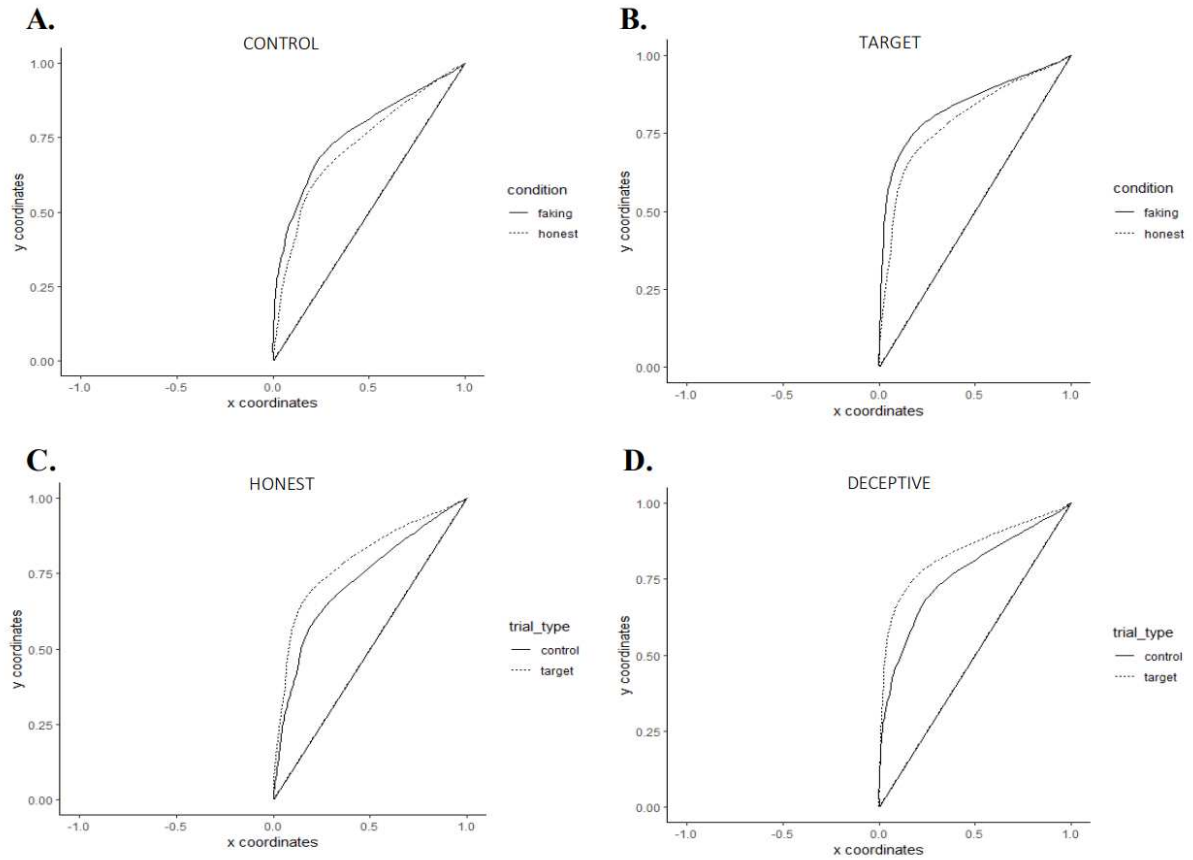


Figure 3.3 Trajectories for control [Panel A], target [Panel B], honest [Panel C], and deceptive group [Panel D].

3.8.2 Features Selection

A feature selection process was implemented to explore those independent variables that exhibited the most significant correlation with the dependent variable (condition: honest or deceptive) while simultaneously ensuring minimal inter-correlation among these variables. Initially, the point-biserial correlation coefficient method was applied to evaluate the relationships between the variables and the condition. The analysis revealed several key variables: error (r_{pb} = 0.30), MAD (r_{pb} = 0.16), AD (r_{pb} = 0.16), MD_above (r_{pb} = 0.15), and AUC (r_{pb} = 0.16). These variables demonstrated the most substantial correlations with the condition variable. However, it is essential to emphasize that certain variables exhibit extremely high intercorrelations. Given the significant correlation between MAD and MAD_above variables (r = 0.996, p < 0.001), it was decided to retain only MAD and exclude MAD_above. A similar approach was adopted for the variables idle_time and hover_time (r = 1.000, p < 0.001), which measure the duration of inactivity or lack of movement during the entire trial.

Subsequently, the Correlation-based Feature Selector (CFS) was employed to evaluate the attributes. This method assesses each feature's merit by considering its predictive capacity

and the extent of redundancy between features. The Greedy Stepwise algorithm, known for its effectiveness in feature selection, was utilized as the search strategy. It sequentially selects features, starting with the most strongly correlated with the model's output. From the independent variables analyzed, the error was ultimately selected.

3.8.3 Statistical Analysis

The selected measurements were examined to gain a deeper understanding and explore the disparities between the two conditions. After verifying the normal distribution of the data, a Mann-Whitney test was employed to elucidate the disparities between the two independent groups. This non-parametric approach was deemed appropriate for comparing the central tendencies of these groups in the context of non-normally distributed data. Results reported in Table 3.8 suggest that average trajectories in deceptive condition tend to deviate more from the direct path, cover a greater total distance, and are associated with a more significant error rate than those in the honest condition.

Table 3.8 Descriptive and Statistical Analysis of the honest and faked conditions.

Variables	Descriptives			
	Mean	**SD**	**W**	**Rank-Biserial Correlation**
MAD_H	0.465	0.245	8764*	0.178
MAD_F	0.546	0.260		
AD_H	0.172	0.098	8652 *	0.163
AD_F	0.203	0.106		
AUC_H	0.280	0.133	8705*	0.170
AUC_F	0.320	0.137		
total_dist_H	2.224	0.507	8554*	0.149
total_dist_F	2.367	0.568		
error_H	0.287	0.787	8883.5 ***	0.194
error_F	0.885	1.890		
xpos_min_H	0.257	0.160	8753 ***	0.176
xpos_min_F	0.313	0.181		

Note. Mean, Standard Deviation (SD), the value of the W-statistic (W), *p*-value, and Effect Size of the significant variables in honest (H) and faked (F) groups. Each variable is represented in two rows. For example, the first two rows of the table should be read as follows: MAD_H defines the Maximum Absolute Deviation in the honest (H) condition, while F is in the faked condition. Effect size is given by the matched rank biserial correlation.
* $p \leq 0.05$ ** $p \leq 0.01$ *** $p \leq 0.001$

Subsequently, we observed the comparison between the two experimental groups and the question type: control and target. Considering the nature of the variables in our study, we adopted a tailored statistical approach. For those variables, we conducted a distribution analysis and a test of residual normality to confirm data normality distribution.

We employed a Mixed Model to analyze variables with a normal distribution. We opted for the Generalized Linear Mixed Model (GLMM) with the Gamma distribution with the "Log" link function for the continuous variables with a non-normal distribution. For the count distribution variables, such as errors, we used a Poisson GLMM. We considered the condition (honesty vs. deception) and sentence type (control vs. target) as fixed factors in all models. Participants were entered as random effects. All analyses were performed using Jamovi (Version 2.4) [Computer Software] (2023) and are reported below, divided by variables.

Maximum Absolute Deviation. The results of the fixed effect omnibus tests obtained running a Mixed Model indicated that the sentence type factor had a highly significant effect, F (1, 120) = 94.983, $p < 0.001$. At the same time, the condition did not show significance F (1, 120) = 3.706, $p = 0.057$. Also, the interaction between them, F (1, 120) = 0.155, $p = 0.694$, did not reach statistical significance. The parameter estimates for the fixed effects revealed that sentence type had a substantial positive impact on the dependent variable, with "target" sentences yielding higher values than "control" sentences (Estimate = 0.1358, SE = 0.0139, $p < 0.001$). The Bonferroni-corrected post hoc tests indicated significant differences in the deceptive condition between target and control sentences (Difference = -0.1303, SE = 0.0197, t = -6.61, $p_{bonferroni} < 0.001$), with control sentences yielding lower values. The same trend was also significant in comparing control and target sentences within the honest group (Difference = -0.1413, SE = 0.0197, t = -7.71, $p_{bonferroni} < 0.001$).

Average Deviation. Mixed Model analysis showed a significant main effect for trial type F (1, 120) = 89.4350, $p < 0.001$, while condition F (1, 120) = 3.4014, $p = 0.068$, and their interaction, F (1, 120) = 0.0735, $p = 0.787$, did not reach statistical significance. The parameter estimates for the fixed effects revealed that sentence type had a substantial positive impact on the dependent variable, with "target" sentences yielding higher values than "control" sentences (Estimate = 0.05334, SE = 0.00564, $p < 0.001$). The Bonferroni-corrected post hoc tests indicated significant differences in the deceptive condition between target and control sentences (Difference = -0.0549, SE = 0.00798, t = -6.88, $p_{bonferroni} < 0.001$), with control sentences yielding lower values. The same trend also significantly compared control and target sentences within the honest condition (Difference = -0.0518, SE = 0. 00798, t = -6.50, $p_{bonferroni} < 0.001$).

Area Under the Curve. Mixed Model analysis showed a significant main effect for trial type F (1, 120) = 82.5773, $p < 0.001$, while condition F (1, 120) = 3.3261, $p = 0.071$, and their interaction, F (1, 120) = 0.0348, $p = 0.852$, did not reach statistical significance. The parameter estimates for the fixed effects revealed that sentence type had a substantial positive impact on the dependent variable, with "target" sentences yielding higher values than "control" sentences (Estimate = 0.07650, SE = 0.00842, $p < 0.001$). The Bonferroni-corrected post hoc tests indicated significant differences in the deceptive condition between target and control sentences (Difference = -0.0781, SE = 0.00119, t = -6.56, $p_{bonferroni} < 0.001$), with control sentences yielding lower values. The same trend also significantly compared control and target sentences within the honest condition (Difference = -0.0749, SE = 0. 00119, t = -6.50, $p_{bonferroni} < 0.001$).

Total distance: The distance traveled by trajectories comparison was made running a GLMM with a Gamma distribution and a "Log" link function to account for the positive skewness and non-normal distribution. The fixed effects analysis demonstrated significant main effects for trial type (χ^2=137.36, df = 1, $p < 0.001$), while condition (χ^2 =1.22, df = 1, $p = 0.269$) and their interaction (χ^2 =1.18, df = 1, $p = 0.278$) were not significant. Parameter estimates for fixed effects indicated that the target sentences (B = 0.1197, SE = 0.0102, exp(B) = 1.127, z = 11.72, $p < 0.001$) affected increasing the distance traveled by the trajectories. Post hoc tests with Bonferroni correction highlight these specific significant comparisons within the honest (Difference = 0.877, SE = 0.0127, z = -9.050, $p < 0.001$) and deceptive (Difference = 0.897, SE = 0.0130, z = -7.524, $p < 0.001$) group between control and target sentences.

xpos_min: The minimum -x position comparison was made running a GLMM with a Gamma distribution and a "Log" link function to account for the positive skewness and non-normal distribution. The fixed effects analysis demonstrated significant main effects for trial type (χ^2 = 35.061, df = 1, $p < 0.001$), while condition (χ^2 = 1.809, df = 1, $p = 0.179$) and their interaction were not significant (χ^2 = 0.530, df = 1, $p = 0.467$). Parameter estimates for fixed effects indicated that the target sentences (B = 0.424, SE = 0.0716, exp(B) = 1.528, z = 5.921, $p < 0.001$) affected an increase in the lowest coordinate value achieved for the target sentences. Post hoc tests with Bonferroni correction highlight these specific significant comparisons within the honest (Difference = 0.622, SE = 0.0627, z= -4.710, $p < 0.001$) and deceptive group (Difference = 0.689, SE = 0.0696, z = -3.682, $p = 0.001$) between control and target sentences.

Errors: A GLMM with a Poisson distribution was applied to analyze the count of errors as a function of condition, type of question, and their interaction, with random effects accounted for by individual differences (ID). The fixed effect omnibus tests didn't show a significant main effect for fixed effects and their interaction. However, post hoc comparisons with Bonferroni

correction revealed several significant differences in the within-subject error rate comparisons in the deceptive group (exp(B)= 0.0485, SE= 0.0229, z = -6.40684, $p_{bonferroni} < 0.001$) between the control and target questions. In particular, the errors for the target sentences were notably higher. The same discrepancy is not observed within the honest group. Moreover, the deceptive and honest groups were found to be statistically different in the number of errors for target questions (exp(B)= 3.0634, SE= 1.0268, z = 3.34005, $p_{bonferroni} = 0.005$), with the deceptive group making more errors.

3.8.4 Classification Task

This section presents the logistic regression analysis results to assess the relationship between the predictor variables and the likelihood of deceptive behavior. The model summary (see Table 3.9) reveals a progressive improvement in model fit by adding variables across models 1 to 9. The predictive power of the models, as indicated by Nagelkerke R^2, consistently increased with each subsequent model, peaking at model 9 (Nagelkerke $R^2 = 0.343$). Table 3.10 delineates the coefficients for the finalized model (Model 9). Notable predictors encompass error_target, xpos_min_control, ypos_reversals_control, ypos_max_control, and ypos_flips_target. An escalation in the frequency of errors within target sentences is linked to an increased likelihood of the "deceptive" condition, as evidenced by an odds ratio of 1.484. This indicates that a higher incidence of errors correlates with a greater probability of the "deceptive" condition. Lower minimum values on the x-axis in control sentences (xpos_min_control) are strongly associated with a reduced probability of "deception." Low values are meant to be close to zero on the x-axis, hence closer to the ideal trajectory. In other words, the closer the minimum values on the x-axis are to zero and thus to the ideal trajectory, the lower the likelihood that they belong to the deceptive condition. Furthermore, an increase in the number of y-axis crossings in control sentences (ypos_reversals_control) significantly elevates the likelihood of "deception," with an odds ratio of 11.460. This implies that frequent crossings on the y-axis in control sentences are associated with the "deceptive" condition. However, "ypos_flips_target" appears negatively associated with the likelihood of the "deceptive" condition. As the number of directional changes along the y-axis in target sentences increases, the probability of the condition being "deceptive" decreases. Finally, elevated maximum values on the y-axis in control sentences are correlated with a decrease in the probability of "faking." Multicollinearity diagnostics indicated acceptable levels of multicollinearity for all predictors (VIF < 2).

Table 3.9 Model Summary

Model	Deviance	AIC	BIC	df	ΔX^2	p	Nagelkerke R^2
1	169.128	171.128	173.932	121			
2	157.108	161.108	166.716	120	12.020	< 0.001	0.125
3	151.894	157.894	166.306	119	5.214	0.022	0.176
4	147.342	155.342	166.558	118	4.552	0.033	0.218
5	142.899	152.899	166.919	117	4.443	0.035	0.258
6	140.675	152.675	169.499	116	2.224	0.136	0.277
7	138.637	152.637	172.265	115	2.039	0.153	0.295
8	135.119	151.119	173.551	114	3.517	0.061	0.324
9	132.817	150.817	176.053	113	2.302	0.129	0.343

Table 3.10 Coefficients

					Wald Test		
Parameters	**Estimate**	**Standard Error**	**Odds Ratio**	**z**	**Wald Statistic**	**df**	**p**
(Intercept)	16.231	18.305	$1.120\times10^{+7}$	0.887	0.786	1	0.375
error_control	16.507	1006.439	$1.475\times10^{+7}$	0.016	2.690×10^{-4}	1	0.987
error_target	0.395	0.167	1.484	2.364	5.588	1	0.018
xpos_min_control	-4.472	1.643	0.011	-2.723	7.413	1	0.006
ypos_reversals_control	2.439	1.074	11.460	2.272	5.160	1	0.023
acc_min_target	647.553	544.819	$1.694\times10^{+281}$	1.189	1.413	1	0.235
ypos_max_control	-44.641	19.466	4.098×10^{-20}	-2.293	5.259	1	0.022
ypos_flips_target	-0.638	0.308	0.528	-2.072	4.291	1	0.038
ypos_max_target	27.906	18.982	$1.317\times10^{+12}$	1.470	2.161	1	0.142

Note. Condition level "deceptive" coded as class 1.

The model correctly predicts 68.85% of deceptive cases and 75.41% of honest cases, resulting in an overall correct classification rate of 72.13%. The performance metrics further support the model's effectiveness, with an accuracy of 72.1%, an AUC of 0.79, a sensitivity of 0.75, a specificity of 0.69, and a precision of 0.71. In summary, the logistic regression analysis indicates that several predictor variables are significantly associated with deceptive behavior in the given condition. The model performs well in classifying deceptive and honest responses, suggesting its potential utility in identifying dishonest behavior.

A Machine Learning model also used these selected features in a final classification task. To ensure that the model's performance generalizes well to new data, the standard k-fold cross-validation technique (with k = 10) was employed. As displayed in Table 3.11, the results highlight the performance of the two most effective algorithms, with the Light Gradient Boosting Machine identified as the top-performing model.

Table 3.11 Performance Indices for the Top Five Algorithms.

Model	Performance Indices				
	Accuracy	**AUC**	**Recall**	**Precision**	**F1 Score**
Light Gradient Boosting Machine	0.65	0.61	0.65	0.65	0.64
Quadratic Discriminant Analysis	0.61	0.68	0.61	0.64	0.60

Note. The table presents the values of the performance indices for the top two algorithms used in the study. The indices listed across the columns include Accuracy, Area Under the Curve (AUC), Recall, Precision, and F1 Score. The corresponding values for each algorithm are listed in the respective rows.

3.9 Discussion

The current research comprises two studies investigating deception in workplace drug use through mouse movement analysis. In both experiments, participants were categorized as honest or deceptive based on their history of drug use. Each group was tasked with responding to a series of true/false sentences concerning their drug usage in the workplace. Surprisingly, in the first experiment, deceptive participants responded more rapidly than their honest counterparts, yet they exhibited more directional changes, indicating a degree of indecision in their choices. The second experiment implemented changes, including heightened time pressure and improved participant engagement, to gain a deeper understanding. Unlike the first study, which involved participants who had used drugs in their workplace at least once (which may have diminished their engagement and the cognitive effort needed for deceit), the second study exclusively focused on habitual drug users at work (those who used drugs at least once or twice a week) for the deceptive group. The aim was to more clearly distinguish between mouse movement patterns in honest and deceptive responses, thereby providing insight into the unexpected results of the initial experiment.

Similar to the first study and according to our Preregistration, the analyses conducted in the second study encompassed visual comparisons, feature selection, statistical analyses, and classification methods. Compared to the conclusions of the initial research, this second experiment has undoubtedly yielded intriguing results. The visual analysis of mouse trajectories unveiled distinct patterns between honest and deceptive responses, with the deceptive group exhibiting a more significant deviation from the ideal path. This pattern was consistent in both control and target scenarios, suggesting a systematic difference in the strategies of individuals when being deceptive. A notable shift in results emerged between the two experiments due to

the altered inclusion criteria for participant selection. Specifically, a visual comparison of the average trajectories from both experiments (Figures 3.2 and 3.3) shows a more pronounced divergence between honest and deceptive responses in the second experiment, especially in response to target questions. Statistical analyses also corroborated this more significant discrepancy. While initial differences in the primary study were confined to reaction times and on the y-axis, the second study identified various spatial variables and error rate as helpful in differentiating between the groups. The analyses have indeed demonstrated that the deceptive group committed more errors compared to the honest group, thus enabling effective discrimination between the two groups, particularly when examining the target sentences. Moreover, the investigations within the deceptive group have revealed that they commit significantly more errors in responding to the target sentences than to the control ones, thereby confirming the general difficulty associated with lying. This discrepancy was not observed in the honest group. Furthermore, the conducted statistical analysis revealed that the average trajectories in the deceptive condition tend to deviate more from the direct path. This deviation is indicated by significant differences between the two groups in variables such as Area Under the Curve (AUC), Maximum Absolute Deviation, Average Deviation from the direct path, and the minimum value reached on the x-axis, which is significantly larger in the group of liars. Additionally, the deceptive group was shown to cover a greater total distance than those in the honest condition, once again indicating that they trace broader trajectories compared to the control group.

Logistic regression analysis indicates that specific performance characteristics in control and target sentences can be essential deception indicators, with spatial changes (measured by maximum and minimum values and directional shifts on the y-axis) and error frequency playing a pivotal role. Moreover, the accuracy and other performance parameters of logistic regression improved from the first experiment (64% accuracy) to the second (72% accuracy), though they still fall short of high accuracy levels (approximately 80%). Despite training various machine learning models, the classification task did not meet our expectations, remaining stable at 64% accuracy. However, this outcome aligns with the findings reported in the existing literature. Monaro utilized mouse tracking as a technique for fake identity deception and achieved similar accuracy levels in classifying the two groups via machine learning for control (50-62.5%) and target (65-67.5%) questions while reaching around 90% accuracy using only unexpected questions (Monaro, Gamberini, et al., 2017).

Although the modifications implemented in this exploratory study have yielded intriguing findings, not all our hypotheses were fully confirmed in this second experiment.

Specifically, the group of liars did not exhibit longer durations in temporal variables, such as reaction times, suggesting that they do not take more time to respond to sentences. The variables analyzed also demonstrated a general difficulty perceived by our participants in responding to target sentences compared to control ones. However, with the exception of the number of errors, this difference emerged in both the group of liars and the honest group, thus precluding the attribution of this effort solely to the act of lying. Finally, regarding the analyses conducted on the classification task, although the trajectories of the liars were shown to deviate more from the ideal trajectory as evidenced by statistical analysis, the data obtained from the investigated variables do not seem to be sufficiently predictive of group membership.

While the results facilitated better discrimination between the two groups compared to the previous study, they appear to be still somewhat hid. Indeed, certain limits may potentially diminish the impact of deception on the investigated variables. These limitations primarily include the artificial setting of the experiments, which may only partially replicate real-world workplace deception conditions. Notably, while advantageous, the ambition to conduct the research entirely online raises several considerations, such as the need for more control over the conditions under which participants responded to questions. Despite various post-hoc measures, such as excluding participants who took excessively long or short times to respond or those suspected of not using a mouse, it was impossible to control the response conditions, which is critical when using a sensitive tool like a mouse. Additionally, the online recruitment of participants, though improved from the first to the second study, warrants clarification. Despite verifying participants' statements and asking them before and after the practical mouse experiment whether they had truthfully declared their workplace drug use, it is impossible to ascertain the veracity of self-reports. Some participants might have lied during the pre-screening questionnaire, thus being placed in the wrong group. Lastly, the low risk for participants who face no consequences for lying in the study must be mentioned. Various researchers recognize the importance of specific "moderators" that can emphasize deception signals (O'Sullivan et al., 2009; Porter & ten Brinke, 2010). These signals of deception become more or less diagnostic depending on the situation. Although drug use in a workplace can lead to significant consequences, in an experimental environment such as this, participants may feel secure knowing they face no risks. This perceived safety could create differences in the collected measurements and behavior under real workplace scrutiny.

Despite these considerations, the results from these experiments provide valuable insights into the subtleties of deceptive behaviors and the potential of mouse movement analysis as a tool for detecting deception in real-world settings.

3.10 Conclusion

This explorative study was born in a complex and multifaceted workplace environment, aiming to identify deception about drug use in this setting. The research topic is essential, as it directly affects individual health, workplace safety, and economic efficiency.

Recognizing the profound and destructive consequences of workplace drug use, several countries have implemented a range of policies, from stringent legal frameworks to more permissive approaches. Among these, a key focus is the current debate about the biological practice of drug testing, which raises critical discussions about the intricate balance between ensuring workplace safety and upholding employees' privacy rights. This issue underscores the ethical, moral, and legal challenges of implementing workplace drug testing protocols.

However, although alternative testing methods are considered less invasive by nature, such as self-reporting via paper-and-pencil questionnaires, these approaches are not without limitations, as their transparency regarding detecting drug use can facilitate deceit. This vulnerability underscores the need for more effective and reliable testing methods to accurately assess drug use while managing the perceived invasiveness of some testing procedures. The study's innovative approach in employing kinematic analysis of mouse movements as an implicit measure to detect deception related to drug use in the workplace contributes to this ongoing dialogue and marks a significant advancement in this field, offering new insights to address the challenges posed by workplace drug use.

Furthermore, this approach does not require particular technical skills for administration and data collection. On the contrary, biological testing techniques, such as blood draws and urine analysis, require suitably equipped laboratories and highly specialized technical personnel. This also demonstrates the potential of non-invasive, cost-effective methods in addressing complex issues like deception. Last but not least, the application of this approach has significant implications across a multitude of sectors. Its utilization applies in numerous contexts, both within and beyond the workplace sphere. The inherent simplicity characterizing this procedure allows for its straightforward implementation in any context deemed appropriate to detect deception.

This study paves the way for further research and innovation in the field, opening new horizons for understanding and mitigating the complexities associated with workplace drug use. Integrating technological advancements with ethical considerations sets a precedent for future endeavors to enhance individual well-being and organizational integrity.

Chapter 4.
Uncovering Lies During an Interrogation

4.1 Abstract

This study proposes an innovative approach to detect deception during face-to-face investigative interviews by integrating response latency analysis with the unexpected question technique. Sixty participants were assigned to an honest or deceptive group. The deceptive group was instructed to memorize false biographical details linked to a fictitious identity. Throughout the interviews, participants were presented with a randomized sequence of control, expected, and unexpected open-ended questions about identity. Responses were audio-recorded for detailed examination. Our findings indicate that while response latencies and error rates did not significantly differ between groups for control questions (which necessitated truthful answers), deceptive participants showed markedly longer latencies and higher error rates for both expected (requiring deception) and unexpected questions (where premeditated deception was not possible). Moreover, through a within-subject analysis, the unexpected queries significantly impaired the performance of individuals when compared to control and expected questions. Leveraging machine learning algorithms, our approach attained a classification accuracy of 98% in distinguishing deceptive and honest participants. Additionally, a classification analysis on single response levels was conducted. The findings underscore the effectiveness of merging response latency metrics with unexpected questioning as a robust method for identity deception detection in face-to-face investigative interviews. Significant implications for enhancing interrogation strategies are discussed.

4.2 Introduction

Forensic psychology has developed various methodologies to distinguish between truth and deception under controlled experimental conditions. Notable examples include the Concealed Information Test (CIT; Verschuere et al., 2011) and the autobiographical Implicit Association Test (aIAT; (Sartori et al., 2008). Despite these advancements, there is a noticeable gap in the development of methods for assessing the credibility of statements during investigative interviews. Ideally, these methods should be integrated smoothly into the flow of an investigative interview and remain imperceptible to the subject under examination, thereby avoiding any explicit indication of credibility assessment (Vrij & Granhag, 2012). Current techniques such as CIT, aIAT, and traditional polygraph testing tend to make the subjects overtly aware of the evaluation process, potentially affecting their responses. An attempt to address these limitations has been recently made by combining these behavioral techniques with eye tracking (Ogawa et al., 2021; Zangrossi et al., 2024), given the relation between eye movements and endogenous brain dynamics (Celli et al., 2022; Zangrossi et al., 2021). However, using aIAT or similar tools is not always applicable within police procedures or legal systems. This study introduces a novel methodology that relies on the analyses of audio recordings of investigative interviews to accurately distinguish between deceptive and honest responses.

4.2.1 Clues of Lies

A multitude of studies suggests that deception is cognitively more demanding than truth-telling, given the additional cognitive processes engaged during lying (Christ et al., 2009; Duran et al., 2010; Gombos, 2006; Hadar et al., 2012; Langleben et al., 2002; Luan Phan et al., 2005; Spence et al., 2001). One metric extensively used to probe these cognitive demands is reaction times (RTs). For instance, while truthful responses generally occur within 400 milliseconds (Walczyk et al., 2003, 2009), deceptive responses are typically associated with longer RTs (Johnson et al., 2004; Spence et al., 2001; Vendemia et al., 2005; Walczyk et al., 2003). This difference in RTs is thought to reflect cognitive load (De Boeck & Jeon, 2019), hinting that longer response durations could be pivotal in distinguishing deception from honesty.

Recognizing the potential of this phenomenon, there has been a significant shift in research focus toward enhancing deception detection by highlighting signs of cognitive strain (Blandón-Gitlin et al., 2014; Masip & Herrero, 2015; Parkhouse & Ormerod, 2018; Vrij et al., 2006, 2012; Walczyk et al., 2013). This is achieved by amplifying the cognitive load on

respondents using several innovative techniques. For instance, respondents might be asked to narrate events in reverse order (Vrij et al., 2008, 2012), maintain uninterrupted eye contact during interviews (Vrij, Mann, et al., 2010), or impose an additional task. This exploits the dual-task paradigm, where juggling two simultaneous tasks can be particularly taxing (Gawrylowicz et al., 2016; Goto & Hakoda, 2020; Lancaster et al., 2013; Pashler, 1994).

Task switching is a crucial high-level cognitive ability that enables individuals to direct and allocate their attentional resources across multiple sequential cognitive tasks (Draheim et al., 2016). In typical deception paradigms, task switching occurs even when transitioning from truth-telling to lying or vice versa, such as when the control and target questions are presented randomly. The individual engaged in deceit must switch between these two cognitive tasks, thereby influencing task performance (Foerster et al., 2017). Within the task-switching literature, the well-established presence of a performance switch cost is evidenced by participants being slower and more error-prone when switching tasks than when repeating the same task (Kiesel et al., 2010; Monsell, 2003; Vandierendonck et al., 2010). Therefore, exploring switch costs in task switching presents a potential, albeit underexplored, avenue for lie detection (Foerster et al., 2017).

It is worth noting, however, that the cognitive demands of lying are not uniform. Whether challenging to fabricate or relatively straightforward, the lie's complexity plays a role in this variability (Vrij & Heaven, 1999). Additionally, individuals vary in their proficiency at lying (Vrij, Granhag, et al., 2010), with many liars rehearsing their deception in advance (Clemens et al., 2011). In the deception detection landscape, a clear strategy divergence is naturally observed between truth-tellers and liars. Truth-tellers rely on genuine memories and seldom employ overt strategies (Gilovich et al., 1998; Hartwig et al., 2007), whereas liars tend to script their narratives in advance (Clemens et al., 2011). This training allows them to rely on memory during questioning, thus reducing cognitive overhead. Consequently, they often stick to core story elements and avoid incorporating unconventional details (DePaulo et al., 2003). Given a choice, liars tend to use actual events over fabrications to reduce cognitive effort (D. A. Leins et al., 2013). This strategy helps them maintain narrative coherence and avoids arousing suspicion. Intriguingly, premeditation significantly alters the cues indicative of deceit and associated latencies. Well-crafted lies often show reduced RTs compared to spontaneous deceptions (DePaulo et al., 2003; Walczyk et al., 2009, 2012; Zuckerman et al., 1981). This intricacy in deception presents significant challenges for detectors, especially when confronted with carefully crafted lies (Littlepage & Pineault, 1985).

However, the strategy of reducing cognitive load by preparing for an interview—that is anticipating potential questions and formulating answers in advance—is effective only if the liar accurately predicts the questions that will be asked. In other words, this strategy falters when the interviewer poses questions the liar has not anticipated (Vrij, 2016). In this perspective, the "unexpected question technique" can be fruitfully implemented into the investigative interview to enhance lie detection accuracy. When faced with unanticipated questions, truth-tellers respond instinctively, while deceivers struggle to fabricate answers without rehearsed responses (Vrij et al., 2009). Such questions can span a range of topics, from planning future behavioral intentions, such as *"I want you to think back to when you planned your errand. I want you to tell me about your planning, and I want you to be as detailed as possible?"*(Sooniste et al., 2015, 2016) or spatial-temporal details and shifts, e.g., *"In relation to the front door, where did you and your friend sit?"* (Lancaster et al., 2013; D. Leins et al., 2011; Vrij et al., 2009) to specific information from identification cards, such as *"Is Venezia the chief town of your residence region?"* (Monaro et al., 2021). The main benefit of using these questions is that they effectively disrupt a liar's narrative. To differentiate between truth-tellers and deceivers, their reactions to unexpected questions are compared to responses to verifiable control questions, such as *"Are you in front of a computer?"* (Palena et al., 2018; Verigin et al., 2021; Vrij, 2016). These control questions serve as a baseline for genuine reactions, highlighting how liars typically take longer to respond to unexpected questions (Monaro et al., 2021). The efficacy of this technique is well-supported (Lancaster et al., 2013; D. Leins et al., 2011; Sooniste et al., 2015, 2016; Vrij et al., 2011; Warmelink et al., 2012), demonstrating a 71% success rate in distinguishing truth from deceit. This marks a significant improvement over the 56% success rate achieved by conventional methods (Vrij et al., 2017).

4.2.2 Identity Deception

The main focus of the present study is to evaluate a lie detection method that integrates RTs analysis and the unexpected question technique to identify a specific form of lying, identity deception. Identity deception, as articulated by G. A. Wang et al., (2006), pertains to a distinct form of deception wherein individuals intentionally conceal their identity, impersonate another individual, or use counterfeit identity documents, undermining the credibility of identity information through deliberate deceit. Alarmingly, this type of deceit is widespread. For example, in a case study, A. G. Wang et al., (2005) found that nearly a third of police suspects adopted false identities, predominantly by modifying their names. The advent of the digital era

has magnified this challenge, with frequent alterations observed in attributes such as images, names, age, and gender (Utz, 2005; van der Walt et al., 2018). A study by Drouin et al., (2016) revealed that only between 16% and 32% of participants maintained consistent honesty across different websites, with a scant 0.2% believing that others were universally truthful online. Regrettably, online identity deception can also arise from malevolent motives. A particularly disconcerting manifestation of this deceit can be seen in heinous activities such as child grooming (Cano et al., 2014).

On a large scale, identity deception poses a critical threat to national security. Forged documents provide terrorists and criminals an avenue to bypass security protocols, a reality underscored by the proliferation of counterfeit documents at Europe's borders (Frontex, 2021) and incidents like the Twin Towers bombing or the more recent attack on the Brussels Maelbeek metro on March 22, 2016, where a terrorist used several false identities to pass through different states. At the time of his transit through Italy, for example, he entered the country under the name of a former Inter Milan football player (Agenzia Giornalistica Italiana (AGI), 2016). Various criminal activities, including terrorism, human and drug trafficking, and money laundering, often involve counterfeit identification (Boongoen et al., 2010; Gordon et al., 2003). The 2021 EU Organized Crime Report highlights how technological advancements have facilitated the ease with which counterfeit documents can be procured (Europol, 2021). In light of this information, it is evident that deceptive identities are becoming an increasingly pervasive issue. The rapid evolution of these false identities accentuates the pressing need for practical solutions to detect and counter identity deception.

Recent research efforts have employed techniques such as the unexpected question method with mouse and keyboard movement analyses to detect identity deception (Monaro, Gamberini, et al., 2017). These methods differ significantly from those that require prior knowledge of an individual's identity, such as the autobiographical Implicit Association Test (Sartori et al., 2008) and the Reaction Times-based Concealed Information Test (RT-CIT). In many real-world scenarios, examiners might not have access to comprehensive identity information, rendering specific methods less practical. Furthermore, patterns in mouse and keyboard movements can be seen as implicit behavioral indicators; individuals engaging in a computer-based task are often unaware that these specific patterns are being observed and analyzed (Giot et al., 2009). Studies suggest that unintentional behaviors are among the most reliable indicators of deception (Sartori et al., 2018), underscoring the potential of covert deception detection. In an investigative interview context, the subject remains unaware of the objective to detect deception and the specific metrics under observation. The subject's lack of

control over metrics such as latencies in responding to unexpected questions thus increases the effectiveness of the technique.

To the best of our knowledge, no study has thus far explored the combined efficacy of the unexpected question technique, response latency analysis, and error measurements in detecting identity deception during face-to-face investigative interviews. This research aims to evaluate the efficacy of this integrated approach to lie detection. A distinctive contribution of this study is its application within the context of investigative interviews, a setting in which investigators routinely and extensively use audio recordings. Notably, the analysis of these recordings has traditionally been focused primarily on the verbal content of the interviewee's responses, thereby neglecting a detailed examination of response latencies. While interviewees might be aware of the recording procedure, they cannot anticipate the in-depth analysis of the timing of their responses to specific queries. These response latencies act as implicit indicators, providing data less prone to intentional manipulation by the interviewee. This approach has three main advantages. First, it allows to conduct lie detection without the individual's awareness of the credibility assessment, thereby ensuring that the continuity of investigative interviews is not compromised; second, it can be potentially applied retrospectively to interviews that have been conducted, allowing to exploit data that are routinely acquired; third, it does not impact standard procedures of the judicial or police system, since no *ad hoc* tests are administered.

4.3 Materials and Methods

4.3.1 Participants

The study sample comprised 60 native Italian participants (57% female) recruited through the experimenter’s network and word-of-mouth.

An a priori power analysis determined that a sample size of 60 is sufficiently large to achieve a statistical power of $(1-\beta) = 0.95$ in a repeated-measures ANOVA (within-between interaction) involving two groups, given a significance level (α) of 0.05 and an effect size of 0.25. This sample size is consistent with other lie detection research based on response latencies that employed a similar experimental design (Monaro et al., 2021; Monaro, Gamberini, et al., 2017, 2018).

All participants were university students aged between 21 and 29 ($M = 23.68$, $SD = 1.32$) with educational backgrounds ranging from 16 to 19 years ($M = 17.82$; $SD = 0.59$). The two experimental groups were statistically homogeneous regarding age, education, and gender

distribution, ensuring comparability in subsequent analyses. No external incentives (monetary rewards or university credits) were provided. Participants were randomly assigned to one of two experimental conditions: half (30 participants) were placed in the honest condition and instructed to answer interview questions truthfully using their personal information. The other half were in the faking condition and directed to respond with false details consistent with a fictitious identity provided to them.

4.3.2 Experimental Procedure

Ethical clearance was obtained from the ethics committee at the Department of General Psychology, University of Padua (Protocol n. 5399). All methods were performed in accordance with the Declaration of Helsinki. Participants were informed about the confidentiality of their data and their right to withdraw at any time without repercussions. Informed consent was obtained from all participants.

Participants in the honest condition initially completed a file with their details, such as name, surname, date of birth, the first six letters of their tax code, place of birth, residence, address, phone number, email, completed degree, and year of degree completion. Conversely, participants in the faking condition were given five minutes to memorize a fictitious identity with the same categories of information. Following this, both groups performed five arithmetic tasks to serve as a cognitive distraction, thereby facilitating the retention of the false identity. Participants then orally recounted their details. Those in the deceptive group only advanced to the next phase upon accurately recalling their assigned identity. If inaccuracies occurred, they reviewed the details for another five minutes, completed another set of arithmetic tasks, and attempted recall again.

Each participant underwent a standardized, face-to-face interview of 36 open-ended questions, detailed in the Supplementary Materials (S1). Interviews lasted approximately five minutes and began with a reminder that questions pertain to previously provided personal information. The questions were categorized into 12 control, 12 expected, and 12 unexpected items and were presented in a randomized order. Control questions related to the experiment's context or verifiable physical attributes, allowing interviewers to verify responses directly. Response latencies to these questions established a benchmark for the typical duration required for a participant to answer truthfully. Expected questions were based on information directly supplied in the file. Both honest and deceptive participants could readily access answers from their genuine personal details or the memorized fictitious information. On the contrary,

unexpected questions probed information not explicitly documented or recalled but inferred from primary details. Honest participants would respond based on their knowledge, whereas deceptive participants had to fabricate responses on the spot. For example, an individual born on September 22nd would intuitively recognize their zodiac sign as Virgo. Similarly, a resident of Padua would know its regional capital. In contrast, liars, not well-acquainted with their assigned false identities, faced the challenge of quickly deducing the correct answer from their fabricated data.

All sessions were audio-recorded using the iPhone 13's Voice Memo app (iOS 17) to analyze latencies, precisely the time gap between the conclusion of the interviewer's question and the onset of the participant's response, as well as errors. Post-interview, participants completed two Google Forms questionnaires to assess the perceived difficulty in answering each question (rated on a 1 *"very easy"* to 5 *"very difficult"* Likert scale) and the level of anticipation (rated on a 1 *"slightly anticipated"* to 3 *"highly anticipated"* Likert scale) for both expected and unexpected questions.

The total duration of the experiment was approximately 40 minutes. At the end of the session, participants provided written consent. The primary objective of this study was to investigate the effects of unexpected questions on information reconstruction, comparing latencies and error rates between the two groups when presented with unexpected questions. Within the group identified as "deceptive," we aimed to identify differences in their responses to expected versus unexpected questions. To validate the efficacy of our design, the concluding questionnaires were administered to gauge whether participants perceived the unexpected questions as unanticipated and more challenging compared to the expected ones.

4.4 Data Analysis

The Audacity Software (www.audacityteam.org) was employed to extract RTs, providing precision at the millisecond level. The latencies were measured from the end of the interviewer's question to the onset of the interviewee's first meaningful word, excluding fillers such as *'umm.'* Errors were categorized as misaligned answers or claims of ignorance regarding personal details.

As data consisted of a series of repeated individual responses (each participant responded to control, expected, and unexpected questions), the analysis was performed by fitting the Generalized Linear Mixed Model (GLMM). In particular, two GLMMs were built to analyze latencies and error rates separately. In both GLMMs, condition (honest vs. deceptive), sentence type (control vs. expected vs. unexpected), and their interaction were entered as fixed

factors, while latencies and errors were the dependent variables. Participants were entered as random effects in all models. A post-hoc test was conducted to determine which groups and types of questions were responsible for the significant differences identified.

The latency comparison within and between groups was made by running a GLMM with a Gamma distribution and a "Log" link function. This approach was chosen to address the positive skewness and non-normal latency distribution. It also aligns with the actual distribution of latencies, which exhibit variation on a logarithmic scale, with increasing variance corresponding to longer RTs. In contrast, a GLMM with a Poisson distribution was applied to analyze the non-normal count of error distribution as a function of condition, type of question, and their interaction, with random effects accounted for by individual differences (ID). This model choice is appropriate for count data, and we used a "Log" link function to accommodate the multiplicative nature of the effects.

The same analysis was performed to ascertain the influence of condition and type of question on the perceived difficulty of the participants in answering the questions. Given the positive skewness and non-normal distribution, the perceived difficulty across experimental conditions and question categories was evaluated using a GLMM with a Gamma distribution and "Log" link function. This model adeptly accounted for fixed effects (condition and question type) and random effects to accommodate individual subject variability (ID). In contrast, an independent-sample *t*-test assessed the level of anticipation perceived by participants when responding to expected and unexpected questions. These results were extracted from the two final questionnaires. All analyses were performed using Jamovi (Version 2.4) [Computer Software], (2023).

We further examined the feasibility of classifying individual subjects under the two conditions, quantifying the accuracy of such classification. For this aim, we trained and validated a series of Machine Learning classification models through a 10-fold cross-validation procedure using PyCaret, an open-source low-code machine learning library implemented in Python (https://pycaret.org). On applying Pycaret various classifiers having different accuracies are produced and shown in the result section.

The use of machine learning in the forensic classification of individual subjects is both promising and significant. Traditional analyses typically rely on statistical methods that compare two or more groups to discern patterns or differences. Although these statistical analyses provide a comprehensive view of group differences, they fall short in accurately classifying individual subjects. In contrast, machine learning offers an approach that can uncover intricate patterns and relationships often overlooked by conventional statistical

methods. This precision is especially vital in forensic settings, where individual accuracy at a single subject level is required.

4.5 Results

4.5.1 Latencies Analysis

The GLMM built for the latency comparison showed a good fit with a low Chi-squared/DF ratio of 0.0680, indicating no overdispersion. The fixed effects analysis demonstrated significant main effects for condition (χ^2=31.4, df = 1, p <.001), type of question (χ^2=820.3, df = 2, p <.001), as well as for their interaction (χ^2=83.2, df = 2, p <.001), suggesting that both factors and their combination significantly predict RTs. Parameter estimates for fixed effects indicated that the deceptive condition (B = 0.77012, SE = 0.1374, exp(B) = 2.160, z = 5.6041, p < .001) and the unexpected type (B = 1.19839, SE = 0.0482, exp(B) = 3.315, z = 24.8554, p <.001) had particularly strong effects on increasing latencies. The interaction terms were also significant; notably, the interaction of deceptive condition with unexpected type showed a substantial effect (B = 0.84121, SE = 0.0964, exp(B) = 2.319, z = 8.7250, p < .001). Random Components analysis indicates variability in the intercepts across individuals (SD = 0.302, Variance = 0.0914), and the residual standard deviation (SD = 0.309, Variance = 0.0952) suggests additional variation in RTs not explained by the model.

Post hoc tests with Bonferroni correction highlighted significant results for the comparison between honest and dishonest participants in the expected (z = -4.17, p < .001) and unexpected questions (z = -8.55, p < .001), showing that identity liars took considerably longer time to respond.

Additionally, when examining the within-subject latencies in the deceptive group, significant differences were observed between the latencies to control and unexpected questions (z = -23.74, p <.001) and between the expected and unexpected questions (z = -22.42, p < .001). In both instances, the latencies for the unexpected questions were notably higher. The same pattern was also observed within the honest group between the latencies to control and unexpected questions (z = -11.41, p <.001) and between the expected and unexpected questions (z = -12.82, p <.001). Please refer to Figure 4.1 for a graphic representation of the average response latencies between groups.

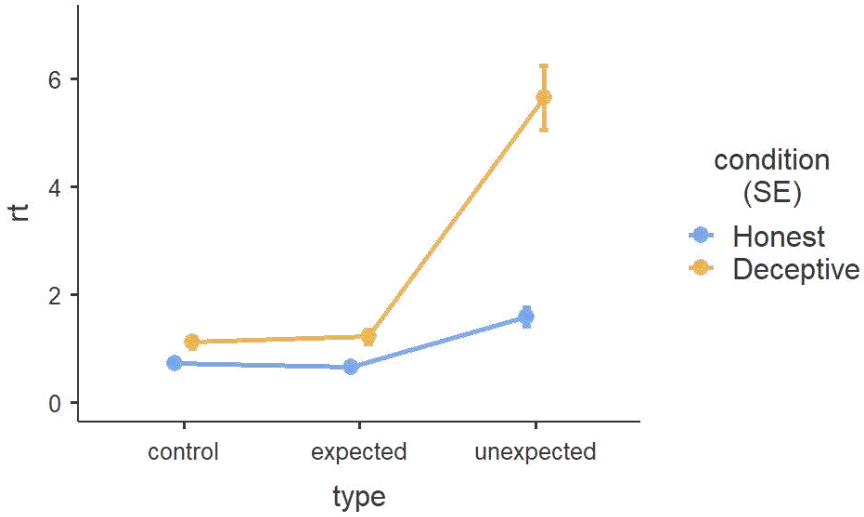


Figure 4.1 Descriptive plots of the average response latencies for honest and deceptive participants in control, expected, and unexpected questions. Error bars displayed in this chart represent the Standard Error.

4.5.2 Errors Analysis

The GLMM built for analyzing the error distribution demonstrates a satisfactory goodness-of-fit, as evidenced by a Chi-squared/DF ratio of 0.884, indicating no overdispersion, and a high marginal R-squared value of 0.814, suggesting a strong relationship between the predictors and the outcome variable.

The fixed effect omnibus tests indicated a significant main effect for condition (χ^2= 5.79, df = 1, p = 0.016) and a for type (χ^2 = 65.62, df = 2, p < .001). However, the interaction between condition and type was not significant (χ^2 =1.89, df = 2, p = 0.388). Examining the fixed effects parameter estimates, the model revealed that the deceptive condition (B = 1.29, SE = 0.538, exp(B) = 3.646, z = 2.41, p = 0.016) significantly increased the error count. Similarly, the expected type (B = 1.79, SE = 0.799, exp(B) = 6.000, z = 2.24, p = 0.025) and the unexpected type (B = 4.10, SE = 0.715, exp(B) = 60.332, z = 5 .73, p <.001) had strong effects on increasing the count of errors. The random effects analysis showed an intercept variance of zero, which suggests no variability between individuals at the baseline level.

Post hoc comparisons with Bonferroni correction revealed several significant differences between levels of condition and type. The comparison between honest and deceptive participants in the unexpected questions (z = -7.884, p < .001) showed that identity liars committed more errors than the honest group. The same emerged also when comparing honest end deceptive participants in expected questions (z = -2.948; p =0.048).

Additionally, within-subject error rate comparisons in the deceptive group revealed a significant difference between the control and unexpected questions (z = -4.925, p < .001) and between the expected and unexpected questions (z = -8.192, p < .001). In both instances, the errors for the unexpected questions were notably higher. The same pattern was also observed within the honest group between the control and unexpected questions (z = -3.197; p = 0.021) and between the expected and unexpected questions (z = -3.495, p < .007), demonstrating that unexpected questions yielded a higher error rate. Please refer to Figure 4.2 for a graphic representation of these results.

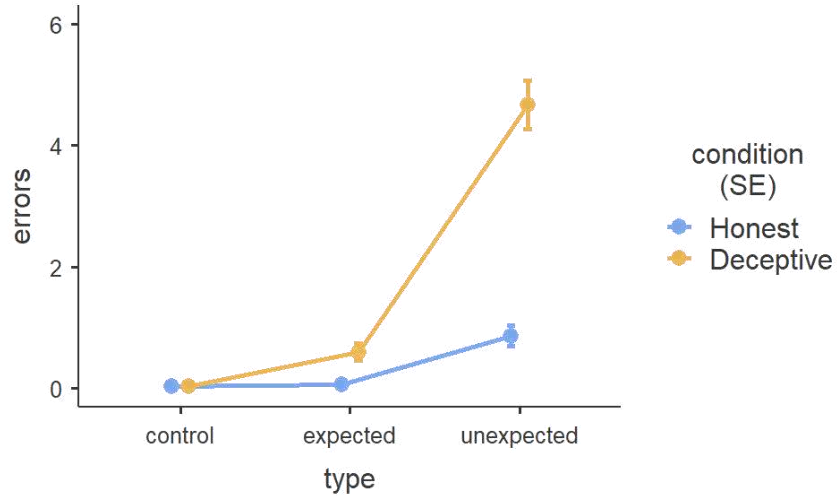


Figure 4.2 Descriptive plots of the sum error rates for honest and deceptive participants in control, expected, and unexpected questions are reported. Error bars displayed in this chart represent the Standard Error.

4.5.3 Questionnaires Analysis

As previously mentioned, participants were instructed to assess each question's degree of anticipation and perceived difficulty by responding to two questionnaires, the first using a 5-point Likert scale and the second using a 3-point Likert scale.

The GLMM demonstrated a good fit with a low Chi-squared/DF ratio of 0.0217, suggesting no overdispersion and endorsing the model's suitability for this data set.

Omnibus tests for fixed effects yielded a significant main effect for condition (χ^2 = 9.76, df = 1, p = 0.002) and for question type (χ^2 = 697.54, df = 2, p < .001), as well as a significant interaction effect between condition and type (χ^2 = 56.56, df = 2, p < .001). These findings assert that both the experimental condition, the type of question, and their interaction significantly contribute to predicting perceived difficulty. Delving into the parameter estimates for fixed effects, the model disclosed that the deceptive condition (B = 0.185, SE = 0.0592, exp(B) =

1.20, $z = 3.12$, $p = 0.002$) and the unexpected question type (B = 0.654, SE = 0.0262, exp(B) = 1.92, $z = 24.99$, $p < .001$) substantially augment perceived difficulty. The interaction between deceptive condition and unexpected question type also exhibited a considerable impact (B = 0.388, SE = 0.0524, exp(B) = 1.47, $z = 7.41$, $p < .001$). The random effect for intercepts across subjects indicated variability (SD = 0.128) with a residual standard deviation (SD = 0.164), hinting at unexplained variance not captured by the model.

Post hoc analyses with Bonferroni correction delineated significant pairwise differences within the condition by type interaction, especially between honest and deceptive participants for expected ($z = -3.364$, $p = .012$) and unexpected questions ($z = -7.354$, $p < .001$), suggesting that identity deceivers perceive greater difficulty than their honest counterparts. No differences emerged when comparing control questions. Comparable significant disparities were also observed within the deceptive group between responses to control and expected ($z = -7.003$, $p < .001$), between control and unexpected questions ($z = -22.865$, $p < .001$), and between expected and unexpected questions ($z = -15.870$, $p < .001$). A similar pattern emerged within the honest group between control and unexpected ($z = -12.454$, $p < .001$) and between expected versus unexpected questions ($z = -12.221$, $p < .001$). However, honest participants did not perceive more difficulty answering expected questions than control questions.

Given the non-normal distribution, a Mann-Whitney Test was conducted to investigate anticipation effects, with the question type as the grouping variable. The results indicated a significant difference in the anticipation scores for expected *(*$M = 2.75$, $SD = 0.31$) versus unexpected questions *(*$M = 1.49$, $SD = 0.23$), $W = 3514.00$, $p < .00$, $r = 0.95$, $SE = 0.11$.

4.5.4 Detection of Deception at Single Subject Level

The concluding analysis was carried out using PyCaret to assess the accuracy of different machine learning classifiers in distinguishing between honest and deceptive participants. For this classification task, latencies and the number of errors were employed as features. Each was further differentiated into control, expected, and unexpected, resulting in a total of six distinct features. The six features were entered into a set of three machine learning algorithms: Logistic Regression, KNN, and Random Forest Classifier. We chose multiple algorithms to prevent the selection of only the highest-performing model by chance, thus checking whether there is substantial variation in classification accuracy across various types of classifiers (Orrù et al., 2020). Moreover, we chose these specific algorithms to test the robustness of our classification performance, since they are based on different underlying

assumptions. Indeed, when machine learning models, based on fundamentally different principles, yield comparable outcomes, it can be reasonably assumed that the results are not reliant on particular assumptions.

Leveraging machine learning models allows for predicting individual behaviors rather than generalizing the collective behavior of a group (Zago et al., 2019). To ensure the out-of-sample generalization, the standard K-fold cross-validation technique (K=10) was employed (Orrù et al., 2020). This technique is critical, mainly when working with limited data samples that require out-of-sample accuracy. By applying this technique, the data sample was divided into 10 unique subsets. The algorithm was trained on 9 subsets to learn to differentiate between honest and deceptive individuals and then validated on the remaining subset to assess its performance. This process was iterated 10 times, with each iteration featuring different training and validation subsets, guaranteeing that every data point (in this case, each of the 60 participants) is utilized for training and validation. This approach provides a comprehensive understanding of the model's performance across different subsets, minimizing the risk of overfitting and allowing a balanced and reliable evaluation of the model's ability to generalize.

At the end of the process, various performance metrics—such as accuracy, Area Under the Curve (AUC), recall, precision, and F1 score—were calculated for each algorithm to reflect the model's overall efficacy. Table 4.1 presents the results from the three algorithms, with Logistic Regression identified as the top-performing algorithm.

Table 4.1 Performance Indices for the three Algorithms in Correctly Identifying Deceptive and Honest Participants.

Model	**Performance Indices**				
	Accuracy	**AUC**	**Recall**	**Precision**	**F1 Score**
Logistic Regression	0.98	1.00	1.00	0.97	0.98
KNN	0.96	0.98	0.95	0.95	0.95
Random Forest Classifier	0.95	1.00	1.00	0.93	0.96

Note. The table presents the values of the performance indices for the top five algorithms used in the study. The indices listed across the columns include Accuracy, Area Under the Curve (AUC), Recall, Precision, and F1 Score. The corresponding values for each algorithm are listed in the respective rows.

Further analysis evaluated the frequency of false positives and false negatives across the ten separate folds. A false positive denotes an instance where an honest individual, as per

ground truth, is erroneously identified as deceptive by the algorithm. In contrast, a false negative occurs when a deceptive individual, according to the ground truth, is incorrectly classified as honest. Across the ten folds, there was one false positive, resulting in an average rate of 0.1 for false positives and 0 for false negatives. Consequently, the Logistic Regression algorithm exhibited high effectiveness with a low misclassification rate. This achievement is especially noteworthy given the relative simplicity of Logistic Regression as a classification method. It managed to deliver robust results, bypassing the need for more intricate approaches like Support Vector Machines (SVM), Bayesian methods, or deep neural networks.

The Feature Importance Plot, illustrated in Figure 4.3, delineates the impact of each input variable on the Logistic Regression's predictions. This visual representation aids in comprehending which factors most significantly influence the predictions and provides insights into the interplay among variables. Notably, the 'errors on unexpected questions' variable was the most indicative in distinguishing between honest and deceptive individuals, with 'response times to unexpected questions' being a close second.

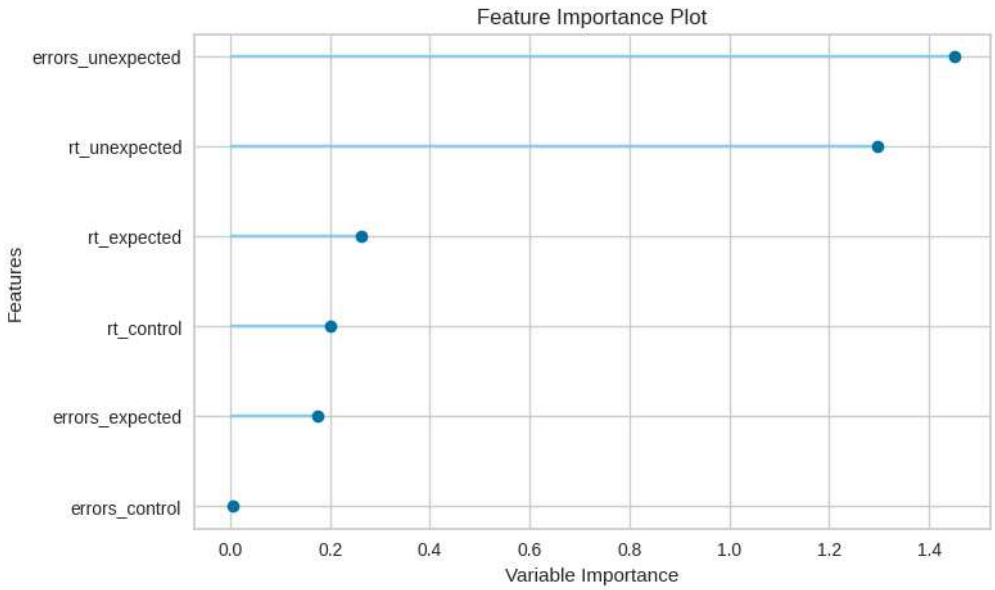


Figure 4.3 Feature Importance Plot. The x-axis of the plot quantifies the importance of each variable, reflecting its influence on the predictive model. The y-axis lists the six input variables in descending order of importance, from the most influential to the least.

4.5.5 Detection of Deception in Single-Faked Answers

Despite the accuracy in distinguishing between truthful and deceptive subjects is notably high, the scenario where a guilty suspect lies to all questions is exceedingly rare in investigative contexts. Typically, a guilty individual tends to lie primarily to directly incriminating questions,

thereby complicating the detection process. Nonetheless, adopting machine learning algorithms enables the estimation of probabilities for accurately categorizing each unexpected response as either truthful or deceptive. By employing the K-fold cross-validation method, response latencies and error rates for unexpected questions (for both honest and deceptive groups) were used as the input variables for training and testing the three machine learning models in categorizing each answer as honest or deceptive. The performance metrics of the models are detailed in Table 4.2.

Table 4.2 Performance Indices for the Three Algorithms in Classifying Single Unexpected-Faked Answers.

Model	Performance Indices				
	Accuracy	**AUC**	**Recall**	**Precision**	**F1 Score**
Logistic Regression	0.74	0.85	0.64	0.81	0.71
KNN	0.74	0.81	0.73	0.75	0.74
Random Forest Classifier	0.71	0.79	0.70	0.71	0.70

Note. The table presents the values of the performance indices for the top five algorithms used in the study. The indices listed across the columns include Accuracy, Area Under the Curve (AUC), Recall, Precision, and F1 Score. The corresponding values for each algorithm are listed in the respective rows.

This outcome demonstrates that the proposed methodology can identify the individual subject and detect deception in each response given by the deceiver, albeit with a lower level of accuracy.

4.5.6 Relative Measures

The method proposed here differs from other lie detection techniques as it can be applied offline to audio recordings of investigative interviews. However, it is expected to encounter subjects from various age groups in practical applications. Considering the well-documented phenomenon that latencies tend to increase with advancing age (Tun & Lachman, 2008; Yan & Tourangeau, 2008) and that our group of participants did not include all age groups, it is important to evaluate the accuracy in identifying deception by computing indices that compare the responses to unexpected questions with expected and control questions within subjects. Such indices are particularly valuable as they diminish participant variance by comparing the subjects against themselves, thus using a subject-specific baseline (e.g., RTs to control

questions). This process may still accurately identify deceptive responses, albeit with a slight reduction in accuracy. As an illustration, we incorporated additional measures, which we can designate as relative, by computing all conceivable data combinations. In this procedure, for each participant, the mean of one latency is subtracted from the mean of another, and the resultant outcome is subsequently divided by the mean of a third latency. The list with all the tested possible combinations is reported in the Supplementary Materials (S2).

Subsequently, these variables were employed to classify our participants, employing the identical procedure detailed in the preceding paragraph. The outcomes reported in Table 4.3 revealed the KNN as the top-performing algorithm.

Table 4.3. Performance Indices for the three Algorithms Using the Relative Measures.

Model	Performance Indices				
	Accuracy	AUC	Recall	Precision	F1
KNN	0.81	0.87	0.80	0.84	0.80
Logistic Regression	0.79	0.90	0.80	0.79	0.78
Random Forest Classifier	0.79	0.81	0.75	0.83	0.77

Note. The table presents the values of the performance indices for the three algorithms used in the study. The indices listed across the columns include Accuracy, Area Under the Curve (AUC), Recall, Precision, and F1 Score. The corresponding values for each algorithm are listed in the respective rows.

Furthermore, an analysis was conducted to ascertain the impact of each feature on the classification task, which the analysis unveiled as (control-unexpected)/control, denoting the result of the latencies to expected questions subtracted from that of unexpected questions, and then divided by the latencies to expected questions. Employing the same machine learning procedure, this single variable reaffirmed the previously obtained results, yielding an accuracy rate of 81% for the top-performing algorithm (Table 4.4).

Table 4.4. Performance Indices for the Three Algorithms Only Using the Most Informative Relative Measures.

Model	Performance Indices				
	Accuracy	AUC	Recall	Precision	F1
Logistic Regression	0.81	0.92	0.80	0.84	0.80
KNN	0.81	0.82	0.75	0.88	0.79
Random Forest Classifier	0.70	0.79	0.70	0.68	0.67

Note. The table presents the values of the performance indices for the three algorithms used in the study. The indices listed across the columns include Accuracy, Area Under the Curve (AUC), Recall, Precision, and F1 Score. The corresponding values for each algorithm are listed in the respective rows.

4.6 Discussion

This study introduces a novel technique for evaluating the credibility of responses obtained during investigative questioning. The technique consists of computing the latencies from the end of the question and the beginning of the response in a recorded interview. This technique addresses a significant gap in existing scientific literature by offering a method ensuring that:

1. The individual being assessed is unaware of the credibility evaluation.
2. The evaluation process does not disrupt the flow of investigative interviews.

Our developed procedure only requires strategically incorporating three types of questions: control questions, which the respondent is compelled to answer truthfully; expected questions, where fake responses can be premeditated; and unexpected questions, which are challenging to anticipate. A crucial aspect of our methodology's practical application is identifying unexpected questions. This is due to the observation that response latencies and error rates for expected questions closely resemble control questions. In contrast, the latencies and error rates associated with unexpected questions provide valuable insights, proving to be an indirect indicator of deception as difficult to manipulate consciously. By comparing the RTs and errors of unexpected questions to those of expected and control questions, we can accurately determine the individual's truthfulness.

Another goal of our proposed methodology is to differentiate truthful responses from deceptive ones. In forensic practice, it is rare for a guilty suspect to provide false answers to all

questions. Typically, a guilty individual is more likely to lie selectively, particularly in response to incriminating inquiries, thereby complicating the process of identifying deceit. Our developed method addresses this challenge by employing machine learning models. These models enable the estimation of the probability that a given response is truthful or deceptive.

These goals were achieved by dividing 60 participants into two distinct groups—honest and deceptive. Before the main interview, participants assigned to the deceptive group were instructed on a fictitious identity to ensure their familiarity with the false information. Following this orientation phase, participants from both groups were subjected to a series of questions that included "control" queries requiring verifiable answers, "expected" questions delving into the specifics of identity, and "unexpected" questions that probed for information that was not directly provided but could be inferred from basic details. For instance, an unexpected question for a participant could be, *"What is your zodiac sign?"* based on their provided date of birth. While truthful participants could instinctively answer such questions based on authentic personal knowledge, deceptive participants were compelled to fabricate responses on the spot, likely increasing their cognitive load, prolonging their RTs, and introducing more errors.

The questions were randomized to challenge the participants, introducing a cognitive switch cost, particularly for deceptive individuals. This cost was incurred as they navigated the mental shift required to alternate between fabricating answers for the expected and unexpected questions and providing truthful responses to the control questions. All interviews were audio-recorded, enabling a detailed analysis of response latencies and inaccuracies. Post-interview, participants completed questionnaires assessing their anticipation level and perceived difficulty for each question category.

The analysis indicated that when confronted with unexpected questions, deceptive participants exhibited prolonged response latencies and a higher frequency of errors than their honest counterparts. These findings align with the study conducted by Monaro, Fugazza, et al., (2017). In a similar vein to our research, the experiment by the authors required participants to learn and adopt a fictitious identity profile. This was followed by a series of control and target queries, both expected and unexpected, pertaining to their assumed identities, to which the participants responded via a computerized task. In this task, the authors scrutinized mouse-tracking trajectories. Among various metrics, RTs and error rates were particularly revealing in differentiating between honest and deceptive responses. Notably, the results indicated that the responses of deceptive individuals to unexpected questions were markedly distinct from those of truth-tellers in terms of RTs, with deceivers exhibiting slower responses. Additionally,

deceivers committed a higher number of errors compared to truth-tellers. This finding aligns with existing literature, which suggests an augmented cognitive burden borne by individuals engaged in deceit, particularly when they lack premeditated responses (Clemens et al., 2011; Derrick et al., 2013; Monaro et al., 2017, 2018, 2019, 2021; Vrij et al., 2009).

Our results also indicated that individuals feigning their identity exhibited significant discrepancies in latencies and error rates when responding to expected versus unexpected questions. This was characterized by an increase in errors and delayed responses, particularly to the latter. Furthermore, when control questions were juxtaposed with unexpected questions, differential response patterns emerged, with the unexpected questions consistently eliciting longer response latencies and a greater number of errors. These findings are consistent with the study by Monaro et al., (2021), in which the authors focused on detecting individuals fabricating their identities. However, instead of conducting face-to-face interviews, participants in Monaro's study responded to questions through a computerized task. The authors recorded and analyzed both RTs and the Inverse Efficiency Score (IES), a composite measure that integrates both speed and accuracy. In their study, individuals who were lying exhibited longer RTs and higher IES scores when responding to unexpected questions compared to expected ones, further corroborating our observations. Such outcomes are predicated on the theoretical underpinning of the unexpected questions technique, which posits that the deliberate augmentation of cognitive load on deceivers disrupts their ability to sustain the deceptive act (Masip & Herrero, 2015; Vrij et al., 2006, 2012; Walczyk et al., 2013).

However, a similar trend was also observed within the honest group: unexpected questions engendered extended latencies and an increased error incidence, corroborating the findings of Parkhouse & Ormerod's study (2018) that the act of responding to unexpected questions inherently imposes a higher cognitive demand, even on individuals providing truthful answers. The study by Monaro et al. in 2021 also revealed that truth-tellers exhibited longer RTs and higher scores on the IES when responding to unexpected questions as opposed to expected ones.

Notably, dishonest participants did not significantly differ from honest ones in terms of latency and error rates when responding truthfully to control questions, corroborating the literature's results, which found no differences between honest and deceptive participants in control questions (Monaro, Fugazza, et al., 2017; Monaro, Galante, et al., 2018).

However, a notable divergence from Monaro's findings regarding expected questions was observed in our study. Monaro's studies did not report significant differences in responses to expected questions. In contrast, our research identified distinct differences between honest

and deceptive responses even in the context of expected questions, thereby highlighting the impact of deception. In fact, our deceptive participants demonstrated variances when deception was a factor, both in expected and unexpected questions. Particularly in responding to unexpected questions, where they could not premeditate a deceptive response, these differences were more pronounced. Consistent with the literature, unexpected questions have proven efficacious in evoking deception cues from liars (Knieps et al., 2013; Lancaster et al., 2013; D. Leins et al., 2011; Monaro et al., 2021; Sooniste et al., 2015, 2016; Vrij et al., 2009; Warmelink et al., 2012).

This finding is corroborated by participant perceptions of the interview's difficulty and anticipation. No significant differences were detected between groups in the perceived difficulty of control questions. However, unexpected questions were rated as considerably more unpredictable than expected by both groups. Nevertheless, participants engaging in deception reported a higher level of challenge in addressing both expected and unexpected questions than their truthful counterparts, mirroring the outcomes of Parkhouse and Ormerod (2018), where deceptive subjects perceived the interview questions as more difficult regardless of the question type (expected or unexpected). A notable divergence between the experimental groups emerged in their mental taxation when responding to expected questions instead of control questions. Deceivers reported a greater cognitive challenge with expected questions—a complexity not paralleled in the honest group, who exhibited no significant difference in perceived difficulty between these question types. The perceived increased cognitive load for deceivers in response to expected questions as compared to control questions may be attributed to the necessity of reconciling known truths with the fabrication of their responses, introducing an additional layer to the response formulation process (Sheridan & Flowers, 2010).

In conclusion, our investigation assessed the ability of machine learning models to differentiate between individuals who provide deceptive responses regarding their identity and those who are truthful. We selected machine learning due to its demonstrated proficiency in classifying subjects on an individual basis, which aligns with the objectives of our research and the imperatives of forensic applications. Traditional statistical methods often focus on group-level variations and may lack the precision required for individual case analysis. In contrast, machine learning models offer a robust framework for detailed assessments of an individual's likelihood of being truthful or deceptive—a crucial aspect within the forensic domain, where individual accuracy at a single subject level is required.

As presented in Table 4.1, our models attained an accuracy rate of 98%, with a mean of only 0.1 false positives and 0 false negatives. The two most indicative features for

discriminating between truth-tellers and deceivers were the frequency of errors and the latencies to unexpected questions. Furthermore, our methodology demonstrated an accuracy rate of 74% in distinguishing between deceptive and honest unexpected answers. This makes it possible to identify whether the single answer, given by the subject to an unexpected question is lying or honest.

Notwithstanding the successful execution of our experiment, it is crucial to acknowledge and address its limitations. The demographic constraints of the participant pool, comprised exclusively of university students aged between 21 and 29 years, limiting the generalizability of our findings to a broader population. Further research should investigate these dynamics in both older and younger populations. Moreover, our methodology primarily relies on reaction times and error rates as indicators of deception. However, these metrics may be influenced by a myriad of factors not directly related to deceit, such as anxiety levels, cognitive load from unrelated sources, or genuine difficulties in recalling specific details. Future research should aim to develop more nuanced indicators of deception that can effectively distinguish between deceitful behaviors and those influenced by external or internal factors.

Additionally, the laboratory setting may have affected participants' propensity for deception. Unlike real-world forensic interviews, where the consequences of being caught are significant, the low-risk environment of a laboratory may not compel participants to employ their most convincing deceptive strategies. Future studies should consider incentivizing participants for deception to emulate more realistic high-stakes conditions. Supporting this notion, Vrij et al., (2010) found that the complexity of a lie is related to the cognitive load involved; hence, participants might exhibit more pronounced cues of deception under higher stakes. Therefore, further research should explore the impact of increased stakes on deceptive behavior. These considerations highlight the complexity of interpreting deception-based metrics and underscore the importance of adopting a more nuanced and varied methodological framework to accurately address the multifaceted nature of human behavior.

Another notable limitation was the exclusive use of a single experimenter for both conducting and coding the audio recordings of the interviews. Employing a double-blind methodology would enhance the credibility of the results and mitigate potential biases. Moreover, it is essential to recognize that the technique of posing unexpected questions is not universally applicable. This situation may arise because there may simply be no unforeseen questions that can be appropriately asked concerning the subject of the investigation. Indeed, in some instances, the nature of the topic under discussion might not accommodate the possibility of unexpected questions.

In conclusion, the application of machine learning models for the identification of deceptive behaviors demonstrates promising results, characterized by high accuracy rates. However, it is imperative to acknowledge the ethical dilemmas that arise, particularly concerning false positives and negatives. The generalizability of machine learning models beyond the specific datasets and conditions of the study poses a significant challenge that demands greater attention, casting doubt on their reliability and universal applicability. Therefore, it is crucial that future research not only focuses on enhancing the accuracy of these models but also on exploring and implementing robust ethical frameworks that can guide their development and utilization responsibly. Although in this study, the rate of false positives is exceedingly low, and no false negatives were detected, the misclassification of innocent individuals as guilty, or vice versa, is not merely a technical issue but has profound ethical implications, especially in investigative contexts where it can alter the lives of the individuals involved.

Despite these limitations, identifying them is essential for improving experimental designs and directing subsequent inquiries. Future research should endeavor to replicate and extend the findings of this study by employing a more diverse and representative sample that encompasses various ages and socio-cultural backgrounds. This approach is crucial for enhancing the external validity of our results and ensuring their applicability across a broader spectrum of the population. Additionally, integrating non-verbal cues that may indicate deception, such as facial expressions, eye movements, body language, or physiological responses, into the research methodology is suggested. The inclusion of these non-verbal signals would provide a more holistic understanding of the mechanisms underlying deceptive behavior, acknowledging the multifaceted nature of communication. Such a comprehensive approach would not only deepen our insights into deception but also potentially improve the accuracy and reliability of deception detection techniques. Additionally, investigating the effectiveness of the unexpected questions technique in scenarios where deceivers make only minor alterations to their actual biographical information, as commonly occurs (G. Wang et al., 2004), would be beneficial. Such investigations would clarify the effectiveness of this technique in detecting subtle identity deceptions.

4.7 Conclusion

This study highlights the considerable promise of employing audio recordings and analysis of response latencies to improve deception detection within forensic interview contexts. Although individuals may be aware that their spoken responses are being captured,

they frequently do not realize that their response latencies, subtle as they may be, can convey critical information. Our research has shown that these latencies yield valuable insights, revealing cognitive discrepancies between truthful individuals and those who are not, as evidenced by differences in their response latencies. Our focus was specifically on detecting false identities, a widespread and concerning type of deception. In today's digital era, with the ease of obtaining counterfeit documents and the increasing manipulation of personal attributes like images, names, and birth dates, the issue of false identity is pressing and necessitates urgent attention due to its association with various criminal activities and security breaches.

The results indicate that those impersonating others demonstrate significantly longer latencies and higher error rates than their truthful counterparts when answering unexpected questions that catch them unprepared. This difference highlights the increased cognitive burden on deceivers who must quickly fabricate responses on the spot, without the aid of premeditated replies. Their lack of preparedness for unforeseen inquiries affects their performance adversely. The research confirms that leveraging this vulnerability exposes distinct implicit cues indicative of deceit. Additionally, by randomly interspersing control, expected, and unexpected questions, the study found that deceivers are forced to endure an added cognitive load as they oscillate between honesty and deception.

In forensic environments, where discerning the truth is critical, the potential of this method is evident. While overt verbal cues can be consciously controlled, involuntary latencies serve as more stable and dependable markers of veracity. Investigators can track these time stamps via audio recordings, thus gaining a window into the cognitive processes of the individual and the veracity of their account. Considering the complexities associated with deceit in legal settings, this methodology presents a scientifically based tool for distinguishing between truthfulness and deception. This investigation is an initial step toward creating more sophisticated defenses against the growing menace of identity fraud.

Data availability statement. The datasets generated and analyzed during the current study are available in our public repository: https://osf.io/r5z67/

Conclusion

This dissertation explores the multifaceted nature of deception, integrating insights from psychology and computer science. It provides a thorough historical analysis and employs state-of-the-art methods to enhance our comprehension of the intricate dynamics of human deceit. This research introduces novel strategies applicable across diverse scenarios, addressing current challenges and offering pragmatic solutions.

The project's application of the Term Frequency-Inverse Document Frequency (TF-IDF) technique marks a groundbreaking advancement in detecting deceptive answers in psychological assessments. By achieving a highly refined level of deception detection, this method opens new avenues for increasing the trustworthiness of questionnaires, especially in forensic contexts. A key feature of this approach is its capacity to consider individual variances and the diverse circumstances under which dishonesty may occur. As detailed in Chapter 2, this method allows for nuanced interpretation due to the interplay between each participant's unique response pattern throughout the questionnaire and the collective response trends to the same questions. Although the prerequisite for applying this technique to any test requires data from a representative sample that answers truthfully, a unique aspect of this technique, compared to other lie detection methods, is its reliance solely on data from an honest sample for classifying responses. This feature streamlines its implementation, yet it underscores the necessity for additional research.

Further investigation is essential to apply this method to specific clinical samples, strengthening its validity and practicality. In scenarios like the one proposed, where individuals may present themselves overly positively (faking good), there isn't a corresponding clinical group for comparison. However, in situations where individuals might simulate symptoms (faking bad), such as feigning depression, the dynamics change due to the presence of three distinct groups: a truthful healthy sample, a deceitful healthy sample feigning depression, and a potential group genuinely suffering from the condition. Applying the TF-IDF technique to these distinct groups could yield fascinating insights, enabling not only the differentiation of truth from falsehood but also the distinction between deceitful individuals and those genuinely affected by the disorder under investigation.

The second project (Chapter 3) focuses on the workplace environment, specifically aiming to identify the presence of lies related to drug use in this context. This issue is crucial due to its direct impact on individual worker health, workplace safety, and overall company economic efficiency. It represents a significant, and in some cases, still underestimated

problem. In this framework, the kinematic mouse movement analysis was introduced as an implicit and non-invasive tool for detecting deceptive responses in drug tests. This approach is particularly intriguing because it highlights the complex ethical, moral, and legal issues that emerge in applying drug testing protocols in the workplace. These issues include balancing workplace safety with respect for employee privacy. Ultimately, the application of this approach could contribute to this ongoing dialogue and marks a significant advancement in this field, offering new insights for addressing the challenges posed by drug use at work.

Despite the preliminary nature of the study, which requires further optimization, this methodology presents significant cost and ease of implementation advantages. With the COVID-19 pandemic, many work practices, including job interviews, have moved online. In this context, proposing a methodology that can be easily managed remotely without requiring specific technical expertise becomes even more relevant. To guide future research, it is crucial to build upon the preliminary conclusions of this study, focusing on enhancing the ecological validity and refining online data collection methods.

This study paves the way for further research and innovations in the field, opening new horizons for understanding and mitigating the complexities associated with drug use at work. By integrating technological advancements and ethical considerations, it establishes a precedent for future initiatives aimed at improving both individual well-being and organizational integrity.

Finally, the fourth chapter introduces a technique combining unexpected questions with a detailed analysis of errors and response times during face-to-face investigative interviews. Although the use of unexpected questions and the analysis of errors and response times are concepts already known and explored in forensic literature, their combined application in the analysis of audio recordings of interrogations represents a significant novelty. This technique fills an important gap in the existing scientific literature, offering a method that stands out for its ability to integrate seamlessly into the interrogation flow without the interviewee realizing that they are being subjected to a credibility assessment.

Another valuable element of this methodology is its retrospective applicability. This means that responses can be recorded and analyzed later, ensuring that the individual is unaware of the ongoing evaluation. This feature is particularly relevant since audio recording is a standard procedure during investigative interrogations. Therefore, even if the individual is aware of the recording practice, the evaluative aspect of credibility remains concealed, preserving the integrity and effectiveness of the interview.

A further aim of this methodology is to differentiate not only individuals who lie but also identify truthful responses from deceptive ones. In forensic practice, it is rare for a guilty suspect to provide false answers to all questions. Typically, a guilty individual tends to lie selectively, especially when faced with incriminating questions, thereby complicating the task of identifying deception. To tackle this challenge, this study utilizes machine learning models, which can estimate the probability that a specific response is truthful or deceptive. This research opens promising perspectives for fighting the growing problem of false identities, proposing an approach that could revolutionize interrogation techniques and credibility analysis in forensic investigations.

To conclude, this dissertation has offered a multidisciplinary analysis of lie detection, a phenomenon as ancient as humanity, exploring and applying theories and methodologies ranging from cognitive psychology to machine learning. Through innovative and well-established techniques, the work has expanded the understanding of human deception and proposed tailored solutions to address challenges in lie detection across various domains, such as personality questionnaires, workplace drug tests, and investigative interviews.

The research highlighted how integrating traditional approaches with technological innovations can effectively address deception, shedding light on the complex and multifaceted nature of deceitful practices. Significant advancements in computational tools, representing a strength of innovation in this field, were examined. Besides being technologically advanced, these tools can improve accuracy and efficiency in lie detection across various contexts. The synthesis of these insights opens the path for future research and practical applications, offering a promising route for research in the field.

Supplementary Materials

S1: Standardized order of questions

At the end of the question, the initials "C," "E," or "U" enclosed within parentheses categorize the query as either control, expected, or unexpected, respectively.

1. What is the capital city of your region of birth? (U)
2. Where were you born? (E)
3. What is the color of the shirt you are wearing? (C)
4. What is the capital city of your region of residence? (U)
5. What is your residential address? (E)
6. In which city do you reside? (E)
7. What is your zodiac sign? (U)
8. How many odd numbers are in your date of birth? (U)
9. In which region of Italy are you currently located? (C)
10. What are the last three digits of your phone number? (E)
11. What year is it? (C)
12. What color are your eyes? (C)
13. What is your first name? (E)
14. In which month were you born? (E)
15. How old are you? (U)
16. What are your initials? (U)
17. What undergraduate degree did you complete? (E)
18. What are the first three digits of your phone number in reverse order? (U)
19. In which season are we currently? (C)
20. In which year were you born? (E)
21. What is your gender? (C)
22. What color are your shoes? (C)
23. What is your email address? (E)
24. In which year did you earn your bachelor's degree? (E)
25. How tall are you? (C)
26. What shoe size do you wear? (C)
27. In which region were you born? (U)

28. How old were you in the year [...] when you earned your bachelor's degree? (U)
29. In which month are we currently? (C)
30. Name two regions that border the region where you reside. (U)
31. What is your last name? (E)
32. What color is your hair? (C)
33. Is your birthdate closer in time to Christmas or Easter? (U)
34. How old will you be in 2025? (U)
35. In which city are you located at this moment? (C)
36. What are the first six letters of your tax identification code? (E)

S2: ***Relative Measures List***

List of combinations employed in the analysis as relative measures. Each was computed by subtracting the mean of one latency from the mean of another, and the resulting value was then divided by the mean of a third latency.

1. (control-expected)/expected,
2. (control-expected)/control,
3. (control-unexpected)/control,
4. (control-unexpected)/unexpected.
5. (expected-control)/control,
6. (expected-control)/expected,
7. (expected-unexpected)/expected,
8. (expected-unexpected)/unexpected,
9. (unexpected-control)/control,
10. (unexpected-control)/unexpected,
11. (unexpected-expected)/expected,
12. (unexpected-expected)/unexpected,

Bibliography

Abney, D. H., McBride, D. M., Conte, A. M., & Vinson, D. W. (2015). Response dynamics in prospective memory. *Psychonomic Bulletin & Review*, *22*(4), 1020–1028.

Agenzia Giornalistica Italiana (AGI). (2016). *Bruxelles: kamikaze uso' identita' ex giocatore dell'Inter*.

Allen, L. M., Conder, R. L., Green, P., & Cox, D. R. (1997). CARB'97 manual for the computerized assessment of response bias. *Durham, NC: CogniSyst*.

American Psychiatric Association. (2013). *DSM 5*. American Psychiatric Association.

American Psychiatric Association, A. P. A., & A. P. A. (2013). *Diagnostic and statistical manual of mental disorders: DSM-5* (5th ed.).

Anolli, L. (2003). *Mentire* (Il Mulino, Ed.).

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval* (Vol. 463, Issue 1999). ACM press New York.

Barca, L., & Pezzulo, G. (2012). Unfolding visual lexical decision in time. *PloS One*, *7*(4), e35932.

Beaber, R. J., Marston, A., Michelli, J., & Mills, M. J. (1985). A brief test for measuring malingering in schizophrenic individuals. *The American Journal of Psychiatry*.

Ben-Porath, Y. S., & Tellegen, A. (2020). *MMPI-3: Minnesota Multiphasic Personality Inventory-3 Technical Manual*. University of Minnesota Press.

Benussi, V., & Antonelli, M. (2002). Die atmungssymptome der lüge. In *Psychologische Schriften, Band 1* (pp. 365–394). Brill.

Binder, L. M., & Willis, S. C. (1991). Assesment of motivation after financially compensable minor head trauma. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(2), 175.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335.

Blandón-Gitlin, I., Fenn, E., Masip, J., & Yoo, A. H. (2014). Cognitive-load approaches to detect deception: searching for cognitive mechanisms. *Trends in Cognitive Sciences*, *18*(9), 441–444. https://doi.org/10.1016/j.tics.2014.05.004

Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation. *Journal of Traumatic Stress*, *28*(6), 489–498.

Böckenholt, U., & Meiser, T. (2017). Response style analysis with threshold and multi-process IRT models: A review and tutorial. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 159–181.

Bond Jr, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234.

Boongoen, T., Shen, Q., & Price, C. (2010). Disclosing false identity through hybrid link analysis. *Artificial Intelligence and Law*, *18*(1), 77–102. https://doi.org/10.1007/s10506-010-9085-9

Braun, H. I., Jackson, D. N., & Wiley, D. E. (2001). Socially Desirable Responding: The Evolution of a Construct. In *The Role of Constructs in Psychological and Educational Measurement* (Issue 1958, pp. 49–69). L. Erlbaum Associates. https://doi.org/10.4324/9781410607454-10

Breiman, L., Cox, D., & Breiman, L. (2001). Comment-statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Brett, A., Phillips, M., & Beary, J. (1986). Predictive power of the polygraph: can the" lie detector" really detect liars? *The Lancet*, *327*(8480), 544–547.

Cano, A. E., Fernandez, M., & Alani, H. (2014). *Detecting Child Grooming Behaviour Patterns on Social Media* (pp. 412–427). https://doi.org/10.1007/978-3-319-13734-6_30

Celli, M., Mazzonetto, I., Zangrossi, A., Bertoldo, A., Cona, G., & Corbetta, M. (2022). One-year-later spontaneous EEG features predict visual exploratory human phenotypes. *Communications Biology*, *5*(1), 1361. https://doi.org/10.1038/s42003-022-04294-9

Christ, S. E., Van Essen, D. C., Watson, J. M., Brubaker, L. E., & McDermott, K. B. (2009). The Contributions of Prefrontal Cortex and Executive Control to Deception: Evidence

from Activation Likelihood Estimate Meta-analyses. *Cerebral Cortex*, *19*(7), 1557–1566. https://doi.org/10.1093/cercor/bhn189

Clemens, F., Granhag, P. A., & Strömwall, L. A. (2011). Eliciting cues to false intent: A new application of strategic interviewing. *Law and Human Behavior*, *35*(6), 512–522. https://doi.org/10.1007/s10979-010-9258-9

Coid, J., Bebbington, P., Jenkins, R., Brugha, T., Lewis, G., Farrell, M., & Singleton, N. (2002). The National Survey of Psychiatric Morbidity among prisoners and the future of prison healthcare. *Medicine, Science and the Law*, *42*(3), 245–250.

Colasanti, M., Melis, G., Monaro, M., Ricci, E., Bosco, F., Rossi, M., Biondi, S., Verrocchio, M. C., Di Domenico, A., & Mazza, C. (2023). Did You Commit a Crime There? Investigating the Visual Exploration Patterns of Guilty, Innocent, Honest, and Dishonest Subjects When Viewing a Complex Mock Crime Scene. *Journal of Nonverbal Behavior*, 1–25.

Conroy, M. A., & Kwartner, P. P. (2006). The definition of malingering. *Applied Psychology in Criminal Justice*, *2*(3), 30–51.

Corral, A., Durán, J., & Isusi, I. (2012). *Use of alcohol and drugs at the workplace*.

Crawford, J., Smith, G., Maylor, E., Della Sala, S., & Logie, R. (2003). The Prospective and Retrospective Memory Questionnaire (PRMQ): Normative data and latent structure in a large non-clinical sample. *Memory*, *11*(3), 261–275.

Crowne, D. P., Crowne, D. P., & Marlowe, D. (1964). *The approval motive: Studies in evaluative dependence*.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D. G., Acharyya, M., Loughead, J. W., Gur, R. C., & Langleben, D. D. (2005). Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *Neuroimage*, *28*(3), 663–668.

De Boeck, P., & Jeon, M. (2019). An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00102

Debey, E., Liefooghe, B., De Houwer, J., & Verschuere, B. (2015). Lie, truth, lie: The role of task switching in a deception context. *Psychological Research*, *79*(3), 478–488.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, *70*(5), 979.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. https://doi.org/10.1037/0033-2909.129.1.74

Derrick, D. C., Meservy, T. O., Jenkins, J. L., Burgoon, J. K., & Nunamaker, J. F. (2013). Detecting Deceptive Chat-Based Communication Using Typing Behavior and Message Cues. *ACM Transactions on Management Information Systems*, *4*(2), 1–21. https://doi.org/10.1145/2499962.2499967

Draheim, C., Hicks, K. L., & Engle, R. W. (2016). Combining Reaction Time and Accuracy. *Perspectives on Psychological Science*, *11*(1), 133–155. https://doi.org/10.1177/1745691615596990

Drouin, M., Miller, D., Wehle, S. M. J., & Hernandez, E. (2016). Why do people lie online? "Because everyone lies on the internet." *Computers in Human Behavior*, *64*, 134–142. https://doi.org/10.1016/j.chb.2016.06.052

Duran, N. D., Dale, R., & McNamara, D. S. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review*, *17*(4), 486–491. https://doi.org/10.3758/PBR.17.4.486

Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, *25*, 69–91.

Edens, J. F., Cruise, K. R., & Buffington-Vollum, J. K. (2001). Forensic and correctional applications of the Personality Assessment Inventory. *Behavioral Sciences & the Law*, *19*(4), 519–543.

Einarsen, S., Hoel, H., & Notelaers, G. (2009). Measuring exposure to bullying and harassment at work: Validity, factor structure and psychometric properties of the Negative Acts Questionnaire-Revised. *Work & Stress*, *23*(1), 24–44.

Ekman, P., & Friesen, W. V. (1972). Hand movements. *Journal of Communication*, *22*(4), 353–374.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system. *Environmental Psychology & Nonverbal Behavior*.

Ekman, P., Friesen, W. V, & Simons, R. C. (1985). Is the startle reaction an emotion? *Journal of Personality and Social Psychology*, *49*(5), 1416.

El-Bassel, N., Schilling, R. F., Schinke, S., Orlandi, M., Sun, W.-H., & Back, S. (1997). Assessing the utility of the Drug Abuse Screening Test in the workplace. *Research on Social Work Practice*, *7*(1), 99–114.

Ellingson, J. E., & McFarland, L. A. (2011). Understanding faking behavior through the lens of motivation: An application of VIE theory. *Human Performance*, *24*(4), 322–337.

European Monitoring Centre for Drugs and Drug Addiction. (2022). *Workplaces and drugs: health and social responses*.

Europol. (2021). *European Union serious and organised crime threat assessment, A corrupting influence: the infiltration and undermining of Europe's economy and society by organised crime.*

Evans, A. D., Xu, F., & Lee, K. (2011). When all signs point to you: lies told in the face of evidence. *Developmental Psychology*, *47*(1), 39.

Fazel, S., & Danesh, J. (2002). Serious mental disorder in 23 000 prisoners: a systematic review of 62 surveys. *The Lancet*, *359*(9306), 545–550.

Foerster, A., Wirth, R., Kunde, W., & Pfister, R. (2017). The dishonest mind set in sequence. *Psychological Research*, *81*(4), 878–899. https://doi.org/10.1007/s00426-016-0780-3

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology*, *2*, 59.

Freeman, J. B., Pauker, K., & Sanchez, D. T. (2016). A perceptual pathway to bias: Interracial exposure reduces abrupt shifts in real-time race perception that predict mixed-race bias. *Psychological Science*, *27*(4), 502–517.

Frone, M. R. (2013). *Alcohol and illicit drug use in the workforce and workplace.* American Psychological Association.

Frontex. (2021). *Risk Analysis for 2021*.

Fullam, R. S., McKie, S., & Dolan, M. C. (2009). Psychopathic traits and deception: functional magnetic resonance imaging study. *The British Journal of Psychiatry*, *194*(3), 229–235.

Ganis, G., & Keenan, J. P. (2009). The cognitive neuroscience of deception. *Social Neuroscience*, *4*(6), 465–472.

Gawrylowicz, J., Fairlamb, S., Tantot, E., Qureshi, Z., Redha, A., & Ridley, A. M. (2016). Does Practice Make the Perfect Liar? The Effect of Rehearsal and Increased Cognitive Load on Cues to Deception. *Applied Cognitive Psychology*, *30*(2), 250–259. https://doi.org/10.1002/acp.3199

Gilovich, T., Savitsky, K., & Medvec, V. H. (1998). The illusion of transparency: Biased assessments of others' ability to read one's emotional states. *Journal of Personality and Social Psychology*, *75*(2), 332–346. https://doi.org/10.1037/0022-3514.75.2.332

Giot, R., El-Abed, M., & Rosenberger, C. (2009). GREYC keystroke: A benchmark for keystroke dynamics biometric systems. *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, 1–6. https://doi.org/10.1109/BTAS.2009.5339051

Gombos, V. A. (2006). The Cognition of Deception: The Role of Executive Processes in Producing Lies. *Genetic, Social, and General Psychology Monographs*, *132*(3), 197–214. https://doi.org/10.3200/MONO.132.3.197-214

Gordon, G. R., Willox, N. A., Rebovich, D. J., Regan, T. M., & Gordon, J. B. (2003). *Identity fraud: A critical national and global threat*. Economic Crime Institute Utica, NY.

Goto, R., & Hakoda, Y. (2020). Effects of cognitive load during lying and personal characteristics on attention. *Journal of Human Environmental Studies*, *18*(2), 99–103. https://doi.org/10.4189/shes.18.99

Green, P., Allen, L. M., & Astner, K. (1996). The Word Memory Test: A user's guide to the oral and computer-administered forms, US Version 1.1. *Durham, NC: CogniSyst*.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*.

Gregg, A. P. (2007). When vying reveals lying: The timed antagonistic response alethiometer. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *21*(5), 621–647.

Grimes, G. M., Jenkins, J. L., & Valacich, J. S. (2013). Assessing credibility by monitoring changes in typing behavior: the keystrokes dynamics deception detection model. *Hawaii International Conference on System Sciences, Deception Detection Symposium*.

Guido, G., Peluso, A. M., Capestro, M., & Miglietta, M. (2015). An Italian version of the 10-item Big Five Inventory: An application to hedonic and utilitarian shopping values. *Personality and Individual Differences*, *76*, 135–140.

Hadar, A. A., Makris, S., & Yarrow, K. (2012). The truth-telling motor cortex: Response competition in M1 discloses deceptive behaviour. *Biological Psychology*, *89*(2), 495–502. https://doi.org/10.1016/j.biopsycho.2011.12.019

Hart, C. M., Ritchie, T. D., Hepper, E. G., & Gebauer, J. E. (2015). The balanced inventory of desirable responding short form (BIDR-16). *Sage Open*, *5*(4), 2158244015621113.

Hartwig, M., Anders Granhag, P., & Strömwall, L. A. (2007). Guilty and innocent suspects' strategies during police interrogations. *Psychology, Crime & Law*, *13*(2), 213–227. https://doi.org/10.1080/10683160600750264

Hartwig, M., Voss, J. A., & Wallace, D. B. (2015). Detecting Lies in the Financial Industry: A Survey of Investment Professionals' Beliefs. *Journal of Behavioral Finance*, *16*(2), 173–182. https://doi.org/10.1080/15427560.2015.1034862

Hibbeln, M., Jenkins, J. L., Schneider, C., Valacich, J. S., & Weinmann, M. (2014). Investigating the effect of insurance fraud on mouse usage in human-computer interactions. *35th International Conference on Information Systems: Building a Better World Through Information Systems, ICIS 2014*.

Hiscock, M., & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology*, *11*(6), 967–974.

Horvath, F. (1982). Detecting deception: the promise and the reality of voice stress analysis. *Journal of Forensic Sciences*, *27*(2), 340–351.

Hurst, A. F. (1940). Treatment of War Hysteria. *British Medical Journal*, *1*(4141), 828.

Ingegnieros, J. (1904). La Simulazione Della Pazzia in Rapporto Alla Criminologia: La Medicina Legale E La Clinica Psichiatrica. *Torino: Fratelli Bocca Editori*.

Inman, T. H., Vickery, C. D., Berry, D. T. R., Lamb, D. G., Edwards, C. L., & Smith, G. T. (1998). Development and initial validation of a new procedure for evaluating adequacy of effort given during neuropsychological testing: The letter memory test. *Psychological Assessment*, *10*(2), 128.

Iverson, G. L., Franzen, M. D., & McCracken, L. M. (1991). Evaluation of an objective assessment technique for the detection of malingered memory deficits. *Law and Human Behavior*, *15*(6), 667–676.

jamovi (Version 2.3) [Computer Software]. (2023). *The jamovi project* (Version 2.3).

Jersild, A. T. (1927). Mental set and shift. *Archives of Psychology*.

Johnson, R., Barnhardt, J., & Zhu, J. (2004). The contribution of executive processes to deceptive responding. *Neuropsychologia*, *42*(7), 878–901. https://doi.org/10.1016/j.neuropsychologia.2003.12.005

Jonason, P. K., & Webster, G. D. (2010). The dirty dozen: a concise measure of the dark triad. *Psychological Assessment*, *22*(2), 420.

Jones, D. N., & Paulhus, D. L. (2014). Introducing the short dark triad (SD3) a brief measure of dark personality traits. *Assessment*, *21*(1), 28–41.

Kapur, N. (1994). The coin-in-the-hand test: a new" bed-side" test for the detection of malingering in patients with suspected memory disorder. *Journal of Neurology, Neurosurgery, and Psychiatry*, *57*(3), 385.

Kassin, S. M., & Gudjonsson, G. H. (2004). The psychology of confessions: A review of the literature and issues. *Psychological Science in the Public Interest*, *5*(2), 33–67.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, *136*(5), 849–874. https://doi.org/10.1037/a0019842

Koop, G. J., & Johnson, J. G. (2011). Response dynamics: A new window on the decision process. *Judgment and Decision Making*, *6*(8), 750–758.

Krueger, R. F., Derringer, J., Markon, K. E., Watson, D., & Skodol, A. E. (2013). Personality Inventory for DSM-5. *Psychiatry Research*.

Kugelmass, S., & Lieblich, I. (1966). Effects of realistic stress and procedural interference in experimental lie detection. *Journal of Applied Psychology*, *50*(3), 211.

Kühn, C. G. (1830). Claudi Galeni opera omnia. *Lipsiae*, *XIX*.

Ladavas, E., & Berti, A. (2002). *Neuropsicologia* (il Mulino, Ed.).

Lancaster, G. L. J., Vrij, A., Hope, L., & Waller, B. (2013). Sorting the Liars from the Truth Tellers: The Benefits of Asking Unanticipated Questions on Lie Detection. *Applied Cognitive Psychology*, *27*(1), 107–114. https://doi.org/10.1002/acp.2879

Langleben, D. D., Schroeder, L., Maldjian, J. A., Gur, R. C., McDonald, S., Ragland, J. D., O'Brien, C. P., & Childress, A. R. (2002). Brain Activity during Simulated Deception: An Event-Related Functional Magnetic Resonance Study. *NeuroImage*, *15*(3), 727–732. https://doi.org/10.1006/nimg.2001.1003

Larrabee, G. J. (1990). Cautions in the use of neuropsychological evaluation in legal settings. *Neuropsychology*, *4*(4), 239.

Leins, D. A., Fisher, R. P., & Ross, S. J. (2013). Exploring liars' strategies for creating deceptive reports. *Legal and Criminological Psychology*, *18*(1), 141–151. https://doi.org/10.1111/j.2044-8333.2011.02041.x

Leins, D., Fisher, R. P., Vrij, A., Leal, S., & Mann, S. (2011). Using sketch drawing to induce inconsistency in liars. *Legal and Criminological Psychology*, *16*(2), 253–265. https://doi.org/10.1348/135532510X501775

Lezak, M. D., Howieson, D. B., Loring, D. W., & Fischer, J. S. (2004). *Neuropsychological assessment*. Oxford University Press, USA.

Littlepage, G. E., & Pineault, M. A. (1985). Detection of Deception of Planned and Spontaneous Communications. *The Journal of Social Psychology*, *125*(2), 195–201. https://doi.org/10.1080/00224545.1985.9922872

Liu, M., & Wronski, L. (2018). Trap questions in online surveys: Results from three web survey experiments. *International Journal of Market Research*, *60*(1), 32–49.

Loconte, R., Russo, R., Capuozzo, P., Pietrini, P., & Sartori, G. (2023). Verbal lie detection using Large Language Models. *Scientific Reports*, *13*(1), 22849.

Lombroso, C., & Leggiardi-Laura, C. (1905). *La perizia psichiatrico-legale: coi metodi per eseguirla e la casuistica penale classificata antropologicamente (con 4 tav. e 48 fig. interc. nel testo)* (Vol. 37). Fratelli Bocca.

Löwe, B., Unützer, J., Callahan, C. M., Perkins, A. J., & Kroenke, K. (2004). Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical Care*, 1194–1201.

Luan Phan, K., Magalhaes, A., Ziemlewicz, T. J., Fitzgerald, D. A., Green, C., & Smith, W. (2005). Neural correlates of telling lies. *Academic Radiology*, *12*(2), 164–172. https://doi.org/10.1016/j.acra.2004.11.023

Luke, T. J., Giolla, E. Mac, Memon, A., Landström, S., Granhag, P. A., & Kassin, S. (2023). What have we learned about cues to deception? A survey of expert opinions. *Psychology, Crime & Law*, 1–20. https://doi.org/10.1080/1068316X.2023.2292504

Luyten, P., Mayes, L. C., Nijssens, L., & Fonagy, P. (2017). The parental reflective functioning questionnaire: Development and preliminary validation. *PloS One*, *12*(5), e0176218.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology*, *43*(6), 385.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163.

Maleck, M., & Gross, T. (2023). AnswerTruthDetector: a combined cognitive load approach for separating truthful from deceptive answers in computer-administered questionnaires. *I-Com*, *22*(3), 241–251.

Maricuțoiu, L. P., & Sârbescu, P. (2016). The relationship between faking and response latencies. *European Journal of Psychological Assessment*.

Masip, J., & Herrero, C. (2015). New approaches in deception detection I. Background and theoretical framework. *Papeles Del Psicólogo*, *36*(2), 83–95.

Mazza, C., Monaro, M., Burla, F., Colasanti, M., Orrù, G., Ferracuti, S., & Roma, P. (2020). Use of mouse-tracking software to detect faking-good behavior on personality questionnaires: an explorative study. *Scientific Reports*, *10*(1), 4835. https://doi.org/10.1038/s41598-020-61636-5

Mechelli, A., Sartori, G., Orlandi, P., & Price, C. J. (2006). Semantic relevance explains category effects in medial fusiform gyri. *Neuroimage*, *30*(3), 992–1002.

Meixner, J. B., & Rosenfeld, J. P. (2011). A mock terrorism application of the P300-based concealed information test. *Psychophysiology*, *48*(2), 149–154.

Merckelbach, H., Smeets, T., & Jelicic, M. (2009). Experimental simulation: Type of malingering scenario makes a difference. *The Journal of Forensic Psychiatry & Psychology*, *20*(3), 378–386.

Michael Connelly. (2008). *The Brass Verdict*. Grand Central Publishing.

Mitchell, T. M. (1997). Machine Learning. *McGraw-Hill Education*, *(1st ed.)*.

Mittenberg, W., Patton, C., Canyock, E. M., & Condit, D. C. (2002). Base rates of malingering and symptom exeggeration. *Journal of Clinical and Experimental Neuropsychology*, *24*(8), 1094–1102.

Monaro, M., Businaro, M., Spolaor, R., Li, Q. Q., Conti, M., Gamberini, L., & Sartori, G. (2019). *The Online Identity Detection via Keyboard Dynamics* (pp. 342–357). https://doi.org/10.1007/978-3-030-02683-7_24

Monaro, M., Fugazza, F. I., Gamberini, L., & Sartori, G. (2017). How human-mouse interaction can accurately detect faked responses about identity. *Symbiotic Interaction*, *115*, 91.

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., & Sartori, G. (2018). Covert lie detection using keyboard dynamics. *Scientific Reports*, *8*(1), 1976. https://doi.org/10.1038/s41598-018-20462-6

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PLOS ONE*, *12*(5), e0177851. https://doi.org/10.1371/journal.pone.0177851

Monaro, M., Gamberini, L., Zecchinato, F., & Sartori, G. (2018). False identity detection using complex sentences. *Frontiers in Psychology*, *9*, 283.

Monaro, M., Toncini, A., Ferracuti, S., Tessari, G., Vaccaro, M. G., De Fazio, P., Pigato, G., Meneghel, T., Scarpazza, C., & Sartori, G. (2018). The detection of malingering: a new tool to identify made-up depression. *Frontiers in Psychiatry*, *9*, 249.

Monaro, M., Zampieri, I., Sartori, G., Pietrini, P., & Orrù, G. (2021). The detection of faked identity using unexpected questions and choice reaction times. *Psychological Research*, *85*(6), 2474–2482. https://doi.org/10.1007/s00426-020-01410-4

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/10.1016/S1364-6613(03)00028-7

Morey, L. C., & Staff, P. (1991). *Personality Assessment Inventory: Professional Manual.* (F. Odessa, Ed.).

Niemeyer, L. M., Grosz, M. P., Zimmermann, J., & Back, M. D. (2022). Assessing maladaptive personality in the forensic context: Development and validation of the Personality Inventory for DSM-5 Forensic Faceted Brief Form (PID-5-FFBF). *Journal of Personality Assessment*, *104*(1), 30–43.

Ogawa, T., Todoriki, N., & Tsuneoka, M. (2021). The Use of Pupillometry in Autobiographical Implicit Association Test. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.729897

Ones, D. S., & Viswesvaran, C. (2001). *Personality at work: Criterion-focused occupational personality scales used in personnel selection.*

Orrù, G., Mazza, C., Monaro, M., Ferracuti, S., Sartori, G., & Roma, P. (2021). The development of a short version of the SIMS using machine learning to detect feigning in forensic assessment. *Psychological Injury and Law*, *14*(1), 46–57.

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology*, *10*, 2970.

O'Sullivan, M., Frank, M. G., Hurley, C. M., & Tiwana, J. (2009). Police lie detection accuracy: The effect of lie scenario. *Law and Human Behavior*, *33*(6), 530–538. https://doi.org/10.1007/s10979-008-9166-4

Otasowie, O. (2020). Application of machine learning in deception detection. *Intelligent Computing: Proceedings of the 2020 Computing Conference, Volume 3*, 61–76.

Palena, N., Caso, L., Vrij, A., & Orthey, R. (2018). Detecting deception through small talk and comparable truth baselines. *Journal of Investigative Psychology and Offender Profiling*, *15*(2), 124–132. https://doi.org/10.1002/jip.1495

Pankratz, L. (1983). A new technique for the assessment and modification of feigned memory deficit. *Perceptual and Motor Skills*, *57*(2), 367–372.

Parkhouse, T., & Ormerod, T. C. (2018). Unanticipated questions can yield unanticipated outcomes in investigative interviews. *PLOS ONE*, *13*(12), e0208751. https://doi.org/10.1371/journal.pone.0208751

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244. https://doi.org/10.1037/0033-2909.116.2.220

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

Pennisi, P. (1998). e Sarlo. *M.: Indici Elettrofisiologici in Psicologia (Cleup Editrice), Padova*.

Porter, S., & ten Brinke, L. (2010). The truth about lies: What works in detecting high-stakes deception? *Legal and Criminological Psychology*, *15*(1), 57–75. https://doi.org/10.1348/135532509X433151

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:2010.16061*.

Quétard, B., Quinton, J.-C., Colomb, M., Pezzulo, G., Barca, L., Izaute, M., Appadoo, O. K., & Mermillod, M. (2015). Combined effects of expectations and visual uncertainty upon detection and identification of a target in the fog. *Cognitive Processing*, *16*(1), 343–348.

Ramachandran, S., Rosenthal, M., Young, J., Holmes, E., & Bentley, J. P. (2019). Subtle scales: An avenue for identification of prescription drug abuse. *Research in Social and Administrative Pharmacy*, *15*(8), 936–942.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212.

Rees, L. M., Tombaugh, T. N., Gansler, D. A., & Moczynski, N. P. (1998). Five validation experiments of the Test of Memory Malingering (TOMM). *Psychological Assessment*, *10*(1), 10.

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique.(Les problems.). *Archives de Psychologie*.

Rey, A. (1958). *L'examen clinique en psychologie.*

Rey, A. (1964). L'examen clinique en psychologie, Paris: Presses Universitaires de France, 1964. *Chemotherapy and Objective Cognitive Functioning*, *95*.

Rheem, H., Verma, V., & Becker, D. V. (2018). Use of mouse-tracking method to measure cognitive load. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 1982–1986.

Rogers, R. (1997). *Researching dissimulation.*

Rogers, R., Bagby, R. M., & Dickens, S. E. (1992). Structured Manual., interview of reported symptoms (SIRS) and professional. *Odessa, FL: Psychological Assessment Resources.*

Rogers, R., Harrell, E. H., & Liff, C. D. (1993). Feigning neuropsychological impairment: A critical review of methodological and clinical considerations. *Clinical Psychology Review*, *13*(3), 255–274.

Roncoroni, L. (1987). Contributo allo studio della simulazione. *Archivio Di Psichiatria*, *fasc. IV*.

Rosenfeld, J. P., & Bessinger, G. T. (1990). Feedback-evoked P300 responses in lie detection. *Psychophysiology*, *27*, S60.

Rotter, M., Way, B., Steinbacher, M., Sawyer, D., & Smith, H. (2002). Personality disorders in prison: Aren't they all antisocial? *Psychiatric Quarterly*, *73*, 337–349.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological Science*, *19*(8), 772–780.

Sartori, G., & Lombardi, L. (2004). Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, *16*(3), 439–452.

Sartori, G., Orrù, G., & Zangrossi, A. (2016). Detection of malingering in personal injury and damage ascertainment. In *Personal injury and damage ascertainment under civil law* (pp. 547–558). Springer.

Sartori, G., Zangrossi, A., & Monaro, M. (2018). *Deception detection with behavioral methods: The Autobiographical Implicit Association Test, concealed information test-reaction time, mouse dynamics, and keystroke dynamics.*

Sartori, G., Zangrossi, A., Orrù, G., & Monaro, M. (2017). Detection of malingering in psychic damage ascertainment. *P5 Medicine and Justice: Innovation, Unitariness and Evidence*, 330–341.

Saxe, L. (1991). Science and the CQT polygraph. *Integrative Physiological and Behavioral Science*, *26*(3), 223–231.

Shahandeh, B., & Caborn, J. (2003). Ethical issues in workplace drug testing in Europe. *Geneva: International Labour Office*.

Sheridan, M. R., & Flowers, K. A. (2010). Reaction Times and Deception - the Lying Constant. *International Journal of Psychological Studies*, *2*(2). https://doi.org/10.5539/ijps.v2n2p41

Shevlin, M., Hyland, P., Ben-Ezra, M., Karatzias, T., Cloitre, M., Vallières, F., Bachem, R., & Maercker, A. (2020). Measuring ICD-11 adjustment disorder: The development and initial validation of the International Adjustment Disorder Questionnaire. *Acta Psychiatrica Scandinavica*, *141*(3), 265–274.

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(6), 679–692.

Skinner, H. A. (1982). The drug abuse screening test. *Addictive Behaviors*, *7*(4), 363–371.

Slick, D. J., Hopp, G., Strauss, E., & Spellacy, F. J. (1996). Victoria Symptom Validity Test: Efficiency for detecting feigned memory impairment and relationship to neuropsychological tests and MMPI-2 validity scales. *Journal of Clinical and Experimental Neuropsychology*, *18*(6), 911–922.

Smith, G. P., & Burger, G. K. (1997). *Detection of Malingering: Validation of the Structured Inventory of Malingered Symptomatology (SIMS)*.

Sooniste, T., Granhag, P. A., Strömwall, L. A., & Vrij, A. (2015). Statements about true and false intentions: Using the Cognitive Interview to magnify the differences. *Scandinavian Journal of Psychology*, *56*(4), 371–378. https://doi.org/10.1111/sjop.12216

Sooniste, T., Granhag, P. A., Strömwall, L. A., & Vrij, A. (2016). Discriminating between true and false intent among small cells of suspects. *Legal and Criminological Psychology*, *21*(2), 344–357. https://doi.org/10.1111/lcrp.12063

Spector, A., & Biederman, I. (1976). Mental set and mental shift revisited. *The American Journal of Psychology*, 669–679.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport*, *12*(13), 2849–2853. https://doi.org/10.1097/00001756-200109170-00019

Stein, M. B., & Sareen, J. (2015). Generalized anxiety disorder. *New England Journal of Medicine*, *373*(21), 2059–2068.

Stracciari, A., Bianchi, A., & Sartori, G. (2010). *Neuropsicologia forense* (Il mulino, Ed.).

Sun, T., Zhang, B., Cao, M., & Drasgow, F. (2022). Faking detection improved: Adopting a Likert item response process tree model. *Organizational Research Methods*, *25*(3), 490–512.

Theodore Millon, S. G. C. M. (2015). *Millon Clinical Multiaxial Inventory-IV (MCMI-IV)*. Bloomington, MN: NCS Pearson, Inc.

Tombaugh, T. N. (1996). Test of Memory Malingering (TOMM). Multi-Health Systems. *New York*.

Tombaugh, T. N. (1997). The Test of Memory Malingering (TOMM): Normative data from cognitively intact and cognitively impaired individuals. *Psychological Assessment*, *9*(3), 260.

Tun, P. A., & Lachman, M. E. (2008). Age differences in reaction time and attention in a national telephone sample of adults: education, sex, and task complexity matter. *Developmental Psychology*, *44*(5), 1421.

Utz, S. (2005). Types of Deception and Underlying Motivation. *Social Science Computer Review*, *23*(1), 49–56. https://doi.org/10.1177/0894439304271534

van der Walt, E., Eloff, J. H. P., & Grobler, J. (2018). Cyber-security: Identity deception detection on social media platforms. *Computers & Security*, *78*, 76–89. https://doi.org/10.1016/j.cose.2018.05.015

van Impelen, A., Merckelbach, H., Jelicic, M., & Merten, T. (2014). The Structured Inventory of Malingered Symptomatology (SIMS): A systematic review and meta-analysis. *The Clinical Neuropsychologist*, *28*(8), 1336–1365.

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin*, *136*(4), 601–626. https://doi.org/10.1037/a0019791

Vendemia, J. M. C., Buzan, R. F., & Green, E. P. (2005). Practice Effects, Workload, and Reaction Time in Deception. *The American Journal of Psychology*, *118*(3), 413–429. https://doi.org/10.2307/30039073

Verigin, B. L., Meijer, E. H., & Vrij, A. (2021). A within-statement baseline comparison for detecting lies. *Psychiatry, Psychology and Law*, *28*(1), 94–103. https://doi.org/10.1080/13218719.2020.1767712

Verschuere, B., Ben-Shakhar, G., & Meijer, E. (2011). *Memory detection: Theory and application of the Concealed Information Test*. Cambridge University Press.

Verschuere, B., Spruyt, A., Meijer, E. H., & Otgaar, H. (2011). The ease of lying. *Consciousness and Cognition*, *20*(3), 908–911.

Viglione, D. J., Giromini, L., & Landis, P. (2017). The development of the Inventory of Problems–29: A brief self-administered measure for discriminating bona fide from feigned psychiatric and cognitive complaints. *Journal of Personality Assessment*, *99*(5), 534–544.

Vrij, A. (2000). *Detecting lies and deceit: The psychology of lying and implications for professional practice*. Wiley.

Vrij, A. (2004). Why professionals fail to catch liars and how they can improve. *Legal and Criminological Psychology*, *9*(2), 159–181.

Vrij, A. (2016). Baselining as a Lie Detection Method. *Applied Cognitive Psychology*, *30*(6), 1112–1119. https://doi.org/10.1002/acp.3288

Vrij, A., Akehurst, L., & Knight, S. (2006). Police officers', social workers', teachers' and the general public's beliefs about deception in children, adolescents and adults. *Legal and Criminological Psychology*, *11*(2), 297–312. https://doi.org/10.1348/135532505X60816

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2006). Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, *10*(4), 141–142. https://doi.org/10.1016/j.tics.2006.02.003

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1–21.

Vrij, A., & Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked. *Journal of Applied Research in Memory and Cognition*, *1*(2), 110–117.

Vrij, A., Granhag, P. A., Ashkenazi, T., Ganis, G., Leal, S., & Fisher, R. P. (2022). Verbal lie detection: its past, present and future. *Brain Sciences*, *12*(12), 1644.

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Outsmarting the Liars: Toward a Cognitive Lie Detection Approach. *Current Directions in Psychological Science*, *20*(1), 28–32. https://doi.org/10.1177/0963721410391245

Vrij, A., Granhag, P. A., & Porter, S. (2010). Pitfalls and Opportunities in Nonverbal and Verbal Lie Detection. *Psychological Science in the Public Interest*, *11*(3), 89–121. https://doi.org/10.1177/1529100610390861

Vrij, A., & Heaven, S. (1999). Vocal and verbal indicators of deception as a function of lie complexity. *Psychology, Crime & Law*, *5*(3), 203–215. https://doi.org/10.1080/10683169908401767

Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R. P., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and Human Behavior*, *33*(2), 159–166. https://doi.org/10.1007/s10979-008-9143-y

Vrij, A., Leal, S., Mann, S., & Fisher, R. (2012). Imposing cognitive load to elicit cues to deceit: inducing the reverse order technique naturally. *Psychology, Crime & Law*, *18*(6), 579–594. https://doi.org/10.1080/1068316X.2010.515987

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior*, *32*(3), 253–265. https://doi.org/10.1007/s10979-007-9103-y

Vrij, A., Mann, S., Leal, S., & Fisher, R. (2010). 'Look into my eyes': can an instruction to maintain eye contact facilitate lie detection? *Psychology, Crime & Law*, *16*(4), 327–348. https://doi.org/10.1080/10683160902740633

Walczyk, J. J., Griffith, D. A., Yates, R., Visconte, S. R., Simoneaux, B., & Harris, L. L. (2012). Lie detection by inducing cognitive load: Eye movements and other cues to the false answers of "witnesses" to crimes. *Criminal Justice and Behavior*, *39*(7), 887–909.

Walczyk, J. J., Igou, F. P., Dixon, A. P., & Tcholakian, T. (2013). Advancing Lie Detection by Inducing Cognitive Load on Liars: A Review of Relevant Theories and Techniques Guided by Lessons from Polygraph-Based Approaches. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00014

Walczyk, J. J., Mahoney, K. T., Doverspike, D., & Griffith-Ross, D. A. (2009). Cognitive Lie Detection: Response Time and Consistency of Answers as Cues to Deception. *Journal of Business and Psychology*, *24*(1), 33–49. https://doi.org/10.1007/s10869-009-9090-8

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: response time as a cue to deception. *Applied Cognitive Psychology*, *17*(7), 755–774. https://doi.org/10.1002/acp.914

Wang, A. G., Atabakhsh, H., Petersen, T., & Chen, H. (2005). *Discovering Identity Problems: A Case Study* (pp. 368–373). https://doi.org/10.1007/11427995_30

Wang, G. A., Chen, H., Xu, J. J., & Atabakhsh, H. (2006). Automatically detecting criminal identity deception: an adaptive detection algorithm. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *36*(5), 988–999. https://doi.org/10.1109/TSMCA.2006.871799

Wang, G., Chen, H., & Atabakhsh, H. (2004). Automatically detecting deceptive criminal identities. *Communications of the ACM*, *47*(3), 70–76. https://doi.org/10.1145/971617.971618

Warmelink, L., Vrij, A., Mann, S., Jundi, S., & Granhag, P. A. (2012). The effect of question expectedness and experience on lying about intentions. *Acta Psychologica*, *141*(2), 178–183. https://doi.org/10.1016/j.actpsy.2012.07.011

Warmelink, L., Vrij, A., Mann, S., Leal, S., Forrester, D., & Fisher, R. P. (2011). Thermal imaging as a lie detection tool at airports. *Law and Human Behavior*, *35*, 40–48.

Weiss, D. S. (2007). The impact of event scale: revised. In *Cross-cultural assessment of psychological trauma and PTSD* (pp. 219–238). Springer.

Yan, T., & Tourangeau, R. (2008). Fast times and easy questions: The effects of age, experience and question complexity on web survey response times. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *22*(1), 51–68.

Yang, C., Niu, C., & Shen, X. (2022). The Master of Detecting Deception: Machine Learning. *Proceedings of Sixth International Congress on Information and Communication Technology: ICICT 2021, London, Volume 4*, 803–814.

Young, G. (2015). Malingering in forensic disability-related assessments: Prevalence 15±15%. *Psychological Injury and Law*, *8*, 188–199.

Zago, S., Difonzo, T. (2018). La simulazione di stati patologici mentali: aspetti generali e nel contesto penitenziario. In M. Materia Medica (Ed.), *Dipendenze patologiche in area penale. Interventi multidisciplinari di cura e contesti legislativi.*

Zago, S., Piacquadio, E., Monaro, M., Orrù, G., Sampaolo, E., Difonzo, T., Toncini, A., & Heinzl, E. (2019). The Detection of Malingered Amnesia: An Approach Involving Multiple Strategies in a Mock Crime. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00424

Zago, S., Sartori, G., & Scarlat, G. (2004). Malingering and retrograde amnesia: The historic case of the Collegno amnesic. *Cortex*, *40*(3), 519–532.

Zangrossi, A., Cona, G., Celli, M., Zorzi, M., & Corbetta, M. (2021). Visual exploration dynamics are low-dimensional and driven by intrinsic factors. *Communications Biology*, *4*(1), 1100. https://doi.org/10.1038/s42003-021-02608-x

Zangrossi, A., Gatto, L. C., Lanfranchi, V., Scarpazza, C., Celli, M., & Sartori, G. (2024). Autobiographical Implicit Association Test and eye movements: fixations topography enables detection of autobiographical memories. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1268256

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). *Verbal and Nonverbal Communication of Deception* (pp. 1–59). https://doi.org/10.1016/S0065-2601(08)60369-X