Detecting Concealed Information on a Large Scale: Possibilities and Problems

Kleinberg, B.; van der Toolen, Y.; Arntz, A.; Verschuere, B.

DOI
10.1016/B978-0-12-812729-2.00016-1
Publication date
2018
Document Version
Final published version
Published in
Detecting Concealed Information and Deception

CHAPTER 16

Detecting Concealed Information on a Large Scale: Possibilities and Problems

Bennett Kleinberg, Yaloe van der Toolen, Arnoud Arntz, Bruno Verschuere
University of Amsterdam, Amsterdam, The Netherlands

In 2014, one of the main nuclear power stations in Belgium, "Doel 4," was sabotaged (DeClercq, 2014). Investigators quickly realized that the perpetrator must have opened a valve that caused overheating in one of the turbines. The search for the perpetrator was, therefore, narrowed down to those engineers, technicians, and workers that were present at the time of the incident. Forty employees who had been in the machine chamber at the time of the sabotage were suspected, denied access to the nuclear plant for months, and were asked to take a polygraph test. This request raised such indignation that the vast majority of the employees refused to take the polygraph test. To date, no arrests have been made nor has the case been solved. This is an example of a setting in which a larger number of examinees need to be assessed, and in which some of the standard methods of deception detection are limited.

With a few exceptions (e.g., Meijer, Bente, Ben-Shakhar, & Schumacher, 2013), deception research has largely targeted the detection of deception in individual cases. From an applied perspective, however, there is an increasing need to detect concealed information at a larger scale, be it security clearance of staff working with sensitive information, border security (Honts & Hartwig, 2014), insurance claims (e.g., Harvey, Vrij, Nahari, & Ludwig, 2017), terrorism prevention screening (see Kleinberg, Arntz, & Verschuere, 2017), or employee theft. In this chapter, we review the dominant methods of detecting concealed information as to their potential of being suitable for large-scale purposes. For the remainder of this chapter, we do not restrict concealed-information detection to a specific test (i.e., the guilty knowledge test or Concealed Information Test [CIT];

Detecting Concealed Information and Deception
ISBN 978-0-12-812729-2
https://doi.org/10.1016/B978-0-12-812729-2.00016-1

Lykken, 1959; Verschuere, Ben-Shakhar, & Meijer, 2011) nor specific instances of concealed information.

The aim of this chapter is to outline and discuss the methods that facilitate large-scale deception detection. First, we provide an overview of those methods from a broad spectrum of deception detection that could be applied on a large scale. Each method will be outlined, evaluated, and discussed. We also propose a framework that might help to establish the potential of a deception detection method when applied to large groups. Specifically, we focus on (1) the theoretical foundations of the methods, (2) the possibility of using the method in quick procedures, and (3) the flexibility of the method for various contexts. For all criteria, we provide a brief evaluation of the feasibility within 5 years to avoid speculation about long-term technological developments. We conclude this chapter with an outlook on the future of large-scale deception detection.

METHODS FOR THE DETECTION OF CONCEALED INFORMATION

This section provides an outline of various methods that could be used to detect concealed information and discriminate between truthful and deceptive people. We start with an overview of methods that are already applied in large-scale settings, such as behavioral observation, and the analysis of paraverbal speech properties. Next, we describe approaches that are currently not yet implemented but could be applied in large-scale screening procedures within the next 5 years: thermal imaging, reaction time tests, and verbal methods. It should be noted that we have made a selection for this chapter and that several other methods to detecting deception exist but are not part of this chapter, for instance, because they are not suitable for large-scale applications. Among these methods are brain imaging (e.g., Ganis, Kosslyn, Stose, Thompson, & Yurgelun-Todd, 2003; Ganis, Morris, & Kosslyn, 2009; Ganis, Rosenfeld, Meixner, Kievit, & Schendan, 2011), and electroencephalography (EEG) (e.g., Rosenfeld, Hu, Labkovsky, Meixner, & Winograd, 2013; Rosenfeld, Soskins, Bosh, & Ryan, 2004).

Screening Passengers by Observation Techniques

Outline

As a response to the 2001 terrorist attacks, the US Transportation Security Administration (TSA) started with the development of the "screening passengers by observation techniques" (SPOT) program. The outline presented here uses SPOT as one example from a family of related methods that use suspicious behavior as an indicator for deception. Although we focus on SPOT, the general assumptions and shortcomings apply to a range of suspicious-behavior detection methods.

SPOT is intended to identify passengers who may pose a threat to aviation security (Government Accountability Office, 2013) and was developed based on Paul Ekman's work on facial expressions and deception (Transportation Security Administration, 2009). SPOT assumes that someone displaying suspicious behavior is potentially hiding their real purpose at the airport (e.g., by posing as a passenger) or is flying for an illegal purpose (e.g., to carry out an attack at their destination airport). The development of that TSA program culminated in its United States-wide implementation in 2007, and the program was expanded over the following years (albeit under a different name, *Behavior Detection and Analysis Program*; Government Accountability Office, 2016). One of the core assumptions of SPOT is that security-threatening passengers will behave differently compared to ordinary travelers and display suspicious behavior. That behavior is claimed to result from experiencing emotional states such as stress or fear (see also Honts & Hartwig, 2014). Central to SPOT are so-called Behavior Detection Officers (BDOs): employees trained in identifying passengers that show behaviors allegedly indicative of deception (Government Accountability Office, 2015). BDOs initially observe people waiting in line at the airport to establish a general baseline and then assess the passengers' behavioral cues and appearance. They also engage in "walking the line," a process during which they initiate brief conversations with the waiting passengers to assess them for any suspicious behaviors (Department of Homeland Security, 2013).

In case a passenger does display suspicious behavior and exceeds a certain threshold indicated by SPOT, that person is referred for further screening.

This additional screening process includes a search of the passenger and his/her personal luggage, as well as an examination of identification and travel documents (Department of Homeland Security, 2013). For long, it remained unclear exactly which behaviors the SPOT method regarded as suspicious. Recently, however, a 2009-version of a SPOT Referral Report was published online by the news site *The Intercept* (Winter & Currier, 2015). That document revealed the behavioral cues enlisted by the program, including being pale, sweating, excessive eye blinking, trembling, and whistling during the screening process.[1]

Evaluation

Although the program is widely implemented—with over 3000 BDOs working in the aviation sector in 2013 (Department of Homeland Security, 2013)— there are significant concerns about the effectiveness and efficacy of SPOT. For instance, although BDOs referred over 200,000 passengers for a secondary screening between 2006 and 2009, less than 1% of these referrals led to an actual arrest, of which the majority was for reasons unrelated to terrorism (Weinberger, 2010). This points to a high number of false positives (here: larger than 99%). SPOT advocates would argue that false-positive test outcomes in an airport setting are less severe than those in criminal investigations because the adverse consequences are much smaller (i.e., referral for further questioning rather than a criminal conviction). Moreover, SPOT is used primarily as a screening method so that other deception detection methods such as information-gathering interviewing could be combined with it and applied at a later stage. However, the costs of false positives are high both from the individual passenger's perspective (e.g., missing a flight, being treated like a criminal) as well as from the airport's perspective (e.g., spending time and money for extensive security procedures on ordinary passengers). One could argue that a high false-positive rate (i.e., low specificity; identifying those who are regular passengers) would be outweighed by an exceptionally high sensitivity (i.e., identifying those who are in fact planning a malicious act). There are currently no reports of SPOT identifying terrorists. To the contrary, there are several known cases of terrorists traveling undetected through airports

[1] The TSA stated in 2015 that the protocol was about to be improved, which could mean that the published report is outdated (Transportation Security Administration, 2015). Nevertheless, the TSA also mentions that it continues to use behavioral cues to identify suspicious passengers. The cues outlined in the Intercept document (Winter & Currier, 2015) might therefore still provide the only publicly accessible insight into SPOT.

where the SPOT program was implemented (Government Accountability Office, 2010; Perry & Gilbey, 2011).

Although there are no peer-reviewed reports on the empirical evaluation of behavioral detection methods or its working mechanisms (for a published example of random classification accuracy of a suspicious-signs method, see Wijn, van der Kleij, Kallen, Stekkinger, & de Vries, 2017), the program has been rolled out across multiple countries. SPOT uses behavioral cues to detect deception, yet many of the cues have not been empirically evaluated, and others have typically shown no or weak associations with deception (e.g., Bond & DePaulo, 2006; Vrij, 2008). Behaviors such as avoiding eye contact, fidgeting, or fast eye blinking are all listed as being used as indices of deception, whereas metaanalytical research has shown these behaviors bear no relationship to deception (DePaulo et al., 2003). Another behavioral cue that is being regarded as a suspicious is nervousness. Although there are indeed indications that being nervous is related to deception (DePaulo et al., 2003), the relationship is weak at best, and it could be argued that, in an airport setting, displaying nervousness is rather common. Passengers may experience fear of flying, they may feel uncomfortable with the screening procedures, or they might rush to catch their flight. In other words, the already-weak relationship between nervousness and deception can be expected to be weakened further in airport settings leading mainly to false-positive outcomes.

Finally, there is the question of actual implementation. It is uncertain whether human evaluators can be expected to assess large numbers of passengers on a vast list of cues in a short time (i.e., while casually talking to them in the waiting queue), and to effectively integrate that information to reach a decision. To date, no studies have investigated whether and how BDOs use the cues put forward by the SPOT method. Although BDOs are expected to draw objective conclusions based on behavioral indicators, former security agents have asserted that many of their colleagues used subjective judgments and that the program was vulnerable to racial profiling (Schmidt & Lichtblau, 2012). In May 2016, both academics and government officials pointed out that the TSA has not been able to show convincing proof of the SPOT program and suggested that funding should be reduced. In a reaction, the TSA announced that it would take action to optimize the program by conducting operational tests starting September 2016 (Government Accountability Office, 2016). After obtaining documents from the TSA under the Freedom of Information Act, the American Civil Liberties Union concluded that the foundations of SPOT

are "unscientific and unreliable" and that its validity is overstated in official government documents (American Civil Liberties Union, 2017).

Speech Analysis

Several tools try to detect deception by analyzing people's speech signal (e.g., the pitch and intensity). Within the field of forensic speech analysis, there are two dominant approaches: Computerized Voice Stress Analysis (CVSA) and Layered Voice Analysis (LVA). They share the underlying assumption that differences in voice signals of truthful-versus-deceptive statements are due to the different mental states of the truth tellers and liars who uttered them (Gamer, Rill, Vossel, & Gödert, 2006). Because the assumptions are the same and differences between the two versions marginal, this section will focus on Computerized Voice Stress Analysis.

Outline

The first generation of speech-analysis devices was developed in the 1970s under the name Computerized Voice Stress Analyzers (CVSAs, sometimes referred to as VSAs), or Psychological Stress Evaluators (PSEs). The companies that manufacture these tools state that, because of stress, people who are lying produce a different profile of tiny vibrations than truth tellers. Those vibrations, so-called microtremors, are produced by muscles in the throat or larynx (Horvath, 1982) and are inaudible to the human ear but would be detectable through specialist software. To be able to discriminate between truth tellers and liars, CVSA uses interviewing techniques that are based on a set of roughly 12 questions. Some of these questions function as a baseline measure for the level of stress experienced by an individual. They are in turn compared to responses to relevant questions that are related to the subject of interest, such as a crime (see Truth and Deception Technologies, 2009). In general, people are expected to answer each question with "yes" or "no."

Evaluation

Peer-reviewed studies systematically showed that CVSA lacks validity and does not exceed chance level in differentiating deceptive from truthful statements (e.g., Damphousse, Pointon, Upchurch, & Moore, 2007; Gamer et al., 2006; Hollien & Harnsberger, 2006). However, despite the lack of support, CVSA tools are widely used by several law enforcement agencies (see the website cited in Damphousse et al., 2007). The closest evidence to

any effects of microtremors stems from a metaanalysis (DePaulo et al., 2003) that found that liars often sound tenser and have a higher pitch than truth-tellers. The reported effects were, however, small and not a direct test of the microtremors postulated by CVSA. Moreover, there is no theoretical justification for the specific assumption that microtremors differ in truthful and deceptive statements. Olaf Lippold, a British physiologist, who is said to have discovered microtremors in human muscles in the 1970s, is often mentioned in CVSA manuals. It is noteworthy, however, that Lippold and colleagues never investigated the effect of psychological stress on microtremors, nor is there any proof that the arm-muscle microtremors studied by Lippold can also be found in throat or larynx muscles (Eriksson & Lacerda, 2008; Shipp & Izdebski, 1981). Speech analysis seems not fit for larger applications.

Thermal Imaging

Outline

Thermal-imaging technology aims to detect deception by measuring facial temperature with thermal cameras (Pavlidis & Levine, 2002). Deceptive individuals will experience more stress and anxiety compared to innocent people. These emotional differences would, in turn, result in measurable physiological differences (in facial-heating pattern; see following section) that discriminate between truth tellers and liars (Pavlidis, Levine, & Baukol, 2000).

Deceptive individuals afraid of being caught would show an increased sympathetic nervous system activity, related to a "fight or flight" response. That increased activity results in a redistribution of blood in the human body (Pavlidis & Levine, 2002) visible especially in the periorbital regions of the face. To facilitate rapid eye movement, blood flow to this region will increase, in turn causing a rise in the temperature around the eyes (Pavlidis, Eberhardt, & Levine, 2002b). Other research has suggested that the activation of the sympathetic nervous system due to deception also increases nose temperature (Panasiti et al., 2016). With advanced thermal cameras, it is possible to detect these alterations in blood flow, providing possible cues for deception.

It has been suggested that thermal imaging could serve as a tool in security screening processes in public settings such as airports (e.g., originally in Pavlidis, Eberhardt, & Levine, 2002a; Pavlidis et al., 2002b), although these initial claims were later relativized in an erratum (Pavlidis et al., 2002a).

It is proposed that thermal-imaging cameras could be installed at strategic checkpoints in the airport. Checkpoint agents, who already engage in asking travel-related questions to passengers, could then also measure whether passengers show an increase in temperature in the relevant facial areas (combined possibly with remote heart-rate measures) after being asked about their trip. This information could be used as additional data for deciding whether a passenger should be considered suspicious or not (Pavlidis & Levine, 2002).

Evaluation

A small set of studies have reported thermal imaging to successfully distinguish truthful from deceptive subjects (e.g., Panasiti et al., 2016; Park, Suk, Hwang, & Lee, 2013; Pavlidis et al., 2002b; Rajoub & Zwiggelaar, 2014; Warmelink et al., 2011). It should be noted, however, that thermal imaging did not always outperform human interviewers and that it is, therefore, not clear what the added value of that method is (Warmelink et al., 2011). In one study participants were asked to either conduct a mock crime (stealing a wallet from a lab) or an innocent act (sending an email from a lab; Park et al., 2013). After carrying out their activity, participants were interviewed about whether they had committed the crime, using a Concealed Information Test protocol in which expected knowledge concerning the mock crime was tested. While answering the questions, the participants' faces were analyzed with thermal cameras. The authors report an overall accuracy rate of over 90%.

There are, however, certain aspects of thermal imaging that merit attention. Most importantly, the assumption that an alteration in blood flow is the result of deception is controversial. For example, it is possible that innocent airport passengers will also show an increase in blood flow when being interviewed by security officers because of heightened arousal unrelated to deception (see SPOT). People could be worried about missing their flight, might feel anxious about the upcoming flight, or might find the conversation with a security officer stressful (Warmelink et al., 2011). Furthermore, the accuracy rates might have been influenced by the use of validated interviewing approaches (i.e., the Concealed Information Test, Park et al., 2013) and by increasing the cognitive load in participants (Rajoub & Zwiggelaar, 2014). The apparent successful truth–lie discrimination might, therefore, not be attributable to the ability to pick up stress during deception through thermal imaging. Rather, the results might be due to the use of methods that tap into cognitive differences between lying

and truth telling. Moreover, physiological variables like skin-surface temperature are affected by individual characteristics other than deception, such as illness, body metabolism, and facial expressions (Khan, Ingleby, & Ward, 2006; Rajoub & Zwiggelaar, 2014). Environmental factors like the temperature and humidity level of a room also affect thermal measurements (Park et al., 2013). These noise factors might deteriorate the measurements in settings like an airport where passengers might experience stress in a rather unique environment. Most thermal-imaging studies indeed took place in highly controlled lab settings (e.g., Park et al., 2013; Pavlidis et al., 2000; Rajoub & Zwiggelaar, 2014; Warmelink et al., 2011), and it remains an open question whether the effects are stable in less controlled (e.g., the airport) settings.

Reaction Times

Outline

Since the early 2000s, there has been an increased interest in reaction time (RT)-based deception detection (for reviews see Verschuere, Suchotzki, & Debey, 2014; Suchotzki, Verschuere, Van Bockstaele, Ben-Shakhar, & Crombez, 2017). There are several RT-based deception paradigms (e.g., CIT, autobiographical Implicit Association Test [aIAT], Sheffield Lie Test), which have in common that RTs can shed light on the differences in information processing involved in lying versus truth-telling. These paradigms assume that (1) the truth is the dominant, automatic response and increases the speed of responding to stimuli; and/or that (2) lying puts greater demands on cognitive abilities than truth telling which consequently delays the speed of responding. We focus our discussion on the two paradigms that may be readily applied for deception detection at the individual level: the aIAT and the RT-based CIT.

The aIAT assesses which of two conflicting statements is true, by evaluating their association with true and false propositions. Imagine that a refugee asked about activities in his country of origin, Syria, tells the interviewer that he assisted a volunteer organization called White Helmets. The investigator, however, has reason to doubt that statement and suspects that the refugee was, in fact, selling oil to support ISIS. In that situation, an aIAT could be used to contrast "I assisted the White Helmets" with "I sold oil for ISIS." In an aIAT, the participant pairs statements that are autobiographical (e.g., "I sold oil for ISIS") with objectively true and false statements (e.g., "I am in front of a computer," "I am hiking in the

mountains"). Each statement appears one by one on a computer screen and the participant determines its label by pushing one of two buttons on a keyboard. The general assumption of the aIAT is that the examinee will be faster in associating the true statement with the label *true*, and the false statement with the label *false*. For example, faster associations (i.e., shorter response latencies) for "I assisted the White Helmets" with true (and "I sold oil for ISIS" with false) than "I sold oil for ISIS" with true (and "I assisted the White Helmets" with false) would corroborate the refugee's story (see also Chapters 10 and 11). If the refugee was faster in associating "I sold oil for ISIS" with true and "I assisted the White Helmets" with false, this would contradict the subject's original account and might be incriminating.

The RT-CIT (here simply: CIT), in contrast to the aIAT, assesses recognition of critical pieces of information and can be used as a deception-detection method if knowledge of the critical information is unique to the criminal who committed a crime. Think back to the sabotage at the Belgian nuclear plant. Many employees have knowledge of the site, but to the extent that investigators kept the information private, only the culprit would know the specifics of the sabotage (e.g., the exact series of actions that led to the overheating). Those actions could be the critical informational items for a CIT. Similarly, whereas investigators will typically ask the refugee about intimate knowledge of the White Helmets that real volunteers are likely to have (Veldhuizen, Horselenberg, Landström, Granhag, & Koppen, 2017), a CIT would ask about intimate knowledge an ISIS oil seller would have (e.g., who are the buyers of the oil; the current oil price). In a CIT, one embeds critical information within a series of plausible (yet noncritical) alternatives. For instance, "You told me you do not know what price ISIS has been selling oil in the last 3 months of your stay in Syria. I will ask you whether it was 15, 20, 30, 35, or 40 USD a barrel." Denying any knowledge, the refugee is expected to answer NO to all answer options. A slower response on the correct answer as compared to the alternative answers is taken as an indication of recognition. The CIT is built on the premise that only the examinee with intimate knowledge will recognize the correct answer. Although the underlying cognitive processes leading to the RT slowing remain to be fully explored, a likely candidate explanation is response inhibition (see Suchotzki, Verschuere, Peth, Crombez, & Gamer, 2015; Verschuere & De Houwer, 2011). Specifically, it is assumed that, only in those with intimate knowledge, the response tendency elicited by the correct answer (YES) conflicts with the response denying involvement (NO).

Evaluation

A recently published metaanalysis (Suchotzki et al., 2017) suggests that RTs may be a useful means to assess deception. Across contexts and paradigms, RT-based deception-detection tests had a large standardized-effect size (Cohen's $d = 1.05$), although a subsequent, smaller meta-analysis in the same paper showed that both the CIT and the aIAT seemed susceptible to countermeasures.

The aIAT has further been reviewed by its developers who conclude that classification accuracies (i.e., truth tellers and liars identified as such) range above 90% (Agosta & Sartori, 2013), making it a tool suitable for the detection of deception at the individual level (see also (Hu & Rosenfeld, 2012). Some of the results of the aIAT have been independently replicated, although typically with more modest accuracy (e.g., 81% accuracy; Verschuere & Kleinberg, 2017, 67%–86% in guilty and 61% in innocent participants, Verschuere, Prati, & Houwer, 2009; see also Hu, Rosenfeld, & Bodenhausen, 2012).

The CIT has since its inception (Seymour, Seifert, Shafto, & Mosmann, 2000; for a review, see Verschuere et al., 2011) been applied in different contexts and the response latency difference between critical and noncritical stimuli has been found in numerous studies (Seymour et al., 2000; Suchotzki et al., 2017). Using an espionage scenario, Seymour et al. (2000) found support for the use of reaction times as cue for the detection of concealed information. Noordraven and Verschuere (2013) found the CIT capable of identifying the planning of a mock crime with an area under the curve (AUC) of 0.87. The AUC expresses the diagnostic efficiency of a criterion (e.g., the RT difference) and ranges from 0.5 (random classification) to 1.00 (perfect classification). In contrast to accuracy rates, the AUC is an indication of the diagnostic power of a criterion across all observed cutoff points (i.e., it also becomes larger if very high, or very low values are observed for participants belonging to the respective class, e.g., truthful and deceptive). AUCs offer a more comparable metric across methods and approaches (National Research Council, 2003).

The reaction time-based CIT paradigm has also been applied to fake identity settings and applied as such in online environments. For example, participants in Verschuere and Kleinberg (2016) concealed salient details of their identity (e.g., their first name). Using known moderators of the CIT effect (i.e., using multiple highly salient pieces of information increases accuracy), a state-of-the-art CIT showed an AUC of 0.98 with an accuracy of 86%.

Notwithstanding the promising rationale and findings, RT-based deception detection is facing specific challenges. For the CIT, potential leakage of critical information is a major limitation. If the critical information is accessible to more people than just the true perpetrator, the recognition indicated by the test is not unique to the perpetrator, and hence the conclusions become invalid. Further, both the CIT and the aIAT have been shown susceptible to countermeasures (i.e., faking the test, see Suchotzki et al., 2017; Verschuere et al., 2009). Finally, the associations captured by the aIAT might come about through processes other than deception: the association of false sentence with an ISIS-related proposition might be due to the shared negative connotation of both propositions (i.e., they are both perceived as highly negative and therefore rapidly paired; see Rothermund & Wentura, 2004; Verschuere et al., 2014).

Verbal Content

Outline

The verbal approach to deception is based on the assumption that the content of a truthful statement differs from that of a deceptive statement (e.g., Johnson & Raye, 1981; Köhnken, 2004). Reality Monitoring, for instance, suggests that truthful stories are told differently because they have been stored in the memory through perceptual processes (e.g., smelling, seeing, hearing). Deceptive stories, on the other hand, have never been truly experienced, and are obtained through internal, fabricated cognitive processes (Johnson & Raye, 1981). A meta-analysis supports the usefulness of Reality Monitoring for deception detection (e.g., truthful stories contained more visual, auditory, and temporal detail compared to deceptive ones; Masip, Sporer, Garrido, & Herrero, 2005).

As with most deception cues, the association of verbal content differences and deception is weak (DePaulo et al., 2003). Therefore, it has been recommended to focus on techniques that increase these verbal differences (Vrij, Granhag, & Porter, 2010; for further developments like the Verifiability Approach, see Nahari, Vrij, & Fisher, 2014). A series of methods to enhance truth–lie differences have been proposed (e.g., telling a story in reverse order, maintaining eye contact during the interview, drawings; see Vrij, Fisher, & Blank, 2015). The model statement (i.e., providing a detailed example answer) and the unexpected question technique (i.e., asking questions that the suspects do not anticipate, e.g., about the physical layout of a restaurant) seem most readily applicable. Statements from truth tellers

are richer in detail, but they do not always know how much information they need to include in their stories. In the model statement technique, people are, therefore, asked to read a detailed example statement, so that they know how much information they need to mention (Vrij et al., 2015). This is particularly beneficial to truth tellers because their detailed statements could prove that they are telling the truth. For deceptive persons, however, including more information could lead to cues that reveal their lies. Another technique that is widely used in the verbal approach is asking unanticipated questions. Liars often prepare their story, which makes it easier for them to appear truthful. If, however, an interviewer asks unexpected questions, liars have to fabricate a plausible answer on the spot, thereby enlarging the opportunity of being caught (Vrij, Granhag, Mann, & Leal, 2011).

Contemporary verbal approaches emphasize the need to ask questions that actively elicit verbal differences between truth tellers and liars. Interviewers are encouraged to use an information-gathering interviewing style, rather than using accusatorial questioning techniques (Vrij et al., 2010). By applying cognitive interviewing techniques such as asking open-ended questions and building rapport, it is likely that both truth tellers and liars will provide more information when being interviewed. Truth-tellers' providing extra information could lead to cues that prove that their story is genuine (e.g., the person they claim to have an appointment with can be verified). Liars, on the other hand, could risk mentioning information that suggests that they are deceptive (e.g., contradicting information, Vrij et al., 2015).

Evaluation

There are indications that the verbal approach to deception detection could be useful in security settings. Research from the past few years suggested that the verbal approach is also promising when discriminating between genuine and deceptive intentions. For instance, deceptive accounts of false intentions were found to contain fewer details compared to stories of true intentions when unexpected questions are asked (e.g., Sooniste, Granhag, Knieps, & Vrij, 2013; Warmelink, Vrij, Mann, & Granhag, 2013; but see Fenn, McGuire, Langben, & Blandón-Gitlin, 2015). Liars also appeared less plausible and more contradictory in their statement compared to truth tellers (e.g., Vrij et al., 2011). Despite the promising theoretical rationale behind the verbal approach to detection, it must be noted that the effect sizes are small and that the reported high-accuracy rates are often found in discriminant analysis and are not cross-validated.

The verbal approach to deception detection is promising due to its simplicity and reliance on rather easily gathered data (i.e., spoken or written statements). Nevertheless, it is yet an unsettled issue how this method can address so-called embedded lies (i.e., in which the lie is just a small part of a largely truthful account), practiced lies (i.e., in which the liar is used to telling a detailed and plausible, false account), and complicated truths (i.e., in which the truth is more complex, vaguer, or less plausible than the lie).

POSSIBILITIES FOR LARGE-SCALE APPLICATIONS

Criteria for Large-Scale Applicability

The previous section outlined various deception-detection methods and evaluated their theoretical background, whereas this part focuses on the large-scale potential. To examine how fit various deception detection methods are for the application at scale, we assess those methods that were evaluated as theoretically sound, valid, or at least promising in their application (thermal imaging, reaction times, and verbal content (Table 16.1)) on two requirements. The large-scale fitness and suitability are defined through (1) the possibility of using the method in quick procedures (quick data collection), and (2) the flexibility of applying the method in various contexts and scenarios (flexibility).

Quick data collection is essential for any method applied to larger numbers of people because of the sheer fact that much more than single individuals are subjected to a test poses logistical challenges. For example, a method that requires extended interaction with participants or relies heavily on large apparatuses is less suitable than a simple online method. The flexibility criterion pertains to the ability of a method to be used in different contexts (e.g., airport passenger screening, criminal investigations) and scenarios (e.g., false identity allegations, burglaries, threat assessment). Although one could argue that every application inherently defines a specific context for a deception detection method (e.g., a threat assessment method at the airport), the aim of this section is to evaluate the different methods independent of a specific, well-defined application context.

Thermal Imaging

One potential advantage of using thermal-imaging techniques is that it is noninvasive and—privacy and ethical issues aside—that it does not require cooperation from passengers (Rajoub & Zwiggelaar, 2014). For instance,

Table 16.1 Summary of large-scale criteria per deception-detection method

	Quick data collection	Flexibility	Theoretical foundation
Thermal imaging	Promising if combined with proper interviewing techniques; challenge lies in the equipment needed and an automated analytical pipeline.	Difficult in stress-inducing contexts (e.g., airports); equipment needed is static (e.g., Infrared [IR] sensors installed at specific locations).	Largely determined by the interviewing paradigm.
Reaction times	Promising but still relies on a large number of trials.	Poor. Needs fine-tuned stimulus selection and careful consideration on a case-by-case basis.	Rooted in cognitive psychology.
Verbal content	Currently not fit for large-scale purposes; still relies on face-to-face interviews and/or long texts.	Promising; needs little baselining or calibration.	Theoretically embedded in theories such as Reality Monitoring.

conducting thermal-imaging measurements does not require any interaction (e.g., attaching sensors) with passengers (Arora et al., 2008). Because the thermal cameras largely resemble normal cameras used in an airport, it is possible to scrutinize passengers without their awareness, which might also offer a strategy to prevent passengers' countermeasures (Park et al., 2013). Furthermore, data collection and analysis could be fully automated, an important aspect of the implementation of screening tools in large-scale settings, although developing these automated systems would still prove a challenge.

Although the measurement of thermal changes in the facial regions can in principle be done unobtrusively, it is important to realize that current thermal-imaging deception detection requires participants to hold still (to assess the temperature in certain facial regions) and it relies on interviewing techniques to assess changes in facial temperature while answering questions. Accompanied by the thermal measurements are questions aimed at identifying deceptive passengers before the thermal method would be of any use. The results from the abovementioned study of Park et al. (2013) suggest that CIT techniques (questioning people about critical information specific to the crime scenario under investigation) could be helpful in discriminating between truth tellers and liars. However, security agents at border settings do not possess the necessary critical crime information. Interviewing techniques based on verbal and cognitive principles (e.g., asking questions in reverse order) might be useful as a supplement to thermal imaging techniques. Regarding the first criterion of quick data collection, thermal-imaging techniques seem promising once interviewing techniques and the analytical process are automated.

The second criterion of flexibility proves trickier for thermal imaging. As already mentioned in the evaluation of the method, it could be argued that areas such as airports are not a proper location for the use of thermal imaging. People at the airport can become easily aroused because of innocent stress regarding their flight, and environmental factors could influence the measurements. It remains questionable whether this technique will apply to large-scale security settings.

Reaction Times

The minimal requirements needed to administer RT-based tests makes them attractive for large-scale purposes. The first adaptations of the aIAT (e.g., Greenwald, McGhee, & Schwartz, 1998) and RT-CIT (e.g., Seymour et al., 2000) were conducted with specialist laboratory software

for the accurate timing of stimuli presentations and the precise recording of response latencies. However, recent developments showed that RT tests can be administered reliably and validly online (e.g., Kleinberg & Verschuere, 2015, 2016; Lukács, Kleinberg, & Verschuere, 2017; Verschuere & Kleinberg, 2016, 2017). For example, results of lab studies were replicated with considerably larger sample sizes in a fraction of the time of lab studies, whereas participants merely needed a computer with an Internet connection and a standard web browser (e.g., Kleinberg & Verschuere, 2015). The online replications do depend on the context of deception and seem particularly promising for autobiographical settings (e.g., lying about one's identity), so future investigations will need to show that other settings (e.g., mock crimes) can be replicated online as well. Nevertheless, with rapid technology improvement of web browsers, the difference between lab software and online tool will likely begin to vanish (for a discussion, see van Steenbergen & Bocanegra's, 2016, response to Plant, 2016).

Limitations concerning the applicability of the aIAT and CIT stem from a methodological and an implementation point of view. On a methodological level, the RT-based tests require a minimum duration to derive at participant-level predictions. That is, the length of the tests cannot be shortened drastically because one must instruct the participant, provide some practice with the task, and provide sufficient trials to account for errors, missed trials, and different blocks (aIAT) and stimuli presentation proportions (CIT). Likewise, findings from online studies suggest that populations not used to such fast-paced reaction-time tasks require a detailed introduction procedure to the test. Such a time frame can be considered realistic for some applications (e.g., criminal investigations) but may be too long for others (e.g., large-scale screening at airports). Even a short test will probably take about 10 min. From an implementation perspective, the RT-based tests—even if the length problem could be solved—are likely to be perceived as rather unnatural, for example, by passengers on an airport, and cannot be administered unobtrusively. An additional challenge for possible remote testing settings (e.g., testing participants online) is the verification of the test taker's identity[2]—although this limitation is not unique to RT tests but an impediment for all remote testing settings.

[2] It might be feasible to address the identity verification with a CIT itself (Verschuere & Kleinberg, 2016).

The quick data collection criterion discussed earlier is connected to the flexibility criterion. A key prerequisite for the both the CIT and aIAT is a fine-tuned stimulus selection. For the CIT, it takes information about the criminal context (i.e., which are the critical items) as well as a selection of plausible yet unrelated alternatives (i.e., what are the noncritical, matched items). Similarly, it is essential for the aIAT that the associations indicated by the RT differences are a result of deception and not an irrelevant cognitive process such as preference or similarity. For example, "I am in prison" (=the true/false category) and "I have stolen the money" (=the autobiographical category) both share a negative connotation and are not necessarily only associated by truthfulness or deception. If the stimulus selection problem were solved, both tests would be quicker (because the preparation time before the actual test is drastically reduced). With quicker administration procedures, the tests become more flexible because the context (and hence the stimulus selection) is adaptable. To date, the formulation of good test items for the CIT and aIAT are the key impediment for large-scale applications. In the future, this problem could be addressed with advanced statistical modeling and item calibration (e.g., empirically determining the truthful response profile of an exhaustive set of stimuli—e.g., months of birth—and testing how much truthful and deceptive participants deviate from that profile).

Although a large-scale implementation of these tests has not yet been done, RT-based deception detection seems a candidate worth trying on a large scale. Contrary to other deception detection methods, the RT-based tests are among the few that fulfill the requirement of being applicable with relatively minimal equipment and are still considerably quicker for participants than physiological and brain-based methods.

Verbal Content

When looking at the first criterion, quick data collection, the verbal content analysis seems challenging at first glance. Many studies on verbal deception detection rely on personal face-to-face interviews that are manually coded afterward (i.e., oral statements are being transcribed, read by one or more independent coders, and finally scored on variables such as plausibility or richness of detail; see Nahari, 2016). Despite promising results, in its current state, the verbal content method excludes the possibility of real-time data collection, an essential criterion for quick data collection.

Can the verbal approach be used for large-scale purposes? For that to be the case, two key requirements should be met. First, the interviewing procedures (i.e., the collection of data) must become quicker—either through shorter interviews or through online procedures (e.g., providing a statement online before arriving at a security checkpoint). Second, the analysis of provided statements (i.e., the data analysis) must transition toward fast and scalable procedures. The latter could be achieved with methods that allow quick real-time interviewer judgments or through computer-automated analyses. A recent study shows there may be a possibility of short interviews and on-the-spot assessments (Ormerod & Dando, 2015). This extensive in vivo randomized-controlled trial on several international airports tested whether participants would pass security with fake identity papers. After a few days of preparation, the mock passengers went to the airport and tried to pass security screening without being exposed as a liar by security agents. The critical test was whether security personnel trained in two different methods would be able to detect the mock passengers. Using cognitive questioning techniques and paying close attention to the verbal accounts of passengers (termed the Controlled Cognitive Engagement [CCE] method), it turned out that security agents were able to correctly identify more mock passengers (66% of all mock passengers) compared to agents using a behavioral, suspicious-signs method (3%). Agents focusing on verbal content also interviewed their passengers more quickly than did the behavior-detection agents, implying that the verbal method might even be more time effective than methods presently used in security screening. Considering the criterion of quick data analysis (circa 3 min), CCE sounds promising. Furthermore, developments in Natural Language Processing approaches to deception detection might be a promising way to substitute or approximate parts of the manual coding by fully automated scoring (Bond & Lee, 2005; Ott, Cardie, & Hancock, 2013; Ott, Choi, Cardie, & Hancock, 2011). Regarding the flexibility criterion, there is little reason to assume that other deception contexts (e.g., mock crime, malicious intent) would impede CCE-like methods. The original study must be replicated and extended to diverse contexts to reach a verdict on the flexibility. It also remains an open question how exactly the CCE-trained security agents derive their subjective judgment and which verbal or behavioral cues they pay attention to.

OUTLOOK ON THE FUTURE: WHAT NEEDS TO BE DONE?

To bridge the gap between existing deception detection methods and large-scale applications we identify two requirements for research in the next few years: cross-disciplinary collaboration and rigorous research transparency. First, a shift toward large-scale methods will imply that individual researchers need to form interdisciplinary teams. It will be imperative to overcome islands of expertise that are focused on either a method (e.g., verbal content analysis) or technique (e.g., automated analysis). For example, although the psycho-legal deception research community is paramount for theoretical frameworks underlying deception-detection methods and the provision of theoretically founded deception cues (e.g., the plausibility of statements), their resources for developing large-scale methods are limited. Computational disciplines (e.g., computational linguistics) can add to the integration of cues (e.g., dozens of content-based cues) and methods (e.g., verbal content and thermal imaging), the automation of cues (e.g., automating the scoring of a statement's plausibility), and the development of predictive models (e.g., through [un] supervised machine-learning tasks). An illustration of multimodal approaches—contrary to isolated, discipline-specific efforts (e.g., Psychology, Computational Linguistics, Neuroscience)— is shown in Box 16.1.

Box 16.1 Illustration of multimodal deception methods

Combining Methods: Multimodal Deception Detection

One key feature of the methods discussed earlier in this chapter is that they are unimodal; that is, they assume a relationship between the mental state of deception and one specific outcome measure (e.g., reaction time differences, verbal content differences). There are indications that deception detection might benefit from combining multiple cues (Hartwig & Bond, 2014).

For example, Pérez-Rosas et al. (2015) used verbal and nonverbal (i.e., facial displays and hand gestures) variables to classify cases as deceptive or truthful. They found that machine-learning classifiers based on a combined set of predictors (verbal + nonverbal, 77.11%) were more accurate than verbal (accuracy: 65.25%) or nonverbal (75.42%) predictors alone. Further evidence of multimodal approaches comes from Hu and Rosenfeld (2012), who combined EEG and reaction-time measurements and achieved an AUC of 0.98 (compared to 0.84 and 0.95 for reaction times and EEG alone, respectively). Furthermore, Abouelenien, Pérez-Rosas, Mihalcea, & Burzo, 2014) examined physiological variables (e.g., heart rate, skin conductance), linguistic variables (unigrams and psycholinguistic variables), and thermal variables. The latter resulted in heat

Box 16.1 Illustration of multimodal deception methods—cont'd

maps of the facial area of participants during their deceptive and truthful statements. The combination of linguistic and thermal predictor variables resulted in significant accuracy increases compared to unimodal classifications.

Although these results hint at the benefits of multimodal classifiers over unimodal ones, it remains an open question how much better the accuracy of the former needs to be to justify additional data collection for a new modality.

The work on multimodal methods is relatively young but has yielded promising results. As Abouelenien et al. (2014) have shown, accuracies above 70% can be obtained without any involvement of human annotators. The key challenge for such methods is the time needed to gather reliable physiological, linguistic, and thermal data. On a large scale, time is a premium (see Honts & Hartwig, 2014; Kleinberg et al., 2017) and a short procedure is one of the key requirements. At the same time, a multimodal approach brings about new challenges, including the handling of false positives that come with each method and the generalizability of machine-learning models to various contexts.

Second, for deception research applications on a large scale, we believe it is necessary that the disciplines involved embrace the open science philosophy of sharing data and methods. Be it from the point of view that future large-scale applications are inevitably intertwined with computational methods (e.g., machine learning), or from the perspective that open data will allow metaanalytical research on raw data, the necessity, relevance, and usefulness of deception research will depend on the sharing of data. With the logistic efforts of data sharing minimized (see the Open Science Framework's data repository, www.osf.io), it will be paramount for the future generations of researchers equipped with ever more sophisticated tools, to be able to rely on, learn from, and make use of existing data. Consider, as an illustration, the example of Ott et al.'s (2013, 2011) studies on fake and genuine hotel reviews. Ott et al. compiled a corpus of 1600 hotel reviews and made it publicly accessible to everyone. Their data set resulted in novel methodologies (Feng & Hirst, 2013) and insights (Fornaciari & Poesio, 2014) into verbal deception detection that would otherwise have been slowed down by individual efforts of collecting identical or similar data. The efforts made by the original researchers as well

as the contributions by other research groups should be encouraged and should pave the way for a better, open approach toward deception detection.

CONCLUSION

In this chapter, we discussed a selection of deception detection methods concerning their potential of being applied on a large scale. Although some methods are easily usable on a large scale (e.g., behavioral observation, speech analysis) they lack theoretical underpinnings, empirical validation, or both. Other methods are well founded in theory and validated (e.g., RT-based tests, verbal content tools) but have yet to take final steps to be useful for applications at scale. Although each method has advantages and shortcomings, likely candidates for applications at scale are reaction time-based methods as well as verbal content analysis. The future of deception-detection systems suitable for scenarios in which scores of people are investigated might lie in intelligent methodological integration (e.g., multimodal methods, computational methods with theoretically founded cues) as well as cross-disciplinary collaboration. An important step in that direction can be made through the sharing of data and tools.

ACKNOWLEDGMENTS

Bennett Kleinberg received funding from the Dutch Ministry of Security and Justice.

REFERENCES

Abouelenien, M., Pérez-Rosas, V., Mihalcea, R., & Burzo, M. (2014). Deception detection using a multimodal approach. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 58–65). ACM.

Agosta, S., & Sartori, G. (2013). The autobiographical IAT: A review. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00519.

American Civil Liberties Union. (2017). *Bad trip: debunking the TSA's "behavior detection" program*. Retrieved from https://www.aclu.org/report/bad-trip-debunking-tsas-behavior-detection-program.

Arora, N., Martins, D., Ruggerio, D., Tousimis, E., Swistel, A. J., Osborne, M. P., & Simmons, R. M. (2008). Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer. *The American Journal of Surgery, 196*(4), 523–526. https://doi.org/10.1016/j.amjsurg.2008.06.015.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2.

Bond, G. D., & Lee, A. Y. (2005). Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. *Applied Cognitive Psychology, 19*(3), 313–329.

Damphousse, K. R., Pointon, L., Upchurch, D., & Moore, R. K. (2007). Assessing the validity of voice stress analysis tools in a jail setting. In *This research project was flagrant from the National Institute of Justice (2005-IJ-CX-0047)*. Retrieved from https://web.elastic.org/~fche/mirrors/antipolygraph.org/documents/219031.pdf.

DeClercq, G. (2014). *UPDATE 2-Belgian Doel 4 nuclear reactor closed till year-end*. Reuters. Retrieved from http://uk.reuters.com/article/belgium-nuclear-doel-idUKL6N0QK43R20140814.

Department of Homeland Security. (2013). *Transportation security administration's screening of passengers by observation techniques*. Retrieved from https://www.oig.dhs.gov/assets/Mgmt/2013/OIG_13-91_May13.pdf.

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.

Eriksson, A., & Lacerda, F. (2008). Charlatanry in forensic speech science: A problem to be taken seriously. *International Journal of Speech Language and the Law, 14*(2). https://doi.org/10.1558/ijsll.2007.14.2.169.

Feng, V. W., & Hirst, G. (2013). Detecting deceptive opinions with profile compatibility. In *IJCNLP* (pp. 338–346). Retrieved from ftp://128.100.3.31/dist/gh/Feng+Hirst-IJCNLP-2013.pdf.

Fenn, E., McGuire, M., Langben, S., & Blandón-Gitlin, I. (2015). A reverse order interview does not aid deception detection regarding intentions. *Frontiers in Psychology, 6*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4553365/.

Fornaciari, T., & Poesio, M. (2014). *Identifying fake Amazon reviews as learning from crowds*. Association for Computational Linguistics. Retrieved from http://repository.essex.ac.uk/id/eprint/14591.

Gamer, M., Rill, H.-G., Vossel, G., & Gödert, H. W. (2006). Psychophysiological and vocal measures in the detection of guilty knowledge. *International Journal of Psychophysiology, 60*(1), 76–87. https://doi.org/10.1016/j.ijpsycho.2005.05.006.

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: An fMRI investigation. *Cerebral Cortex, 13*(8), 830–836.

Ganis, G., Morris, R. R., & Kosslyn, S. M. (2009). Neural processes underlying self-and other-related lies: An individual difference approach using fMRI. *Social Neuroscience, 4*(6), 539–553.

Ganis, G., Rosenfeld, J. P., Meixner, J., Kievit, R. A., & Schendan, H. E. (2011). Lying in the scanner: Covert countermeasures disrupt deception detection by functional magnetic resonance imaging. *Neuroimage, 55*(1), 312–319.

Government Accountability Office. (2010). *Aviation security: Efforts to validate TSA's passenger screening behavior detection program underway, but opportunities exist to strengthen validation and address operational challenges*. Retrieved from http://www.gao.gov/assets/310/304510.pdf.

Government Accountability Office. (2013). *Aviation security: TSA should limit future funding for behavior detection activities*. Retrieved from https://www.gao.gov/assets/660/658923.pdf.

Government Accountability Office. (2015). *Aviation security: Improved testing, evaluation, and performance measurement could enhance effectiveness*. Retrieved from https://www.gao.gov/assets/680/673490.pdf.

Government Accountability Office. (2016). *Aviation security: Airport perimeter and access control security would benefit from risk assessment and strategy updates*. Retrieved from https://www.gao.gov/assets/680/677586.pdf.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464.

Hartwig, M., & Bond, C. F. (2014). Lie detection from multiple cues: A meta-analysis: Lie detection from multiple cues. *Applied Cognitive Psychology, 28*(5), 661–676. https://doi.org/10.1002/acp.3052.

Harvey, A. C., Vrij, A., Nahari, G., & Ludwig, K. (2017). Applying the verifiability approach to insurance claims settings: Exploring the effect of the information protocol. *Legal and Criminological Psychology, 22*(1), 47–59.

Hollien, H., & Harnsberger, J. D. (2006). *Voice stress analyzer instrumentation evaluation*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.525.3830.

Honts, C., & Hartwig, M. (2014). Credibility assessment at portals. In D. C. Raskin, C. Honts, & J. Kircher (Eds.), *Credibility assessment: Scientific research and applications* (pp. 37–62). Academic Press.

Horvath, F. (1982). Detecting deception: The promise and the reality of voice stress analysis. *Journal of Forensic Sciences, 27*(2), 11488J. https://doi.org/10.1520/JFS11488J.

Hu, X., & Rosenfeld, J. P. (2012). Combining the P300-complex trial-based concealed information test and the reaction time-based autobiographical implicit association test in concealed memory detection: P300-CIT and RT-aIAT in memory detection. *Psychophysiology*. https://doi.org/10.1111/j.1469-8986.2012.01389.x.

Hu, X., Rosenfeld, J. P., & Bodenhausen, G. V. (2012). Combating automatic autobiographical associations: The effect of instruction and training in strategically concealing information in the autobiographical implicit association test. *Psychological Science, 23*(10), 1079–1085. https://doi.org/10.1177/0956797612443834.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological Review, 88*(1), 67.

Khan, M. M., Ingleby, M., & Ward, R. D. (2006). Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations. *ACM Transactions on Autonomous and Adaptive Systems, 1*(1), 91–113. https://doi.org/10.1145/1152934.1152939.

Kleinberg, B., Arntz, A., & Verschuere, B. (2017). Detecting deceptive intentions: Possibilities for large-scale applications. In T. Docan-Morgan (Ed.), *The handbook of deceptive communication* (in press).

Kleinberg, B., & Verschuere, B. (2015). Memory detection 2.0: The first web-based memory detection test. *PLoS One, 10*(4), e0118715. https://doi.org/10.1371/journal.pone.0118715.

Kleinberg, B., & Verschuere, B. (2016). The role of motivation to avoid detection in reaction time-based concealed information detection. *Journal of Applied Research in Memory and Cognition, 5*(1), 43–51. https://doi.org/10.1016/j.jarmac.2015.11.004.

Köhnken, G. (2004). Statement validity analysis and the "detection of the truth". In *The detection of deception in forensic contexts* (pp. 41–63).

Lukács, G., Kleinberg, B., & Verschuere, B. (2017). Familiarity-related fillers improve the validity of reaction time-based memory detection. *Journal of Applied Research in Memory and Cognition*. https://doi.org/10.1016/j.jarmac.2017.01.013.

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

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime and Law, 11*(1), 99–122. https://doi.org/10.1080/10683160410001726356.

Meijer, E. H., Bente, G., Ben-Shakhar, G., & Schumacher, A. (2013). Detecting concealed information from groups using a dynamic questioning approach: Simultaneous skin conductance measurement and immediate feedback. *Frontiers in Psychology, 4*. https://doi.org/10.3389/fpsyg.2013.00068.

Nahari, G. (2016). When the long road is the shortcut: A comparison between two coding methods for content-based lie-detection tools. *Psychology, Crime and Law, 22*(10), 1000–1014.

Nahari, G., Vrij, A., & Fisher, R. P. (2014). Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology, 19*(2), 227–239. https://doi.org/10.1111/j.2044-8333.2012.02069.x.

National Research Council. (2003). *The polygraph and lie detection*. Committee to Review the Scientific Evidence on the Polygraph. Division of Behavioral and Social Sciences and Education.

Noordraven, E., & Verschuere, B. (2013). Predicting the sensitivity of the reaction time-based concealed information test: Detecting deception with the concealed information test. *Applied Cognitive Psychology, 27*(3), 328–335. https://doi.org/10.1002/acp.2910.

Ormerod, T. C., & Dando, C. J. (2015). Finding a needle in a haystack: Toward a psychologically informed method for aviation security screening. *Journal of Experimental Psychology: General, 144*(1), 76.

Ott, M., Cardie, C., & Hancock, J. T. (2013). Negative deceptive opinion spam. In *HLT-NAACL* (pp. 497–501).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (Vol. 1, pp. 309–319). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2002512.

Panasiti, M. S., Cardone, D., Pavone, E. F., Mancini, A., Merla, A., & Aglioti, S. M. (2016). Thermal signatures of voluntary deception in ecological conditions. *Scientific Reports, 6*(1). https://doi.org/10.1038/srep35174.

Park, K. K., Suk, H. W., Hwang, H., & Lee, J.-H. (2013). A functional analysis of deception detection of a mock crime using infrared thermal imaging and the concealed information test. *Frontiers in Human Neuroscience,* 7. https://doi.org/10.3389/fnhum.2013.00070.

Pavlidis, I., Eberhardt, N. L., & Levine, J. A. (2002a). Erratum: Seeing through the face of deception. *Nature, 415*(6872), 602. https://doi.org/10.1038/415602b.

Pavlidis, I., Eberhardt, N. L., & Levine, J. A. (2002b). Seeing through the face of deception. *Nature, 415*(6867), 35. https://doi.org/10.1038/415035a.

Pavlidis, I., & Levine, J. (2002). Thermal facial screening for deception detection. In *Engineering in medicine and biology, 2002. 24th annual conference and the annual fall meeting of the biomedical engineering society EMBS/BMES conference, 2002. Proceedings of the second joint* (Vol. 2, pp. 1143–1144). IEEE.

Pavlidis, I., Levine, J., & Baukol, P. (2000). Thermal imaging for anxiety detection. In *Computer vision beyond the visible spectrum: Methods and applications, 2000. Proceedings. IEEE workshop on* (pp. 104–109). IEEE.

Pérez-Rosas, V., Abouelenien, M., Mihalcea, R., Xiao, Y., Linton, C. J., & Burzo, M. (2015). Verbal and nonverbal clues for real-life deception detection. In *EMNLP* (pp. 2336–2346).

Perry, M., & Gilbey, A. (2011). The screening of passengers by observation techniques programme. *Aviation Security International, 17*(3), 12.

Plant, R. R. (2016). A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter. *Behavior Research Methods, 48*(1), 408–411. https://doi.org/10.3758/s13428-015-0577-0.

Rajoub, B. A., & Zwiggelaar, R. (2014). Thermal facial analysis for deception detection. *IEEE Transactions on Information Forensics and Security, 9*(6), 1015–1023. https://doi.org/10.1109/TIFS.2014.2317309.

Rosenfeld, J. P., Hu, X., Labkovsky, E., Meixner, J., & Winograd, M. R. (2013). Review of recent studies and issues regarding the P300-based complex trial protocol for detection of concealed information. *International Journal of Psychophysiology, 90*(2), 118–134.

Rosenfeld, J. P., Soskins, M., Bosh, G., & Ryan, A. (2004). Simple, effective countermeasures to P300-based tests of detection of concealed information. *Psychophysiology, 41*(2), 205–219.

Rothermund, K., & Wentura, D. (2004). Underlying processes in the implicit association test: Dissociating salience from associations. *Journal of Experimental Psychology: General, 133*(2), 139–165. https://doi.org/10.1037/0096-3445.133.2.139.

Schmidt, M. S., & Lichtblau, E. (2012). Racial profiling rife at airport, US officers say. *The New York Times, 11*.

Seymour, T. L., Seifert, C. M., Shafto, M. G., & Mosmann, A. L. (2000). Using response time measures to assess "guilty knowledge". *Journal of Applied Psychology, 85*(1), 30–37.

Shipp, T., & Izdebski, K. (1981). Current evidence for the existence of laryngeal macrotremor and microtremor. *Journal of Forensic Sciences, 26*(3), 11391J. https://doi.org/10.1520/JFS11391J.

Sooniste, T., Granhag, P. A., Knieps, M., & Vrij, A. (2013). True and false intentions: Asking about the past to detect lies about the future. *Psychology, Crime and Law, 19*(8), 673–685. https://doi.org/10.1080/1068316X.2013.793333.

van Steenbergen, H., & Bocanegra, B. R. (2016). Promises and pitfalls of web-based experimentation in the advance of replicable psychological science: A reply to plant (2015). *Behavior Research Methods, 48*(4), 1713–1717. https://doi.org/10.3758/s13428-015-0677-x.

Suchotzki, K., Verschuere, B., Peth, J., Crombez, G., & Gamer, M. (2015). Manipulating item proportion and deception reveals crucial dissociation between behavioral, autonomic, and neural indices of concealed information: Concealed information test. *Human Brain Mapping, 36*(2), 427–439. https://doi.org/10.1002/hbm.22637.

Suchotzki, K., Verschuere, B., Van Bockstaele, B., Ben-Shakhar, G., & Crombez, G. (2017). Lying takes time: A meta-analysis on reaction time measures of deception. *Psychological Bulletin, 143*(4), 428–453. https://doi.org/10.1037/bul0000087.

Transportation Security Administration. (2009). *The truth behind the title: Behavior detection officer*. Retrieved from https://www.tsa.gov/blog/2008/02/29/truth-behind-title-behavior-detection-officer.

Transportation Security Administration. (2015). *Scientific substantiation of behavioral indicators*.

Truth and Deception Technologies. (2009). *DecepTech system features*. Retrieved from http://www.tdtvsa.com/features.htm.

Veldhuizen, T. S., Horselenberg, R., Landström, S., Granhag, P. A., & Koppen, P. J. (2017). Interviewing asylum seekers: A vignette study on the questions asked to assess credibility of claims about origin and persecution. *Journal of Investigative Psychology and Offender Profiling, 14*(1), 3–22.

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

Verschuere, B., & De Houwer, J. (2011). Detecting concealed information in less than a second: Response latency-based measures. In B. Verschuere, G. Ben-Shakhar, & E. Meijer (Eds.), *Memory detection: Theory and application of the concealed information test* (pp. 46–62).

Verschuere, B., & Kleinberg, B. (2016). ID-check: Online concealed information test reveals true identity. *Journal of Forensic Sciences, 61*, S237–S240. https://doi.org/10.1111/1556-4029.12960.

Verschuere, B., & Kleinberg, B. (2017). Assessing autobiographical memory: The web-based autobiographical implicit association test. *Memory*, 1–11. https://doi.org/10.1080/09658211.2016.1189941.

Verschuere, B., Prati, V., & Houwer, J. D. (2009). Cheating the lie detector: Faking in the autobiographical implicit association test. *Psychological Science, 20*(4), 410–413.

Verschuere, B., Suchotzki, K., & Debey, E. (2014). Detecting deception through reaction times. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting deception* (pp. 269–291). Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118510001.ch12.

Vrij, A. (2008). Reality monitoring. In *Detecting lies and deceit: Pitfalls and opportunities* (2nd ed.). John Wiley & Sons.

Vrij, A., Fisher, R. P., & Blank, H. (2015). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*. https://doi.org/10.1111/lcrp.12088.

Vrij, A., Granhag, P. A., Mann, S., & Leal, S. (2011). Lying about flying: The first experiment to detect false intent. *Psychology, Crime and Law, 17*(7), 611–620. https://doi.org/10.1080/10683160903418213.

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.

Warmelink, L., Vrij, A., Mann, S., & Granhag, P. A. (2013). Spatial and temporal details in intentions: A cue to detecting deception: Spatial and temporal details in lie detection. *Applied Cognitive Psychology, 27*(1), 101–106. https://doi.org/10.1002/acp.2878.

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*(1), 40–48. https://doi.org/10.1007/s10979-010-9251-3.

Weinberger, S. (2010). Intent to deceive? *Nature, 465*(7297), 412.

Wijn, R., van der Kleij, R., Kallen, V., Stekkinger, M., & de Vries, P. (2017). Telling friend from foe: Environmental cues improve detection accuracy of individuals with hostile intentions. *Legal and Criminological Psychology*. https://doi.org/10.1111/lcrp.12107.

Winter, J., & Currier, C. (2015). Exclusive: TSA's secret behavior checklist to spot terrorists. *The Intercept, 27*.