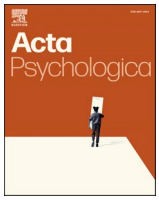
Acta Psychologica 251 (2024) 104548

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Your lies don't leave me cold: Assessing direct, indirect and physiological measures of lie detection

Rima-Maria Rahal [a,b,*], Teun Siebers [c], Willem W.A. Sleegers [a], Ilja van Beest [a]

[a] Tilburg University, Netherlands
[b] Max Planck Institute for Research on Collective Goods, Germany
[c] University of Amsterdam, Netherlands

ARTICLE INFO

Keywords:
Lie detection
Arousal
Thermal imaging
Skin temperature

ABSTRACT

People tend to be bad at detecting lies: When explicitly asked to infer whether others tell a lie or the truth, people often do not perform better than chance. However, increasing evidence suggests that implicit lie detection measures and potentially physiological measures may mirror observers' telling apart lies from truths after all. Implicit and physiological responses are argued to respond to lies as a threatening stimulus associated with a threat response. Subsequently, people who tell a lie should thus be liked and trusted less than those who tell the truth (indirect lie detection measures). In terms of physiology, a threat response should be associated with narrowing blood vessels (vasoconstriction), which should reduce peripheral skin blood flow. Consequently, we expected lower finger temperatures when confronted with a lie compared to the truth. We test lie detection using explicit and indirect measures, as well as using infrared thermal imaging as a physiological measure of lie detection. Participants ($N = 95$) observed videos of people telling lies or the truth about their social relationships, during which participants' fingertip temperature was recorded. Results suggested that the accuracy of explicit categorizations remained at chance level. Judgments of story-tellers' likability and trustworthiness (indirect measures of lie detection) showed no evidence that observers could tell apart liars and truthtellers: Those believed to be truthtellers were liked and trusted significantly more than those believed to be liars, even when this belief was mistaken. Physiological lie detection measured using thermal imaging also failed: Observers' fingertip temperatures did not significantly differ between lies and true stories. If at all, the temperature effects pointed in the opposite direction of the lies-as-threat expectations: Fingertip temperatures increased somewhat while confronted with lies compared to true stories. Results support the impression that people are bad at detecting lies, and cast doubt on whether fingertip temperature responses could be used as lie detection mechanisms.

1. Introduction

People lie. Be it using a white-lie or more serious deception, everyday interactions are laden with dishonesty. Evidence suggests that people lie one to two times per day on average (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996), and that for most people, a lie will slip out even in a short conversation (Tyler & Feldman, 2004; but see Halevy, Shalvi, & Verschuere, 2014). At the same time, people often want to know the truth about social interaction partners and their intentions, especially when others' outcomes may be improved by telling a lie. In outcome-interdependent situations, lies can create disadvantages for those who fail to detect them (e.g., Gneezy, 2005; Schweitzer & Croson, 1999). Therefore, people have to continuously assess whether they deem others' messages truthful or deceptive. It has even been argued that people are equipped with specialized cheater detection abilities (Cosmides, Barrett, & Tooby, 2010): Social contract theory (Cosmides & Tooby, 2000) suggests that in social situations, identifying cheaters (and liars) is an essential evolved specialization of the human mind.

But despite the prevalence and importance of detecting others' lies, people are surprisingly bad at catching liars. When asked to explicitly assess if a statement is a lie, observers' deception detection accuracy remains at chance level (e.g., at 54 % in Bond Jr & DePaulo, 2006, which was not significantly different from chance). In other words, human lie detection appears surprisingly ill suited for the potential importance of

* Corresponding author at: Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Straße 10, 53113 Bonn, Germany.
E-mail address: rahal@coll.mpg.de (R.-M. Rahal).

https://doi.org/10.1016/j.actpsy.2024.104548
Received 2 November 2023; Received in revised form 17 October 2024; Accepted 22 October 2024
Available online 19 November 2024

the task. In this study, instead of only relying on explicit categorizations of messages as lies and truths, we study if other measures of lie detection are more successful. We study indirect veracity judgments as well as observers' implicit lie detection ability mirrored in physiological arousal patterns measured via thermal infrared imaging. We tested if indirect and physiological indicators are better-than-chance predictors of whether participants were observing a lie or a true message.

The elusiveness of lies has led to vivid scientific interest in deception detection and its prerequisites. For instance, there was no evidence that explicit lie detection accuracy improved when people felt more confident in their veracity judgments (DePaulo, Charlton, Cooper, Lindsay, & Muhlenbruck, 1997) or that expertise improved performance (Burgoon, Buller, Ebesu, & Rockwell, 1994; but see Halevy et al., 2014). Even in close relationships, lie detection may be difficult (McCornack & Parks, 1986). Nevertheless, context information (Blair, Levine, & Shaw, 2010) and leaky (i.e., bad) liars (Bond Jr & DePaulo, 2008; Levine, 2010) may make it easier to catch a lie.

Given that explicit categorization into lies and true stories is generally rather unsuccessful, some avenues of research have explored if there are other, non-explicit observer reactions to the veracity of stimuli that have a better hit rate. This approach assumes that there may be reasons why explicit veracity judgments are tainted. It could be that cues for deception are not cognitively processed in a way that would produce awareness for them, or that higher-order processes such as impressions of the target and the situation or observer characteristics interfere. The idea that deception detection is hampered by some higher-order processes activated when explicit veracity judgments are undertaken is supported by findings suggesting that suppressing these competing processes improves deception detection. Concurrent cognitive load, effectively hindering interfering higher-order processes, improved deception detection (Albrechtsen, Meissner, & Susa, 2009; Feeley & Young, 2000). Making indirect judgments of veracity, e.g., by assessing the trustworthiness or likability of liars and truthtellers (DePaulo et al., 2003; Reinhard, Greifeneder, & Scharmach, 2013; van't Veer, Gallucci, Stel, & van Beest, 2015) also improved deception detection compared to explicit veracity judgments. Therefore, although explicit deception detection is a noisy process, at a more basic processing level, cues for the veracity of a signal could be detected and filtered into a less noisy signal. This approach is similar to the argument that visual attention does not necessarily produce explicit cognitive awareness for visual events (for an overview, see O'Regan & Noe, 2001), but could still cue physiological responses.

Similarly, deception detection may have physiological precursors with only limited trickle-down into explicit awareness. For instance, there is a specific set-up of neuro-cognitive functioning to detect errors. Signals of specific brain activation can be detected when actors themselves commit errors (Scheffers & Coles, 2000), or when they observe others committing errors (van Schie, Mars, Coles, & Bekkering, 2004). Detecting lies may share part of the functionalities of such basic error-detection systems. Further, deception detection may be related to anomaly detection. Although anomalous situations can produce physiological reactions, they may not seep into awareness to the degree where they can be explicitly pointed out. For instance, viewing playing cards where colors did not match suits (e.g., black hearts and red spades) led to increased pupil dilation compared to matching colors (Sleegers, Proulx, & van Beest, 2015) but not necessarily to explicit recognition of the anomaly (Bruner & Postman, 1949). Similarly, detecting a lie may cue physiological responses to socially anomalous, or at least undesired, behavior. Closely related is the idea that physiological responses may be cued in risky situations, which thereafter may bias decisions under risk (Bechara, Damasio, Tranel, & Damasio, 1997).

Because others' deception may have costly consequences for observers, lie detection may be related to the detection of threats. Physiological responses to threat detection involve sympathetic nervous system activation, triggering the release of adrenaline and noradrenaline (fight-or-flight response, Cannon, 1932). Among other effects, this fight-or-flight response leads to the narrowing of blood vessels in the extremities (vasoconstriction, Kistler, Mariauzouls, & von Berlepsch, 1998; Sokolov, 1963)]. Consequently, cutaneous blood microcirculation ebbs, which eventually cools down the skin. This effect has been demonstrated in response to horror movies (Kistler et al., 1998) and threatening personal questions (Rimm-Kaufman & Kagan, 1996). The lies-as-threat hypothesis suggests that responding to lies may trigger similar physiological mechanisms as responding to threats. Supporting this hypothesis, van't Veer et al. (2015) provided first evidence of fingertip temperature responses when exposed to deception: When viewing lies compared to true statements, fingertip temperatures decreased in participants aiming to detect lies.

Based on this research, it appears that humans may be able to detect lies after all, although not explicitly. The promise (and threat) of non-explicit mechanisms to detect lies has roused substantial scientific and public interest. If lies could be identified more reliably, social institutions and public security measures could be improved and societies may benefit from an atmosphere of fairness and trust. At the same time, the potential for misuse is substantial (see Honts, Thurber, & Handler, 2021 for commentary in the scope of the polygraph test). Preceding practical applications of non-explicit lie detection mechanisms, a broadly reliable basis of scientific evidence is needed. The current study adds to this goal by comparing explicit, implicit and physical lie detection measures. In particular, we introduce thermal imaging cameras as a novel tool for physiological lie detection research and application.

1.1. Present research

To further study whether lies can be "detected" in physiological responses even when only limited (or no) awareness of veracity exists, we studied observers' responses to lies vs. true stories in fingertip skin temperature via thermal imaging. Building on the lies-as-threat hypothesis, and aiming to replicate the findings of van't Veer et al. (2015), we propose that observers' responses to deception may be mirrored in skin temperature fluctuations: observing lies is expected to lead to decreased finger temperatures compared to observing true stories (preregistered at https://aspredicted.org/8rj9-t22x.pdf).

Whereas van't Veer et al. (2015) studied physiological responses to lies vs. true stories both when participants knew their task was to detect lies ("forewarned" condition) and when they were not explicitly made aware that the story they had watched might be a lie ("not forewarned" condition), the present study focuses only on assessing responses to lies vs. true stories while participants are aware of the possibility that they will encounter a lie. In this "forewarned" setting, van't Veer et al. (2015) showed that fingertip temperatures were lower after observers were exposed to a lie than to a true story. We aimed to replicate the result that fingertip temperatures drop after observing lies compared to true stories.

Further, this study departs from van't Veer et al. (2015) in two general aspects: improving the technology used to measure skin temperature and the stimulus material. First, our aim was to explore the potential of using thermal imaging cameras in lie detection research to improve the measurement of skin temperature as it unfolds. This technique allows continuous high-resolution recordings of skin temperature (Kistler et al., 1998; Pavlidis et al., 2012). Therein, tracking affective processes underlying behavioral outcomes becomes possible in an unobtrusive manner. Affective measures such as the Galvanic Skin Response or the iButton used in van't Veer et al. (2015) require participants to make physical contact with the measurement device at all times. In contrast, infrared thermal imaging can be used non-invasively, i.e., without interfering with the participants' body, reducing potential confounds. Further, infrared thermal imaging does not rely on decoding facial behavior to infer affect (for a discussion of shortcomings of inferring emotions from facial behavior, see Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019). Instead, infrared thermal imaging relies on

tracing thermal responses of the skin sparked by autonomous nervous system reactions, which are comparatively direct and difficult for participants to consciously interfere with. This makes infrared thermal imaging a promising technology for tracking affective processes generally, and in particular to assess autonomous nervous system responses. Here, we use it to measure the expected drop in peripheral skin temperature following a threat response while observing lies vs. true statements.

Moreover, whereas van't Veer et al. (2015) created a small set of novel materials, we used a large set of standardized and freely available stimulus material. Specifically, we make use of the Miami University Deception Detection Database (MU3D, see Lloyd et al., 2018), a standardized set of videos with short stories about people whom the storyteller likes or dislikes, told truthfully or as a lie (like-as-though-dislike and dislike-as-though-like). In addition to having been scored for a number of control variables, the number of videos available offers the opportunity to use a within-subjects repeated measures design, as well as to profit from the advantages of drawing random items for each participant from the larger pool of available stimuli (for an overview, see de Boeck, 2008). By using the MU3D, we deviate from van't Veer et al. (2015) in that storytellers do not tell a story about themselves, but describe another person (following the person-description paradigm (see DePaulo & Rosenthal, 1979). This shift in message contents aligns our study with applications to witness testimony about others, but may limit applications of detecting lies about personal accounts.

Departing again from van't Veer et al. (2015), the present study incentivized participants to catch liars: For correctly identifying truths and lies, participants could earn a monetary reward (lump sum payment if all stories are correctly classified). We chose to incentivize participants' performance to increase their attention and motivation to unmask the stimulus material, and consequently to increase the hit rate. Nevertheless, this deviation from the original study could also backfire, interfering with participants' lie detection abilities by increasing the perceived importance of the task.

Lastly, in this study, Dutch and international respondents are exposed to English stimuli, while van't Veer et al. (2015) used Dutch stimuli for Dutch-speaking participants. While participants in research conducted at Tilburg University are used to taking part in research presented in English and generally have an excellent command English, it is possible that lie detection is more difficult in a foreign language. Moreover, storytellers were American and participants were from mostly European backgrounds, such that differences in mannerisms of lying or cues used for lie detection may further hamper participants' performance. Additional, minor deviations are discussed in the materials section.

Despite these deviations and their potential impact on the results, we expected to conceptually replicate van't Veer et al.'s (2015) main result in the "forewarned" condition: that fingertip temperatures would drop in response to exposure with a lie compared to a true story. In addition to this main, preregistered hypothesis, we further studied whether hit rates would improve when considering indirect veracity judgments. As in van't Veer et al. (2015), we expected that perceived liking and trustworthiness of storytellers would differ between lies and true stories told, but expected no deviation from chance levels in the hit rate of explicit veracity judgments.

2. Method

The study was preregistered, and data, materials and code are available at https://doi.org/10.17605/OSF.IO/RGAWF.

2.1. Participants and design

We ran a 2 (veracity: true story vs. lie) within-subjects design, where in each of 8 target trials, participants were randomly confronted with one of two veracity conditions: a lie or a true story (for an overview of the procedure, see Fig. 1).

As this is a conceptual replication of van't Veer et al. (2015) we based our sample size planning on this paper. van't Veer et al. (2015) sampled 155 participants for a mixed design with two trials per participant. However, we replicate only one of the between-subjects conditions and expose participants to four truths and four lies. Therefore, we required fewer participants.

The target sample size was determined based on a power analysis in G*Power (Faul, Erdfelder, Buchner, & Lang, 2009). We assumed a medium-sized temperature difference between veracity conditions, triangulating between the null effect reported in van't Veer et al. (2015) and the large change in fingertip temperature following threat perceptions in Kistler et al. (1998). Given feasibility concerns regarding data collection by one experimenter with only one thermal imaging camera and a planned experimental duration of 1 h, we assumed that we could sample about 80 participants, which would allow us to detect a medium-sized temperature difference between the veracity conditions of $d = 0.28$ in a paired, one-sided t-test with 80 % power at $\alpha = 0.05$. This analysis disregards the multilevel-structure of the design, which would warrant a higher power estimate. To allow for potential technical failure only uncovered after processing the data, we included an additional 15 participants and preregistered the desired sample size of 95 participants. In total, data from 96 participants was recorded because data from one participant was unusable due to technical malfunctions detected during the data collection period. The final sample size of 95 participants matched the preregistered target sample size.

Participants (81 female, $M_{age} = 20.26$, $SD_{age} = 1.76$, 79 right-handed) were recruited among the first-year psychology students at Tilburg University, and received course credit. In addition, participants were incentivized to correctly identify the veracity of the videos, such that if they identified all videos correctly, they would earn €25. Based on guessing alone, the probability of correctly identifying all 8 videos would be about 0.4 %. No participant qualified for receiving this monetary incentive.

2.2. Procedure

The study received ethics approval from the review board at Tilburg University (reference number EC-2018.114). Upon arrival in the lab, participants gave informed consent.

Participants were seated in a cubicle equipped with the infrared thermal imaging camera (FLIR A655sc, FLIR®Systems, Inc., 640 × 480 pixels, thermal sensitivity <0.03 °C; set to the maximum recording frequency of 30 fps and to emissivity appropriate for human skin (0.98), operated via ResearchIR), where they placed their non-dominant hand on an armrest that allowed them to effortlessly keep their fingers extended (see Fig. 2). Thermal data was collected with the camera pointed to the palm of participants' hands, to avoid interference of nails and hair (Fernández-Cuevas et al., 2015), and with participants' hands positioned under the table on which the computer workstation used for stimulus presentation (via Inquisit 4) was mounted, such that the wooden desk provided a stable background image.

2.2.1. Acclimatization

To ensure that participants' body temperature reached a resting state, participants underwent an acclimatization phase (Ioannou, Gallese, & Merla, 2014). To bridge the time in a controlled manner, they watched a video documentary about Singapore for 7 min and 48 s, which was informative, expected to be affectively neutral and unrelated to the subsequent task. Then, to obtain baseline temperature measurements, participants were asked to view a neutral video of a screensaver in which lines moved vertically across the screen (see Fig. 3) for 60 s. We chose this filler video to allow participants to continue engaging with the study during the temperature measurements without acting on their affective states.

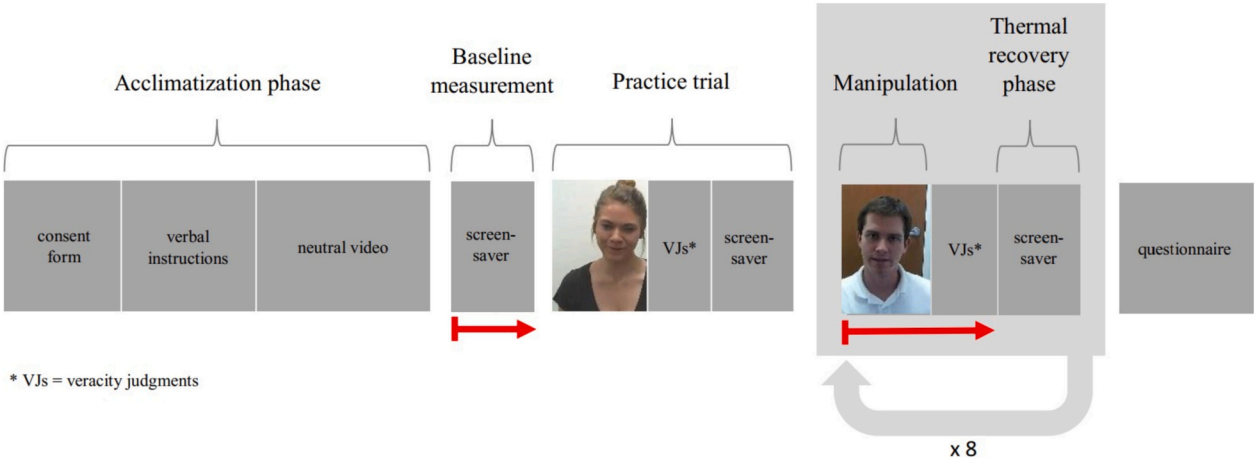


Fig. 1. Overview of the procedure, with red arrows indicating infrared thermal image recording. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 2. Position of the hand on the armrest, while the palm was filmed from below.

2.2.2. *Veracity judgments*

Next, participants were introduced to the veracity judgment task. They underwent one practice trial without thermal imaging recordings and eight target trials during which fingertip temperatures were recorded. In each trial, participants watched one randomly drawn video from the MU3D (Lloyd et al., 2018), which showed a true story or a lie (veracity condition) about how the storyteller felt about a third person. The MU3D consists of 320 videos with 80 storytellers. Here, we only used videos from Caucasian storytellers to match participant demographics, at the potential cost of generalizability. In total, we used a subset of 160 videos from 40 storytellers. On average, the videos lasted 35.86 s (SD = 3.68) and contained 110.92 spoken words (SD = 21.81). In the practice trial, participants randomly saw a female or male storyteller, who told a positive or negative story about a third person, which was either a lie or true story. For the target trials, eight videos were randomly selected from the pool of storytellers who had not been shown in the practice trial, while maintaining an equal number of truths and lies, positive and negative stories, and male and female storytellers, as well as not showing any storyteller more than once. Starting with the onset of the video presentation, temperature data was recorded for 50 s (i.e., longer than the duration of some videos), allowing enough lag for the relatively slow-paced change in cutaneous temperature to unfold (Kistler et al., 1998).

Then, participants were confronted with three questions about each of the target videos, as in van't Veer et al. (2015). We included one *direct question* ("Do you think the story was true?"; yes / no) and two *indirect questions*, one focusing on *liking* ("How much did you like the person who told the story?"; 1 = not at all, 7 = a lot) and the other on *trustworthiness* ("How trustworthy do you think the person who told the story is?"; 1 = not at all, 7 = a lot). The order in which these questions were presented varied randomly between subjects to rule out order effects. This deviates from van't Veer et al. (2015), where the order of these questions was kept constant (liking, trustworthiness, direct veracity judgment).

After each target video, to allow for temperature levels to reset, participants re-watched the filler video of moving lines for 90 s. Since the lag phase of fingertip responses is about 15 s (Kistler et al., 1998), we expected this reset time to be sufficient without unnecessarily prolonging the overall experiment duration. To avoid restlessness, participants were also shown a countdown timer during this recovery period.

2.2.3. *Postexperimental questionnaire*

Participants indicated their age, sex, ability to speak English and whether English was their native language. Finally, we collected data on additional variables potentially affecting body temperature (see Fernández-Cuevas et al., 2015, e.g., physical strain in the recent past, use of hormonal contraceptives, for full materials and data, see online materials). Participants were then debriefed and thanked.

2.3. *Data preprocessing*

Data preprocessing was conducted in multiple steps (which are documented in detail on the OSF). To make the data analyzable, we converted the video recordings of the temperature into individual frames saved as grayscale images with a procedure adapted from

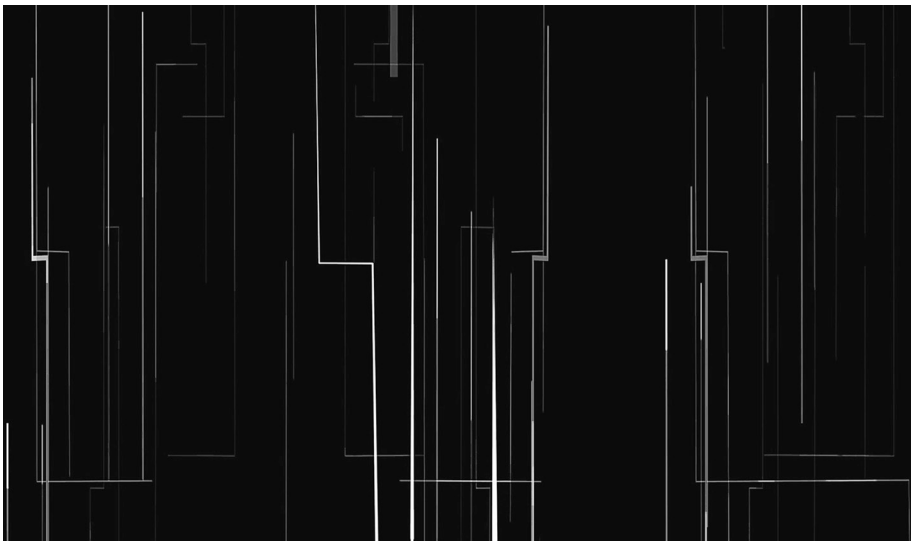


Fig. 3. Frame from the neutral video of geometric shapes showed in the acclimatization phase and as a filler to allow skin temperatures to reset after each target video.

Tattersall (2017). Departing from the preregistration, we then downsampled the data due to processing capacity constraints of the large data files obtained, retaining per trial only every 50th frame (i.e., one temperature recording per 1.67 s, where the total recording duration was 50 s), as well as the first and second to last frame yielding a total of 32 temperature observations per trial). On the remaining images, a background detection algorithm was run, removing data from the wooden desk recorded in the background. Then, an algorithm detecting the fingertips was run, discarding data outside of these defined regions of interest. Finally, we aggregated the data of the five fingertips, retaining information about the minimum, mean and maximum temperature. In this process, 0.42 % of the observations failed to parse and could therefore not be analyzed.

Further, following the criteria of van't Veer et al. (2015) in accordance with the preregistration, we excluded trials based on extreme temperatures (below 18 °C or above 37 °C, 0.01 % of the data).

3. Results

To account for the multilevel structure of the design, where responses were nested within subjects and within videos, we used multilevel models to analyze responses to viewing lies compared to true stories, with random effects for participants, trials and videos. With this strategy, we deviated from the preregistered *t*-tests, in which we would have compared average temperature changes from baseline between lie and truth trials on the participant level only. However, we argue that interpreting the results of the multilevel model is preferable, because these more sophisticated statistical models can take advantage of the repeated measures structure of the data and models the underlying variability accordingly. Further, we had already anticipated this deviation in the preregistration, where we outlined that mixed models would be used to assess the data.

For each hypothesis, we report the critical χ^2 test comparing the full model (with the effect in question) against the null model (without the effect in question). When this comparison suggests a significant difference, we conclude that the effect of interest is significant. We additionally report the estimate of the effect of interest in the full model.

3.1. Direct veracity judgments

First, we analyzed if participants could explicitly point out liars by assessing if participants were more likely to rate a true story as true or as a lie, depending on whether the story was actually true or a lie. A logistic multilevel model with participant-, trial- and video-level random effects indicated no evidence that adding actual veracity to the model as a fixed effect improved model fit ($\chi^2(1) = 1.24, p = 0.27$). There was no evidence that participants categorized true stories as true more often than lies ($OR = 1.81, z = 1.15, p = 0.25$, see Fig. 4). Participants on average only categorized 50.00 % ($SD = 19.04\%$) of videos correctly. Therefore, explicit lie detection largely failed and the hit rate remained on chance level, in line with expectations from the extant literature.

3.2. Indirect veracity judgments: liking and trustworthiness

3.2.1. Depending on veracity

To assess whether indirect veracity judgments would be more favorable, i.e., show higher liking and trustworthiness ratings, when participants saw a true story compared to a lie, we ran linear multilevel models with participant, trial and video random effects. Conditioning on whether stories were actually true or false did not improve model fit regarding rated liking ($\chi^2(1) = 0.13, p = 0.72$), or regarding rated trustworthiness ($\chi^2(1) = 3.11, p = 0.08$). Compared to liars, people who told true stories were rated as somewhat more likable (0.05 points ± 0.13 (standard errors), $t(149.07) = 0.36$, $p = 0.72$) and trustworthy (0.22 points ± 0.13 (standard errors), $t(149.76) = 1.77$, $p = 0.08$). Albeit not statistically significant, the direction of both effects was

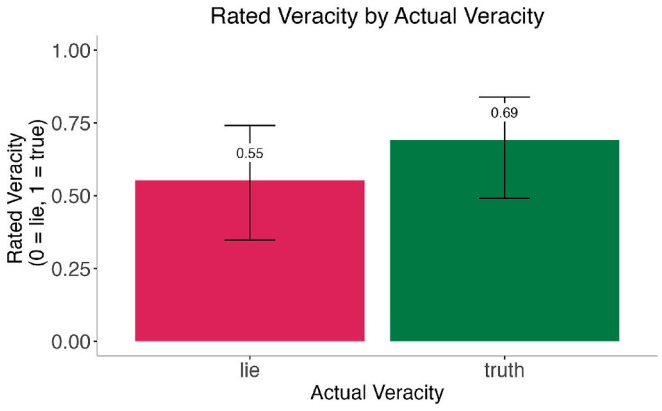


Fig. 4. Rated veracity depending on actual video veracity with 95 % confidence intervals.

consistent with the findings in van't Veer et al. (2015).

3.2.2. Depending on veracity and categorization success

Because lie detection success was low (50.00 % (SD = 19.04%) of videos correctly categorized), we additionally compared the indirect veracity judgments to the videos' veracity separately for correctly identified videos and for videos that were misidentified. We ran analyses investigating indirect veracity judgments after successful lie detection or correct recognition of the truth (which we refer to as "hit" trials) vs. unsuccessful categorization of the veracity of the video (which we refer to as "miss" trials).

Running linear multilevel models with participant, trial and video random effects, we tested whether the effects of the veracity condition on liking and trustworthiness differed between correctly and incorrectly categorized trials. Suggesting that it mattered for rated liking and trustworthiness whether participants had guessed correctly if the video in question contained a lie or a true story, results showed evidence for an interaction effect between veracity condition and categorization correctness (liking: $\chi^2(1) = 3,199.47, p < .001$, trustworthiness: $\chi^2(1) = 6,313.07, p < .001$).

Specifically, in hit trials, conditioning on the type of video shown improved model fit both regarding rated liking ($\chi^2(1) = 32.91, p < .001$) and trustworthiness ($\chi^2(1) = 37.76, p < .001$). In correctly identified true stories, participants reported finding the target more likable (0.98 points ± 0.16 (standard errors), $t(129.07) = 6.12$, $p < .001$, Marginal $R^2 = 0.09$) and trustworthy (1.11 points ± 0.17 (standard errors), $t(131.34) = 6.62$, $p < .001$, Marginal $R^2 = 0.12$) than in correctly identified lies (see Fig. 5, Panels A and C). In miss trials, conditioning the type of video shown also improved model fit, both regarding rated liking ($\chi^2(1) = 29.39, p < .001$) and trustworthiness ($\chi^2(1) = 66.71, p < .001$). Participants reported finding the target less likable (0.98 points ± 0.17 (standard errors), $t(138.97) = -5.72$, $p < .001$, Marginal $R^2 = 0.09$) and trustworthy (1.42 points ± 0.15 (standard errors), $t(137.89) = -9.25$, $p < .001$, Marginal $R^2 = 0.19$) in true stories mistakenly identified as lies compared to false stories mistakenly identified as truths (see Fig. 5, Panels B and D).

Therefore, the results qualified the findings in van't Veer et al. (2015), suggesting that indirect lie detection mechanisms mirrored in evaluations of the storyteller depend on categorization success. Participants liked and trusted people more whom they *believed* to be truthtellers compared to people whom they *believed* to be liars. This was the case even whether the people *believed* to be truthtellers were not, in fact, truthtellers. The observed effects, translating to shifts of about 1 point on a 7-point Likert scale, indicate small yet non-negligible changes in liking and perceived trustworthiness depending on *believed* veracity.

3.3. Implicit lie detection: finger temperature

3.3.1. Sanity check: differences at recording onset

We assumed that there would be no temperature differences between videos containing true stories and lies at the onset of each video (i.e., in the first frame recorded via thermal imaging[1]). This should be the case since the presentation order of true stories vs. lies was determined randomly, so that participants could not anticipate which type of story would be shown. As a sanity check, we assessed if there were temperature differences between conditions at the onset of the recordings. A linear multilevel model with participant, trial and video random effects indicated no evidence that the fingertip temperature differed between true stories and lies at recording onset ($\chi^2(1) = 0.02, p = 0.88$, 0.01 degrees ± 0.07 (standard errors), $t(95.22) = 0.15$, $p = 0.88$). The estimated Bayes factor in favor of the null model over the full model was 9.16.

3.3.2. Depending on veracity

We had hypothesized that finger temperatures while viewing a lie compared to a true story would decrease, mirroring an implicit, physiological lie detection ability. As preregistered, following the argument of Kistler et al. (1998) that it takes approximately 15 s for thermal responses to manifest, data of the first 15 s (i.e., the first 10 temperature observations) of thermal recording was excluded from the analyses corresponding to this hypothesis.

In linear multilevel models with participant, trial and video random effects and predicting finger temperature, results showed no evidence that adding the veracity of the video shown increased model fit ($\chi^2(1) = 0.61, p = 0.44$).[2] If at all, the data suggested that, in contrast to the original hypotheses, observing lies compared to true stories increased finger temperatures by about 0.06 degrees ± 0.08 (standard errors), $t(148.07) = 0.78$, $p = 0.44$, (see Fig. 6). Therefore, the results failed to replicate the findings in van't Veer et al. (2015). The estimated Bayes factor in favor of the null model over the full model was 5.74.

3.3.3. Depending on veracity and categorization success

As for indirect veracity judgments, we compared the finger temperature response to the videos' veracity depending on whether videos were correctly identified (hit trials) or misidentified (miss trials). As for the previous analysis, data of the first 15 s (i.e., the first 10 temperature observations) of thermal recording was excluded from this analysis. Running multilevel models with participant, trial and video random effects, there was no evidence that the effect of the veracity condition on temperature differed between correctly and incorrectly categorized trials, showed by the absence of evidence for an interaction effect between veracity condition and categorization correctness ($\chi^2(1) = 3.78, p = 0.052$). There was no evidence that the type of video shown predicted finger temperature in hit trials ($\chi^2(1) = 0.22, p = 0.64$) or in miss trials ($\chi^2(1) = 0.00, p = 0.95$). The temperature trend pointed in the same direction as for the analyses reported above, in that observing correctly identified lies (0.07 degrees ± 0.14 (standard errors), $t(127.84) = -0.46$, $p = 0.64$, see Fig. 7, Panel A) and incorrectly identified lies (0.01 degrees ± 0.13 (standard errors), $t(132.46) = -0.07$, $p = 0.95$, see Fig. 7, Panel B), if at all, increased fingertip temperatures compared to observing true stories.

4. Discussion

It has long been established that people are bad at detecting lies, and that hit rates for direct veracity judgments rarely deviate from chance (e.g., Bond Jr & DePaulo, 2006). In line with this expectation, our study showed no evidence that lie detection accuracy deviated from chance levels. Corroborating prior literature, we found a hit rate of 50 %, underscoring the assertion that direct lie detection often fails.

But other research had raised hopes that humans could detect lies indirectly (DePaulo et al., 2003; Reinhard, Greifeneder, & Scharmach, 2013; van't Veer et al., 2015), for example via judgments of liking and trustworthiness of liars vs. truthtellers, or that at least implicit, physiological traces of lie detection abilities could be found (see van't Veer et al., 2015). In this preregistered study, we find no evidence of such

[1] Note that we mean the first frame recorded overall, beginning at trial onset, i.e. after 0 s had elapsed.

[2] Using the preregistered paired t-test, there was also no evidence for a temperature difference between conditions ($M_{truth} = 29.34$, $SD_{truth} = 4.02$, $M_{lie} = 29.38$, $SD_{lie} = 4.03$, $M_D = -0.04$, 95 % CI [−0.19, 0.11], $t(93) = -0.52$, $p = .601$). As a robustness check, we further assessed if there were temperature differences between conditions in the first trial. A linear multilevel model with participant and video random effects indicated no evidence that the fingertip temperature differed between true stories and lies in the first trial participants engaged with ($\chi^2(1) = 0.68$, $p = 0.41$, 0.76 degrees ± 0.89 (standard errors), $t(55.95) = -0.85$, $p = 0.4$).

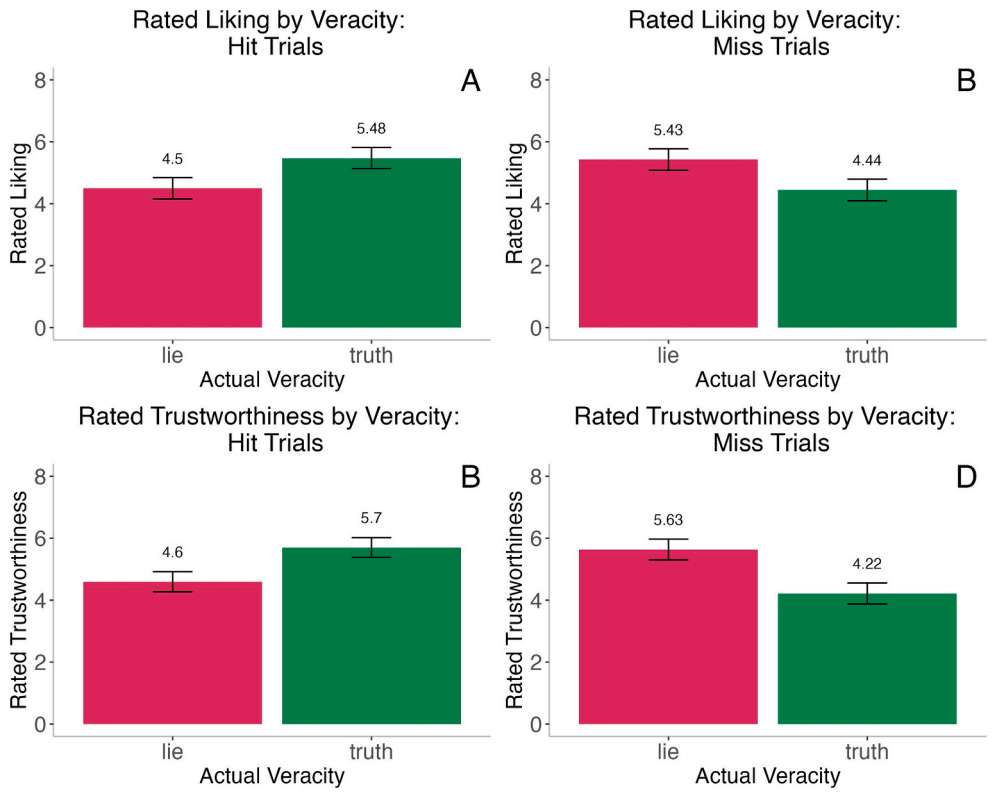


Fig. 5. Rated liking and trustworthiness conditioned on whether participants correctly categorized the video veracity (Panel A and C, hit trials) or did not correctly identify the videos' veracity (Panel B and D, miss trials), for true stories vs. lies, with 95 % confidence intervals.

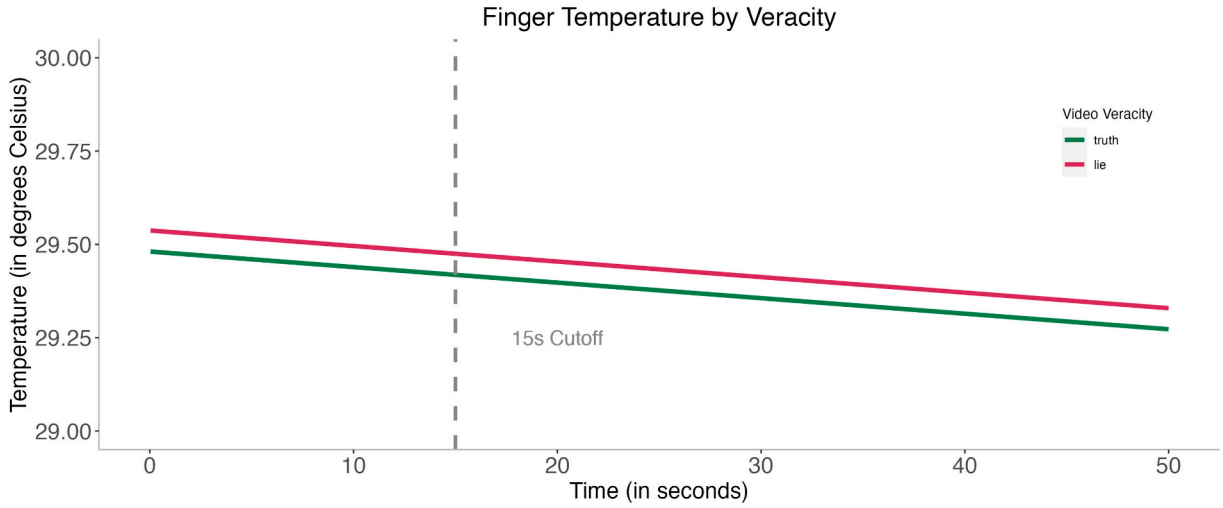


Fig. 6. . Temperature trajectories conditioned on the videos participants viewed. The vertical line indicates the demarcation between the first 15 seconds (excluded from main analyses) and the final 35 seconds (included in main analyses).

indirect or implicit lie detection abilities.

In line with these expectations, we had hypothesized that deception detection would be apparent in indirect judgments of veracity, when assessing the likability and trustworthiness of liars and truthtellers. Qualifying previous findings from van't Veer et al. (2015), indirect measures of veracity judgments showed that participants had more positive evaluations of storytellers whom they *believed* to be truthtellers, even if this assessment was mistaken.

Procedural differences between the study reported here and in van't Veer et al. (2015) could lie at the root of this difference in results: In our study, observers were incentivized to catch liars, as they could earn €25 for correctly guessing the veracity of all presented stories, while van't Veer et al. (2015) used an unincentivized task. Observers could have attended to the incentivized stimuli differently, becoming overconfident (see Lebreton et al., 2018) and consequently overcorrecting their indirect veracity judgments. At the same time, participants may have been more motivated to catch liars, investing greater effort in the endeavor, which could either improve or reduce detection accuracy. Nevertheless, the hit rate of 50 % is in line with expectations from the literature, leading us to assume that the influence of introducing an incentive was negligible for explicit detection performance.

However, it is also possible that these results are driven by

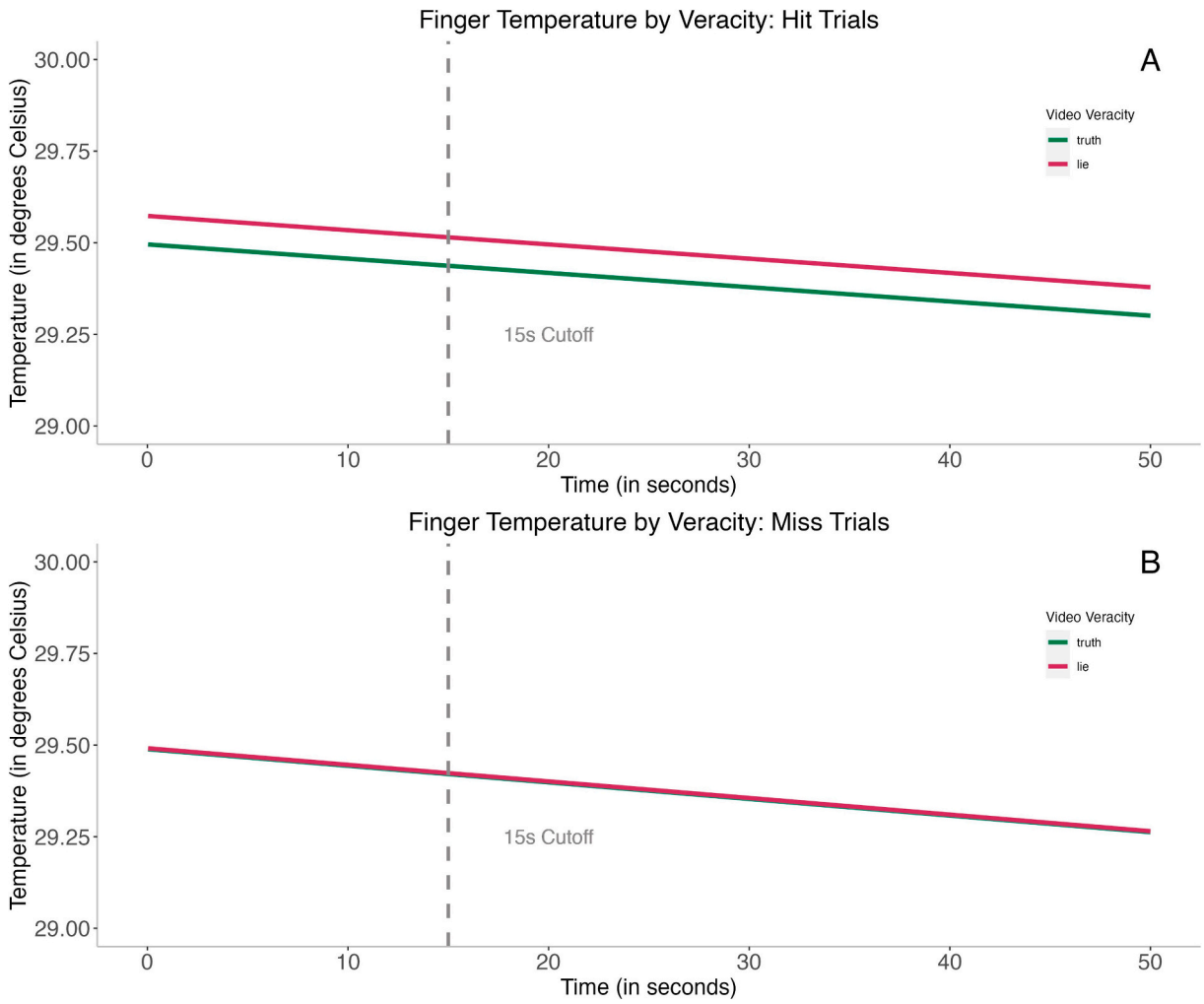


Fig. 7. Temperature trajectories conditioned on whether participants correctly categorized the video veracity (Panel A, hit trials) or did not correctly identify the videos' veracity (Panel B, miss trials), for true stories vs. lies. The vertical line indicates the demarcation between the first 15 s (excluded from analyses) and the final 35 s (included in analyses).

confirmation bias (Nickerson, 1998): Individuals may have judged storytellers as liars (or truthtellers) and interpreted the ambiguous evidence available about their likability and trustworthiness accordingly. Therefore, this result may point to the potential of abusing people's limited ability to distinguish liars and truth tellers, and their willingness to bestow a positive social image on others based on subjective ascriptions of truthfulness. Even if someone is not objectively a truthteller, leading others to *believe* that they are can render them socially more successful, based on the increased trust and likability they receive. Social outcomes, therefore, are not only determined by actual truthfulness but also by individual's ability to create and maintain a subjective impression of truthfulness. Individuals may strategically present themselves as truthtellers to garner trust and favor in various social contexts, such as professional settings, interpersonal relationships, or public interactions.

Further, results in this study showed no evidence for implicit, physiological responses to liars compared to truthtellers. The temperature of observers' fingertips was not found to systematically mirror the veracity of the stories that they saw: Even though fingertip temperatures slightly increased when viewing lies compared to true stories, the effects did not reach statistical significance. These results also held when conditioning fingertip temperature responses on whether participants' beliefs about the observed stories' veracity was correct. In both hit trials, i.e. in trials where participants correctly identified the veracity of the stories they had observed, and in miss trials, we found the same pattern: Both correctly and incorrectly identified true stories were associated with higher fingertip temperatures than correctly and incorrectly identified lies. With regard to implicit, physiological lie detection, the results therefore did not replicate the findings of van't Veer et al. (2015).

Again, differences in the procedure may be a reason why the previous results could not be replicated: While we used a standardized set of stimulus materials presented in English, relying on videos widely used in the lie detection literature (Lloyd et al., 2018) which were shown to a sample of Dutch and international students, van't Veer et al. (2015) relied on a self-created set of stimulus materials that was specifically tailored to Dutch participants and shown to Dutch participants only. It may be the case that physiological lie detection depends on the match of stimulus material and sample in terms of native language and cultural context. However, the literature is largely mute on the specific role of these variables for physiological lie detection responses. Regarding explicit lie detection, there is limited evidence and conflicting evidence about how familiarity could improve explicit lie detection accuracy (see Feeley, deTurck, & Young, 1995; Reinhard, Sporer, & Scharmach, 2013) and that lying in a second language tends to be less successful (e.g., Cheng & Broadhurst, 2005; Elliott & Leach, 2016). Whether assessing native's statements in one's second language obstructs lie detection accuracy is unclear (but see Snellings, 2013 for a null result), although hit rates could be improved through a reduction of overconfidence (DePaulo et al., 1997). Whether the match in terms of language and culture matters for lie detection - both explicit, implicit and

physiological - remains to be clarified. The same holds for assessing if the diverging results could be explained by introducing an incentivization mechanism (see discussion above).

Further, results might diverge from van't Veer et al. (2015) due to the difference in devices used to measure fingertip temperature. While van't Veer et al. (2015) relied on the iButton, which requires participants to touch the device, we assessed fingertip temperatures via remote thermal imaging cameras. In principle, it would be possible that being required to touch the device directly alters participants' (perceived) affective states, potentially affecting the results. Similarly, it would be possible that analytic strategy of assessing the temperature of not only one (as in van't Veer et al., 2015) but all five fingertips leads to differences in results. However, we regard these potential explanations as improbable.

A stronger interpretation of the data in this study would suggest that the results even speak against the lies-as-threat hypothesis. We had hypothesized that lie detection is akin to threat detection and therein triggers an autonomous nervous system response leading to peripheral vasoconstriction (Kistler et al., 1998). Fingertip temperature, as studied in the present research, would be the last link in the chain reaction set into motion during a potential threat response to lies. If this was the case, vasoconstriction triggered by a threat response should lead to decreased fingertip temperatures when observing lies compared to true stories. But when exposing participants to lies, we found a tendency for observers' finger temperature to rise, rather than to drop. This finding may point to different physiological mechanisms involved in response to lies, accompanied by increased blood flow in the extremities. A rise in fingertip temperatures could indicate increased alertness (Vergara, Moënne-Loccoz, Ávalos, Egaña, & Maldonado, 2019; e.g., Vergara, Moënne-Loccoz, & Maldonado, 2017), which may provide a basis for formulating alternative theories of channels by which exposure to lies affects physiological responses through task engagement.

Finally, the null finding could also be interpreted to suggest the absence of an effect of exposure to lies on fingertip temperatures. While a definitive conclusion about the (presence of) physiological mechanisms of lie detection cannot be obtained from the present research, it serves to demonstrate that work on understanding the mechanisms by which cutaneous temperature might reflect affective responses is still in its infancy. Infrared thermal imaging opens a promising avenue for future research by allowing relatively unobtrusive, high-resolution recording of temperature fluctuations. If successfully mapped to affective responses, thermal imaging could be used for tracking affective processes as they unfold not only in lie detection scenarios, but also in other social decision making settings.

In sum, this study provides no evidence that people can catch liars. We found no evidence that observers could explicitly differentiate liars from truthtellers, that observers rated truthtellers as more likeable and trustworthy than liars, or that physiological precursors of lie detection could successfully be identified in observers. Results also cast doubt on the lies-as-threat hypothesis, because fingertip temperatures dropped somewhat in response to true stories, not in response to lies as expected. This work challenges the belief that lies can be accurately detected, if not explicitly then at least through implicit channels. Consequently, while further research is needed to explore other potential lie detection mechanisms and to replicate and substantiate existing attempts, practitioners should critically evaluate the methods and techniques currently employed for detecting deception.

CRediT authorship contribution statement

Rima-Maria Rahal: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Teun Siebers:** Writing – original draft, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Willem W.A. Sleegers:** Writing – original draft, Visualization, Software, Resources, Methodology, Formal analysis, Data curation. **Ilja van Beest:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and materials are shared on the Open Science Framework.

References

Albrechtsen, J. S., Meissner, C. A., & Susa, K. J. (2009). Can intuition improve deception detection performance? *Journal of Experimental Social Psychology, 45*(4), 1052–1055. https://doi.org/10.1016/j.jesp.2009.05.017

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*. https://doi.org/10.1177/1529100619832930

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5304), 1293–1295. https://doi.org/10.1126/science.275.5304.1293

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research, 36*(3), 423–442.

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

Bond, C. F., Jr., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin, 134*(4), 477.

Bruner, J. S., & Postman, L. (1949). On the perception of incongruity: A paradigm. *Journal of Personality, 18*(2), 206–223.

Burgoon, J. K., Buller, D. B., Ebesu, A. S., & Rockwell, P. (1994). Interpersonal deception: V. Accuracy in deception detection. *Communication Monographs, 61*(4), 303–325.

Cannon, W. B. (1932). *Wisdom of the body*. New York, NY, USA: W.W. Norton & Company, Inc.

Cheng, K. H. W., & Broadhurst, R. (2005). The detection of deception: The effects of first and second language on lie detection ability. *Psychiatry, Psychology and Law, 12*(1), 107–118. https://doi.org/10.1375/pplt.2005.12.1.107

Cosmides, L., Barrett, H. C., & Tooby, J. (2010). Adaptive specializations, social exchange, and the evolution of human intelligence. *Proceedings of the National Academy of Sciences, 107*(Supplement 2), 9007–9014. https://doi.org/10.1073/pnas.0914623107

Cosmides, L., & Tooby, J. (2000). The cognitive neuroscience of social reasoning. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 1259–1270). Cambridge MA, USA: MIT Press.

de Boeck, P. (2008). Random item IRT models. *Psychometrika, 73*(4), 533–559. https://doi.org/10.1007/s11336-008-9092-x

DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review, 1*(4), 346–357.

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

DePaulo, B. M., & Rosenthal, R. (1979). Telling lies. *Journal of Personality and Social Psychology, 37*(10), 1713.

Elliott, E., & Leach, A.-M. (2016). You must be lying because I don't understand you: Language proficiency and lie detection. *Journal of Experimental Psychology: Applied, 22*(4), 488–499. https://doi.org/10.1037/xap0000102

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160.

Feeley, T. H., deTurck, M. A., & Young, M. J. (1995). Baseline familiarity in lie detection. *Communication Research Reports, 12*(2), 160–169. https://doi.org/10.1080/08824099509362052

Feeley, T. H., & Young, M. J. (2000). Self-reported cues about deceptive and truthful communication: The effects of cognitive capacity and communicator veracity. *Communication Quarterly, 48*(2), 101–119. https://doi.org/10.1080/01463370009385585

Fernández-Cuevas, I., Marins, J. C. B., Lastras, J. A., Carmona, P. M. G., Cano, S. P., García-Concepción, M.Á., & Sillero-Quintana, M. (2015). Classification of factors influencing the use of infrared thermography in humans: A review. *Infrared Physics & Technology, 71*, 28–55.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review, 95*(1), 384–394. https://doi.org/10.1257/0002828053828662

Halevy, R., Shalvi, S., & Verschuere, B. (2014). Being honest about dishonesty: Correlating self-reports and actual lying. *Human Communication Research, 40*(1), 54–72. https://doi.org/10.1111/hcre.12019

Honts, C. R., Thurber, S., & Handler, M. (2021). A comprehensive meta-analysis of the comparison question polygraph test. *Applied Cognitive Psychology, 35*(2), 411–427. https://doi.org/10.1002/acp.3779

Ioannou, S., Gallese, V., & Merla, A. (2014). Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology, 51*(10), 951–963.

Kistler, A., Mariauzouls, C., & von Berlepsch, K. (1998). Fingertip temperature as an indicator for sympathetic responses. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology, 29*(1), 35–41. https://doi.org/10.1016/s0167-8760(97)00087-1

Lebreton, M., Langdon, S., Slieker, M. J., Nooitgedacht, J. S., Goudriaan, A. E., Denys, D., et al. (2018). Two sides of the same coin: Monetary incentives concurrently improve and bias confidence judgments. *Science Advances, 4*(5), Article eaaq0668. https://doi.org/10.1126/sciadv.aaq0668

Levine, T. R. (2010). A few transparent liars explaining 54% accuracy in deception detection experiments. *Annals of the International Communication Association, 34*(1), 41–61.

Lloyd, E. P., Deska, J. C., Hugenberg, K., McConnell, A. R., Humphrey, B. T., & Kunstman, J. W. (2018). Miami university deception detection database. *Behavior Research Methods*, 1–11.

McCornack, S. A., & Parks, M. R. (1986). Deception detection and relationship development: The other side of trust. *Annals of the International Communication Association, 9*(1), 377–389.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

O'Regan, J. K., & Noe, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939–973. https://doi.org/10.1017/S0140525X01000115

Pavlidis, I., Tsiamyrtzis, P., Shastri, D., Wesley, A., Zhou, Y., Lindner, P., et al. (2012). Fast by nature - How stress patterns define human experience and performance in dexterous tasks. *Scientific Reports, 2*(1), 1–9. https://doi.org/10.1038/srep00305

Reinhard, M.-A., Greifeneder, R., & Scharmach, M. (2013). Unconscious processes improve lie detection. *Journal of Personality and Social Psychology, 105*(5), 721–739. https://doi.org/10.1037/a0034352

Reinhard, M.-A., Sporer, S. L., & Scharmach, M. (2013). Perceived familiarity with a judgmental situation improves lie detection ability. *Swiss Journal of Psychology, 72*(1), 43–52. https://doi.org/10.1024/1421-0185/a000098

Rimm-Kaufman, S. E., & Kagan, J. (1996). The psychological significance of changes in skin temperature. *Motivation and Emotion, 20*(1), 63–78. https://doi.org/10.1007/BF02251007

Scheffers, M. K., & Coles, M. G. H. (2000). Performance monitoring in a confusing world: Error-related brain activity, judgments of response accuracy, and types of errors. *Journal of Experimental Psychology. Human Perception and Performance, 26*(1), 141–151. https://doi.org/10.1037//0096-1523.26.1.141

Schweitzer, M. E., & Croson, R. (1999). Curtailing deception: The impact of direct questions on lies and omissions. *The International Journal of Conflict Management, 10*(3), 225–248. https://doi.org/10.1108/eb022825

W.W.A., Sleegers, Proulx, T., & van Beest, I. (2015). Extremism reduces conflict arousal and increases values affirmation in response to meaning violations. *Biological Psychology, 108*, 126–131.

Snellings, R. (2013). The effect of language proficiency on second-language lie detection (Thesis). Retrieved from https://ir.library.ontariotechu.ca/handle/10155/362.

Sokolov, E. N. (1963). Higher nervous functions: The orienting reflex. *Annual Review of Physiology, 25*(1), 545–580. https://doi.org/10.1146/annurev.ph.25.030163.002553

Tattersall, G. J. (2017). Thermimage. Thermal image analysis. Retrieved from https://CRAN.R-project.org/package=Thermimage.

Tyler, J. M., & Feldman, R. S. (2004). Truth, lies, and self-presentation: How gender and anticipated future interaction relate to deceptive behavior 1. *Journal of Applied Social Psychology, 34*(12), 2602–2615.

van Schie, H. T., Mars, R. B., Coles, M. G. H., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience, 7*(5), 549–554. https://doi.org/10.1038/nn1239

van't Veer, A. E., Gallucci, M., Stel, M., & van Beest, I. (2015). Unconscious deception detection measured by finger skin temperature and indirect veracity judgments - results of a registered report. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00672

Vergara, R. C., Moënne-Loccoz, C., Ávalos, C., Egaña, J., & Maldonado, P. E. (2019). Finger temperature: A psychophysiological assessment of the attentional state. *Frontiers in Human Neuroscience, 13*. https://doi.org/10.3389/fnhum.2019.00066. Retrieved from.

Vergara, R. C., Moënne-Loccoz, C., & Maldonado, P. E. (2017). Cold-blooded attention: Finger temperature predicts attentional performance. *Frontiers in Human Neuroscience, 11*. Retrieved from https://www.frontiersin.org/articles/10.3389/fnhum.2017.00454.