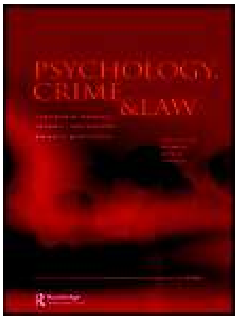

Psychology, Crime & Law

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gpcl20

A trifactorial model of detection of deception using thermography

G. Gálvez-García , J. Fernández-Gómez , C. Bascour-Sandoval , J. Albayay , J. J. González-Quiñones , A. Moliné , R. Pérez-Luco , A. Boso & E. Gómez-Milán

To cite this article: G. Gálvez-García , J. Fernández-Gómez , C. Bascour-Sandoval , J. Albayay , J. J. González-Quiñones , A. Moliné , R. Pérez-Luco , A. Boso & E. Gómez-Milán (2020): A trifactorial model of detection of deception using thermography, Psychology, Crime & Law

To link to this article: https://doi.org/10.1080/1068316X.2020.1815198

Published online: 04 Sep 2020.

PSYCHOLOGY, CRIME & LAW
https://doi.org/10.1080/1068316X.2020.1815198

Routledge
Taylor & Francis Group

A trifactorial model of detection of deception using thermography

G. Gálvez-García [a,b], J. Fernández-Gómez [c], C. Bascour-Sandoval [a,d], J. Albayay [e], J. J. González-Quiñones [f], A. Moliné [c], R. Pérez-Luco [a], A. Boso [g] and E. Gómez-Milán [c]

[a]Departamento de Psicología, Universidad de La Frontera, Temuco, Chile; [b]Département de Psychologie Cognitive, Sciences Cognitives & Neuropsychologie, Institut de Psychologie, Laboratoire d'Étude des Mécanismes Cognitifs, Université Lyon 2, Bron, France; [c]Centro de Investigación Mente, Cerebro y Comportamiento. Universidad de Granada, Granada, Spain; [d]Facultad de las Ciencias de la Salud, Universidad Autónoma de Chile, Temuco, Chile; [e]Dipartimento di Psicologia Generale, Università degli Studi di Padova, Padova, Italy; [f]Universidad de Granada, Granada, Spain; [g]Núcleo de Ciencias Sociales y Humanidades of Social Sciences and Humanities –Centro Butamallin para el Cambio Global, Universidad de La Frontera, Temuco, Chile

ABSTRACT

Most theories of detection of deception relate lying to three factors: (1) cognitive load, (2) subjective arousal, and (3) convincing impression. The evidence suggests that a higher temperature of the forehead is related to cognitive load, a lower temperature of the nose is associated with subjective arousal, and a higher temperature of the cheeks is related to convincing impression. Here, we took into account these three factors and, at the same time, associated the thermal change in specific facial regions of interest (ROIs) with each one of them. More importantly, we studied the combination of the thermal changes in the ROIs to establish the best combination to detect deception. Our results confirm an association between thermal changes in different ROIs and the three factors above. The best combination in the thermal changes of the ROIs for detecting deception (producing 83% accuracy and 13% false alarms in Experiment 1) is the one that was termed 'at least two of the three ROIs' where there is a lower temperature of the nose and/or a higher temperature of the cheeks and/or a higher forehead temperature. This finding constitutes an advance for detecting deception in multiple forensic contexts.

ARTICLE HISTORY
Received 12 October 2019
Accepted 30 July 2020

KEYWORDS
Thermography; facial temperature; subjective experience; detection of deception; forensic assessment

1. Introduction

Thermography has been used as a biological marker of subjective experiences (Clay-Warner & Robinson, 2015; Ioannou et al., 2014a). Facial thermography is a non-invasive tool for the study of thermal radiation, and depends on cutaneous blood perfusion controlled by the autonomous nervous system. The pattern of blood flow changes in the facial blood vessels because of the fight or flight response, producing temperature changes in the forehead, nose and cheeks (Derakhshan et al., 2019). These thermal

CONTACT G. Gálvez-García german.galvez@ufrontera.cl Departmento de Psicología, Universidad de La Frontera, Avenida Francisco Salazar 01145, 4780000, Temuco, Chile

variations are consistent with the activation of specific action units in the facial action coding system (Clay-Warner & Robinson, 2015). Emotion produces changes in the blood flow associated with the activated muscles (Jarlier et al., 2011). For example, in the case of deception, Pavlidis et al. (2002) found that the change in temperature was caused by an increase in the blood flow to the eyes. The periorbital and supraorbital vessels of the face that feed the main muscles surrounding the eyes show heat changes as a result of stressors, and this is also true for the cheeks, the forehead and the tip of the nose (Ioannou et al., 2014a; Kosonogov et al., 2017). Skin blood circulation is regulated by the sympathetic control of the adrenergic vasoconstriction and cholinergic vasodilator systems that, respectively, can produce decrements or increments in the tip of the nose temperature (Cho et al., 2019). However, in the case of the nose, the thermal change is independent of muscle contraction, and depends on the abundant arteriovenous anastomoses in this area (Johnson et al., 2011; Walløe, 2016). The thermal changes in the forehead are associated with the core body temperature because of the vascularization and thin skin in this area (Ioannou et al., 2014a). All in all, the studies cited have reported that the skin temperatures of the nose, forehead and cheeks provide an effective physiological index to evaluate psychophysiological states (Abdelrahman et al., 2017; Engert et al., 2014; Ito et al., 2018; Shastri et al., 2009).

The high sensitivity of the tip of the nose to different emotional situations can be observed in the temperature changes that occur in this area at the same time as emotions, making this first area a suitable reference for thermographic measurement. The predominant temperature stability of the forehead allows this second area to act as an appropriate baseline. Emotion-related studies using thermography have been carried out previously (see Salazar-López et al., 2015 for a review). Lies are considered to be experiences that create high arousal and negative valence (Colwell et al., 2007), have high cognitive load (Vrij et al., 2008) and are associated with language pragmatics (Meibauer, 2018). The use of modern thermographic cameras for the detection of lies is a recent development (Pavlidis et al., 2002). In recent years, thermography and polygraphs have focused on lying as a negative experience that produces stress for the liar (Engert et al., 2014; Panasiti et al., 2016; Pavlidis et al., 2000). However, the primary advantages of thermography over the polygraph are that thermography is a contact-free technology that records the response of the autonomic nervous system (ANS) and can measure not only ANS activity but also cognitive load (e.g. Abdelrahman et al., 2017) and social cognition (e.g. Paolini et al., 2016). Moreover, a crucial advantage of the thermal camera is an average accuracy of detection of deception of 70%, although the accuracy may range between 70% and 90%, depending on the experimental lying scenario (Gołaszewski et al., 2015; Pollina et al., 2006).

Following Vrij and Verschuere (2013), most theories about the detection of deception relate lying to one of three factors: (1) cognitive load, (2) subjective arousal, and (3) the attempt to make a convincing impression. However, to the best of our knowledge, no study has tested this trifactorial model with thermography; in other words, specific facial thermographic regions of interest (ROIs) have not been associated with each of the three factors at the same time (lower temperature of the tip of the nose with arousal) (Kosonogov et al., 2017; Moliné et al., 2017; Salazar-López et al., 2015); higher temperature of the forehead with cognitive load (Abdelrahman et al., 2017; Moliné et al., 2018); and higher temperature of the cheeks with social cognition (Ioannou et al., 2013, 2014b, 2016; Paolini et al., 2016). The absence of previous research studies on this topic could be the result of blurred boundaries between these three ROIs and the three factors associated with deception. For example, the thermal changes in the tip of the nose can be associated not only with high arousal but also with high cognitive load (Abdelrahman et al., 2017) and with social emotions such as guilt (Ioannou et al., 2013).

Concerning the detection of deception in particular, but also concerning stress and emotions in general, inconsistent results have been reported in the literature for the thermal changes in the nose and forehead. Some studies about the detection of deception, mental effort, and emotions with high arousal (and/or negative valence) have revealed a decreased temperature in some facial regions, specifically in the nose (Ioannou et al., 2013; Kosonogov et al., 2017; Moliné et al., 2017; Or & Duffy, 2007; Shastri et al., 2009). This lower nose temperature has been related to stress and increased sympathetic activity. However, in other studies, high arousal has been associated with increased nose temperature in the case of negative emotions, detection of deception, mental effort, crying, ostracism, or direct gaze (Ioannou et al., 2014b, 2016; Panasiti et al., 2016; Paolini et al., 2016). This thermal effect (i.e. higher nose temperature) has frequently been related to a complex autonomic interaction between the sympathetic and the parasympathetic nervous systems, or to residual effects that are probably due to a withdrawal of the sympathetic alpha-adrenergic vasoconstriction effect (Ebisch et al., 2012; Ioannou et al., 2016; Moliné et al., 2017). To clarify these incongruent results, Moliné et al. (2018) carried out a set of experiments, and they also studied the thermographic detection of deception with two tasks: one ecological interview with open questions and a second interview with closed questions. The type of question is a relevant factor in improving the detection of deception (Oxburgh et al., 2010). Moliné et al. (2018) obtained 85% accuracy and 25% false alarms, taking into account only two facial ROIs: higher temperature of the forehead and/or lower temperature of the tip of the nose with respect to a baseline.

In the study by Moliné et al. (2018), the participants performed two tasks to tell a lie: (1) one task with a high mental workload and high anxiety, and (2) a second task with a low mental workload which produced low or no anxiety. In the first task, ten participants were asked to come to the lab after being told that they were to participate in a top-secret research study about which they should tell no one. They were then told they had three minutes to create an alibi, and should fabricate a lie to tell to a friend or partner, whom they would call on their phone after the three minutes had passed. A control group of ten participants had to tell the truth ('I am in a top-secret experiment'). A participant was classified as deceptive if their nose had a lower temperature and/or their left and/or their right forehead a higher temperature. However, it is likely that false alarms were underestimated and the accuracy overestimated, because the authors did not employ an appropriate control condition (i.e. with the participants telling the truth), probably leading to a lower level of arousal in the control condition when compared to the deceptive condition (the authors did not offer data on arousal for the control condition). A better control condition should produce a similar level of arousal.

The present study aimed to use thermography to test the trifactorial model of the detection of deception (Vrij & Verschuere, 2013) based on cognitive load, subjective arousal, and convincing impression. Thus, we aimed to replicate the results of Moliné et al. (2018; i.e. the alibi task). However, we also took into account a third ROI, namely the cheeks (i.e. the factor related to social cognition or convincing impression).

As a secondary aim, we proposed some methodological approaches that were different from previous research (Moliné et al., 2017, 2018), by estimating accuracy and false alarms with higher rigor. We anticipated that this methodology would reveal that Moliné et al. (2018) underestimated false alarms and overestimated the accuracy of the detection of deception in their model (i.e. lower nose temperature and/or higher forehead temperature).

These new methodological approaches (described below) were part of the second aim of this research, so they are presented below. First, and to achieve the previous aim, the cheeks and their relationship with convincing impression were measured. Second, and to avoid the incorrect estimations of accuracy and false alarms, a better control was used, to produce a similar level of arousal for participants who were telling the truth and participants who were lying, as was done by Moliné et al. (2017). Thus the participants were asked to describe images with negative valence and high arousal selected from the IAPS battery (Lang et al., 2008; Moliné et al., 2017). Furthermore, and to give better control of the incorrect estimation of accuracy and false alarms, stricter criteria were employed for the ROIs. Moliné et al. (2018) classified a thermal increment in forehead temperature as a positive point for the detection of deception (+1) using a relaxed criteria, that is, when the temperature of the left or the right side of the forehead increased in conjunction; however, frequent asymmetries were found (sometimes only the left or only the right part of forehead showed thermal change), which resulted in the accuracy being overestimated. In our study we used strict criteria: only when both parts of the forehead showed a higher temperature than the baseline (a difference of at least 0.3°C) was the forehead thermal change classified as a positive point (+1) for the detection of deception. The same criterion was used for the cheeks—both cheeks had to show a higher temperature than the baseline. In the case of the tip of the nose, only one measurement was taken. Third, to improve the participants' uncertainty and the ecological validity of the study, the participants did not know their experimental condition (i.e. whether they would be asked to tell the truth or to lie) until the last minute (after they had arrive at the laboratory); this improved the conditions to be more like real life, and avoided the participants rehearsing their lies or following an automated script. Finally, a chin rest and hand-holders were used to avoid undesirable movements of the body and head that could affect the thermal recording. To avoid the temperature of the hands and the phone interfering with the face temperature, a problem that may have occurred in previous research (Moliné et al., 2017, 2018), we used a hands free phone call.

2. Experiment 1

2.1. Method

2.1.1. Participants

Thirty university students (15 women, age range = 18–32 years old) were recruited for this experiment. All the participants were instructed to read a brief description of this research project; we obtained written informed consent from each participant. After that, each participant answered a series of medical and biographical questions to ensure that they were in good health and not taking medication or drugs that could interfere with the examination results. We employed Moreira et al.'s (2017) checklist for measuring skin temperature, with overall points 1–3, as the exclusion criterion. No participant was excluded. This study was approved by the Research Ethics Committee of the University of Granada, and was carried out in compliance with the principles of the Declaration of Helsinki (World Medical Association, 2013).

2.1.2. Equipment

A thermographic camera, ThermoVision A320G Researcher Infrared Camera, which had a sensitivity difference of 0.07° to 30°C, a sensitivity error of 0.2°C and adequate high resolution for human research and medical thermography, was used (Ring, 2007). The ThermoVision A320G offers different ranges of colors. We used the medical range—named for its use in medical thermography—which gives a clearer view of temperature changes. The work was carried out at a level between 18°C and 40°C (this level is within the sensitivity range of the camera). The camera was placed on a tripod at a distance of 110 cm from the floor and between 60 and 100 cm from the participants. The height was manipulated to capture either the upper body or only the face of each participant. As explained in the introduction, we used a chin rest and hand-holders to avoid undesirable body and face movements.

The automatic focus on the camera was used each time. Researcher ThermaCAM 2.9 by the manufacturer FLIR was used on a PC laptop to capture the signal. ThermaCAM 2.9 is software that allows for the continuous recording of eight photo frames per second.

2.1.3. Procedure and settings

The measures were performed at the same time of day for all participants (Marins et al., 2015). The experiment took place in a tested thermographic laboratory (Fernández-Cuevas et al., 2015; Moreira et al., 2017). The protocol for taking measurements using thermographic cameras (Fernández-Cuevas et al., 2015; Moreira et al., 2017; Ring & Ammer, 2000) that we followed for this study demands specific preparations in order to obtain adequate recordings.

The experiment was carried out in a closed room of about 40 m^2, with a changing room next to it. The thermographic camera, the computer, and the experimenter were positioned in the middle of the room, facing the subject. The area of the skin to be recorded should not be covered by any material. In this way, thermography is able to capture images that accurately reflect the temperature. The subject must be at rest for between 10 and 15 min in order to adapt to room temperature before the temperature of the skin is recorded, and the room temperature must be between 18°C and 24°C ($M = 22$°C in our case). The humidity must also to be controlled ($M = 50\%$ in our case). The procedure required that when the participants entered the room, they remained seated for 10 min on a stool in the changing room adjacent to the studio. After this, they entered the studio and received instructions as follows: 'We are now going to record your face and hands while performing different exercises or tasks. First, a static image at rest will be recorded and then we will indicate the nature of the tasks to be performed.'

To ensure high quality measurements in the ROIs, the whole sequence of movements comprising each experiment was recorded using thermography. The relevant ROIs for this project were the entire forehead, capturing both the left and the right sides, both cheeks, and the tip of the nose; these have proved to be the best delimited ROIs in previous studies, in line with the protocol for delimiting ROIs by Salazar-López et al. (2015) and Ammer and Formenti (2016); see Figure 1. To guarantee a consistent recording of the ROIs, we applied the same square size (3 × 3 pixels) for each ROI in all thermograms for each participant. All participants were their own controls for their ROIs; therefore, all participants were recorded in the same position during all tasks, seated on a stool of variable height, with chin rest and hand-holders.

The participants were asked to come to the lab and, once they had arrived, were told that they were participating in a top-secret research study, about which they should tell no one, including their family, close friends, or partners. In this condition (henceforth, 'deceptive'), once the participants were situated in the thermography room and after the adaptation period, they were told that they had three minutes to create an alibi and that they should fabricate a significant and emotional lie to tell to a friend or partner of theirs, whom they were to call by phone after three minutes. Their goal was to convince their friend or partner of the lie they were telling about their whereabouts and about what they were doing at that moment; they were instructed not to tell the actual truth (i.e. 'I am in a top-secret experiment'). The participants were asked to be as convincing as possible. They would receive five Euros as a reward if the researcher (presented as an expert in non-verbal and verbal behavior) declared that they were compelling and that the person whom they had called believed them. Under the control condition (henceforth,

Figure 1. Main ROIs (tip of the nose, forehead, and cheeks) and secondary-control ROIs (middle fingers, chin, and maxillary mouth).

'truthful'), the same participants spent the waiting time in the experimental room in the same way as in the deceptive condition. After this, each participant called a person important to them and described in detail images that were appearing automatically on a screen in front of them, for a duration of 30 s each, for 6 min. The images were selected from the IAPS battery (Lang et al., 2008). Specifically, we selected the negative valence and high arousal images used by Salazar-López et al. (2015) in their Experiment 1 (set 4). For instance, we selected images of cadavers, mutilated bodies, and violent acts (Numbers 1525, 9265, 3015, 9433), whose average (SD) was 1.70 (1.10) for the valence dimension and 7.35 (1.80) for the arousal dimension. A state of anxiety of 30.25, as measured by the STAI state scale (Spielberger et al., 1999), was produced in people observing these images in a pilot study, which was in line with similar states of anxiety in previous studies (Limonero et al., 2015; Pacheco-Unguetti et al., 2010). The objective was to compare the thermograms of the participant when they were telling the truth and when they were lying, under a similar level of anxiety in both cases; not finding differences in anxiety between the groups would be fundamental to establishing that any possible differences in thermal changes for the deceptive group could not be attributed to an anxiety effect. The alibi task ('deceptive condition') and the control task ('truthful condition') were performed separately on two consecutive days by each participant. Examples of some of the alibis from the participants (the top-secret experiment) were: (1) 'Mum, I am pregnant'; (2) 'Sweetie, I just found a wallet with 1200 Euro, should I call the police or should I keep it?'; (3) 'I just saw your boyfriend walking hand in hand with someone else'; and (4) 'I have been arrested by the police with 10 grams of marijuana'. The deceptive and truthful conditions were counterbalanced across the participants (i.e. half of the participants began with the truthful condition and the other half began with the deceptive condition).

Finally, three subjective questions were employed in order to test the relationship between the three factors of detection of deception and thermal changes, in line with previous studies (Moliné et al., 2017). Thus, 7-point Likert scales ranging from 1 (not at all) to 7 (very much) were used. Cognitive load was measured through the question, 'How difficult did you find the questions or the interview?', subjective arousal was measured through the question, 'How nervous did you feel during the interview (phone call)?', and, finally, convincing impression was measured through the question, 'How convincing do you think you were when telling your alibi?'

2.1.4. Thermographic analysis

Each participant was first recorded for the baseline shot, which shows an initial thermogram of the whole face and hands, and was then recorded during the task and after performing the task. By way of control, the whole sequence of each experiment was recorded as is typically done in medical thermography, and we only considered changes in temperature that were equal to or greater than ± 0.3°C in line with previous research (Moliné et al., 2018). Each ROI was computed as a square of 3 × 3 pixels around the point selected, using a MATLAB algorithm that calculates mean, maximum, minimum, and standard deviation for each ROI at the beginning and at the end of each video thermogram; MATLALB labeled each ROI with ' +1' when the temperature increased by 0.3°C or more, '−1' when the temperature decreased by 0.3°C or more, and '0' when there were no thermal changes or only one side of the face (for cheeks and forehead) showed a thermal change. In addition, the MATLAB algorithm calculated the sequential measurement of one ROI temperature changes. Each time series was corrected for outliers (any point more than 3 standard deviations from the mean of the time series). The resulting signal values were converted to the *z* scores. Also, an expert coding researcher performed these analyses to test the reliability of the MATLAB algorithm. The correlation between human coder and MATLAB algorithm was 0.99, $p < 0.001$ with a difference of 0.02 ± 0.35°C. A second expert coding checked these differences. They were exclusively attributed to small human errors in the codification.

2.1.5. Statistical analyses

We used RStudio (version 1.1.383; RStudio Team, 2016) for all our analyses. The main dependent variable of our study was the change in the thermograms (henceforth 'thermal change') between the baseline and the lying condition (truthful or deceptive). Using a repeated measures analysis of variance (ANOVA), we set the thermal change as the dependent variable, and 'lying condition' (truthful vs. deceptive) and 'ROI' (left forehead, right forehead, left cheek, right cheek, and nose) were included as fixed factors. Multiple comparisons were carried out with a Bonferroni correction and with the α level set at 0.05. Separately, the anxiety levels (STAI) measured before and after the completion of the test were compared in a second ANOVA, where 'time' (pre vs. post) and 'lying condition' were included as fixed factors. Furthermore, a multivariate analysis of variance (MANOVA) was conducted in order to determine whether the lying condition had a multivariate effect on the measures of 'cognitive load', 'subjective arousal', and 'convincing impression'. To calculate the relative evidence for a given alternative hypothesis (H_1) compared to the null hypothesis (H_0), we calculated the Bayesian Factor (*BF*) for the corresponding analyses; $BF > 1$ indicates that the data support H_1, while $BF < 1$ indicates that they support H_0 (Jeffreys, 1961; Raftery, 1995). The *BF* was computed using the *BayesFactor* package (Morey et al., 2018). Partial eta-squared (η^2_p) was reported as an estimate of effect size. Additionally, based on the correlation matrix of our data, we calculated the correlation between the measured tests (i.e. cognitive load, subjective arousal, convincing impression, and anxiety (pre and post)) and the ROIs in both the truthful and the deceptive conditions; the *p*-values were computed using the *Hmisc* package (Harrell Jr & Harrell Jr, 2015). We estimated the mean and standard deviation of each variable for the respective levels of the factors.

The accuracy and false alarms were calculated as follows. A thermal increment for both sides of the forehead or cheeks of at least 0.3°C, or a decrement for the nose of at least −0.3°C was classified as a positive point of detection of deception (+). This was done only when both ROIs of the forehead (or the cheeks) showed a higher temperature, or a lower temperature was found for the nose, as compared to the baseline. A positive point of detection of deception for the truthful condition was considered as a false alarm. For the deceptive condition, a positive point of detection of deception was considered a hit of accuracy.

First, the accuracy and false alarms were computed by considering significant thermal changes for bilateral forehead, bilateral cheeks and nose separately. For the forehead criterion only significant temperature increases for the forehead were counted as evidence of deception. For the cheek criterion, only significant temperature increases for the cheeks were counted as evidence of deception. For the tip of the nose criterion, only significant temperature decreases for the nose were counted as evidence of deception. We then counted the number of participants in the deceptive condition who fulfilled the criteria for each separate ROI model (accuracy), and the number of truthful participants who fulfilled the criteria for each separate ROI model (false alarms). Next, we computed all possible combinations of significant thermal changes (using additive models and dilemma models) to calculate other models, to obtain the accuracy and the proportion of false alarms. In the additive models, we looked for more than one of the possible three thermographic indexes of detection of deception at the same time (these models were based on the coordinative grammar conjunction 'and'). In the dilemma models we looked for at least one of the three thermographic indexes of detection of deception (these models were based on the coordinative grammar conjunction 'or'). In the mixed models, we combined the additive and the dilemma models of detection of deception, based on the conjunction 'and/or'. Examples of additive models are, for two ROIs, forehead and cheeks, forehead and nose, and cheeks and nose. Examples of dilemma models are forehead or nose, cheeks or nose, and forehead or cheeks, and models in which there are all possible combinations of significant thermal changes for the three ROIs (higher temperature of forehead and higher temperature of cheeks and lower temperature of nose, or higher temperature of forehead or higher temperature of cheeks or lower temperature of nose). An example of a mixed model is that termed 'at least two of the three ROIs' (with significant thermal changes at the same time), as explained below.

This model takes looked for higher forehead temperature and/or higher cheek temperature and/or lower tip of the nose temperature, but with at least two indexes at the same time in the thermogram. This was calculated as follows. First, the instances of the combination where there was an increase in the temperature of the forehead and cheeks and a decrease in the temperature of the nose were counted. Second, the instances where there was a combination of an increase in the temperature of the cheeks and a decrease in the temperature of the nose (with a non-significant thermal change in forehead) that had not been counted previously were counted. Third, the instances with a combination of an increase in the temperature of the forehead and a decrease in the temperature of the nose (with a non-significant thermal change in the cheeks) that had not previously been counted were added. Fourth, the instances with a combination between an increase in temperature for both the forehead and the cheeks (with a non-significant thermal change in the tip of the nose) that were not counted in the previous combinations were added. Finally, the sum of all the above combinations was used to compute the accuracy for the deceptive condition and the number of false alarms for the truthful condition.

2.2. Results

2.2.1. Thermal changes

The thermal changes for each ROI were significantly modulated by the lying condition. The main effect of the lying condition was significant, $F(1, 290) = 35.50$, $p = 0.003$, $BF > 100$, $\eta^2_p = 0.109$, with higher values for the deceptive condition (0.26 ± 1.20) than for the truthful condition (−0.26 ± 0.65). The main effect of the ROI was also significant, $F(1, 290) = 30.89$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.299$. Multiple comparisons revealed that the values were significantly lower for the nose (−0.98 ± 1.27) as compared to the left forehead (0.24 ± 0.74), the right forehead (0.22 ± 0.66), the left cheek (0.31 ± 0.76), and the right cheek (0.24 ± 0.82), all *ps* < 0.001; other comparisons were not significant (*ps* = 1.000). The interaction lying condition × ROI reached significance as well (see Figure 2), $F(1, 290) = 16.23$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.183$. Multiple comparisons revealed higher values within the deceptive (vs. truthful) condition for the left forehead (0.73 ± 0.67 vs. −0.25 ± 0.40, $p < 0.001$), the right forehead (0.71 ± 0.51 vs. −0.28 ± 0.35, $p < 0.001$), the left cheek (0.71 ± 0.78 vs. −0.08 ± 0.49, $p = 0.003$), and the right cheek (0.57 ± 0.86 vs. −0.17 ± 0.35, $p = 0.009$), but significantly lower values for the nose (−1.42 ± 1.30 vs. −0.54 ± 1.09, $p < 0.001$).

2.2.2. Anxiety

The anxiety level of the participants before and after performing the tasks differed significantly, but it was not modulated by the lying condition. The main effect of time was significant, $F(1, 116) = 521.195$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.818$, with lower STAI scores for the baseline (20.0 ± 2.1) than just after completion of the task (29.3 ± 2.4). On the other hand, the effect of lying condition was not significant, $F(1, 116) = 0.059$, $p = 0.808$, $BF = 0.19$, $\eta^2_p = 0.001$ (deceptive = 24.6 ± 5.2. truthful = 24.7 ± 5.2), as indicated by both the *p* value and the *BF* supporting H_0. Furthermore, the interaction time × lying condition was not significant, $F(1, 116) = 0.106$, $p = 0.746$, $BF > 100$, $\eta^2_p = 0.001$. However, the *BF* supported H_1, so we explored the differences between the levels of the factors. Multiple comparisons showed differences between the 'pre' and 'post' conditions for both the truthful (20.0 ± 2.0 vs. 29.5 ± 2.1, $p < 0.001$) and the deceptive (20.0 ± 2.1 vs. 29.2 ± 2.7, $p < 0.001$) conditions.

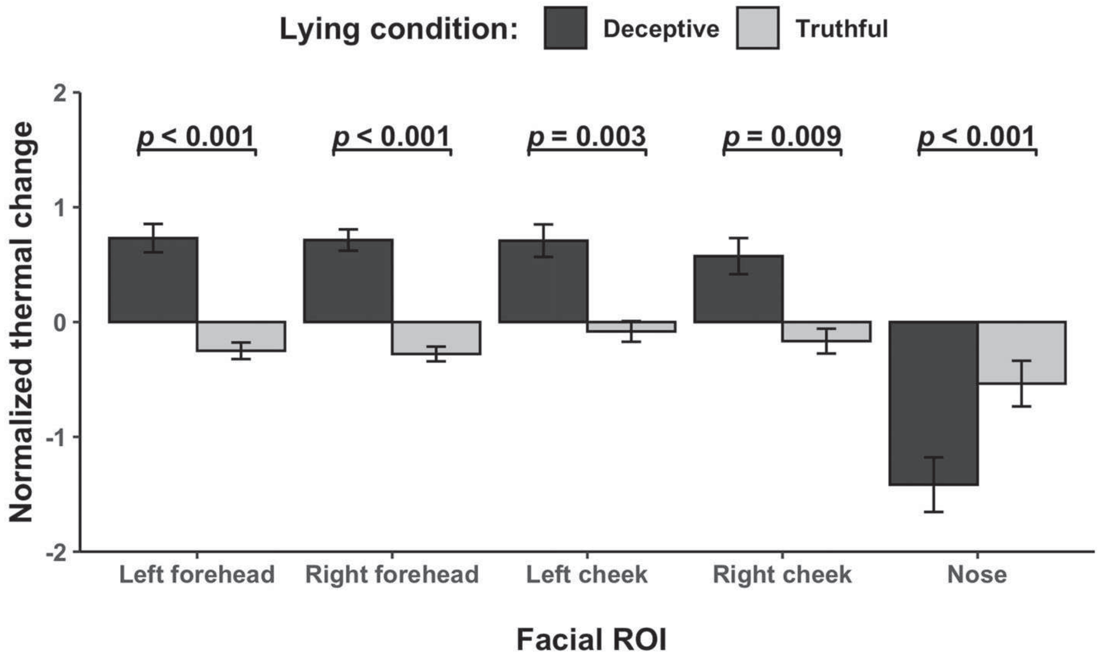

Figure 2. Normalized thermal change for each facial ROI as a function of lying condition (deceptive vs. truthful) in Experiment 1. Error bars represent standard errors of the mean.

2.2.3. Cognitive load, subjective arousal and convincing impression

The MANOVA revealed a significant multivariate effect of lying condition on cognitive load, subjective arousal and convincing impression, $F(1, 298) = 74.183$, $p < 0.001$, $\eta^2_p = 0.429$, Wilks' Lambda = 0.571. Higher cognitive load was revealed in the deceptive condition (4.6 ± 1.7) as compared to the truthful condition (2.5 ± 1.2), $F(1, 298) = 155.93$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.343$. Subjective arousal was also higher for the deceptive condition (5.0 ± 1.2) than for the truthful condition (4.4 ± 0.9), $F(1, 298) = 23.946$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.074$. On the other hand, convincing impression scores were higher in the truthful condition (5.2 ± 1.2) than in the deceptive condition (4.3 ± 1.0), $F(1, 298) = 50.136$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.144$.

2.2.4. Correlations

The correlations between the ROIs and the measured tests (i.e. cognitive load, subjective arousal, convincing impression, and anxiety (pre and post)) for both the truthful and the deceptive conditions are presented in Figure 3. In the truthful condition, we found significant moderate (in the left) and low (in the right) positive associations between cognitive load and thermal changes in the sides of the forehead. Moreover, we found a significant low negative association between subjective arousal and thermal change in the nose. For the deceptive condition, we found significant moderate positive associations between cognitive load and thermal changes in the left and right foreheads. A significant low negative association between subjective arousal and thermal change in the nose was also revealed in this condition. Furthermore, we observed significant low (for the left) and moderate (for the right) negative associations between convincing impression and thermal changes in the cheeks. Importantly, the STAI anxiety tests (pre and post) did not correlate with the thermal change of any of the ROIs, either in the truthful or in the deceptive condition.

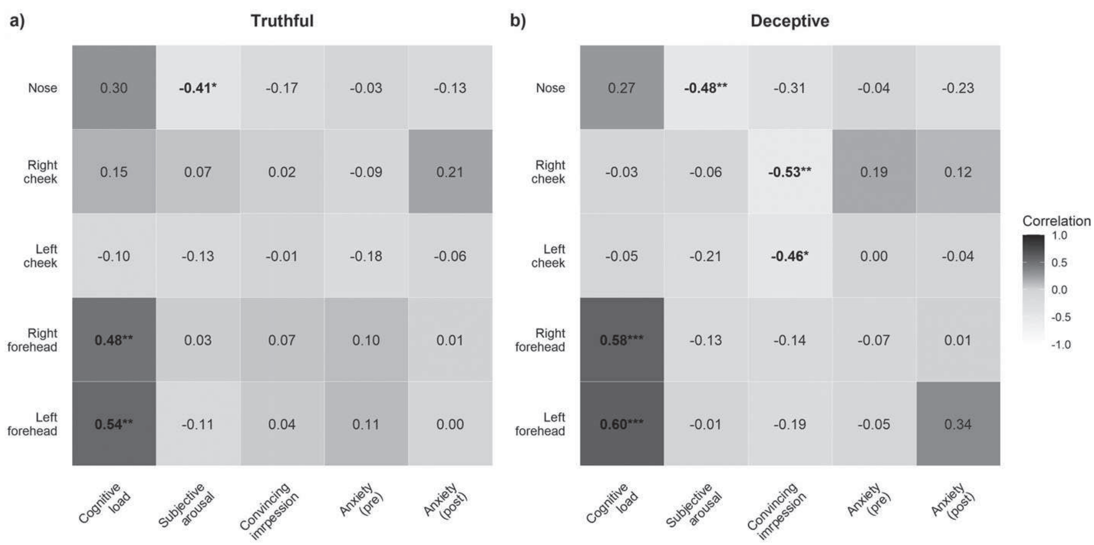

Figure 3. Correlations between the thermal change for each facial ROI and the measured tests within both the truthful condition (panel a) and the deceptive condition (panel b) in Experiment 1. Bold values denote statistical significance ($p < 0.05$*, $p < 0.01$**, $p < 0.001$***).

2.2.5. Accuracy and false alarms

First, accuracy and false alarms were computed from the results for forehead, cheeks and nose separately. Taking into account each ROI described here (see Table 1), the accuracy in detecting a deceptive participant for each ROI calculated in isolation was 77% for the forehead criterion, 80% for the cheeks criterion and 63% for the nose criterion. The proportions of false alarms were 27% for the forehead criterion, 33% for the cheeks criterion and 47% for the nose criterion. Accuracy was higher with the dilemma combination between ROIs, although the number of false alarms was also higher. The criterion with the additive combination between three ROIs gave lower accuracy and fewer false alarms as compared with the other types of combination (e.g. 37% accuracy and 0% for false alarms for the forehead plus cheeks plus nose criteria). The best way to predict who is a liar (high accuracy and a low score for false alarms) is found to consider a criterion with at least two of the three ROIs in a mixed model, which gives 83% accuracy and 13% for false alarms.

2.3. Discussion

This study aimed to test the trifactorial model for the detection of deception (Vrij & Verschuere, 2013), which is based on cognitive load, subjective arousal, and convincing impression, using thermography. Our results support a direct relationship between thermal changes and the factors mentioned above. An increase in forehead temperature is related to cognitive load, a decrease in nose temperature is related to subjective arousal, and, finally, a change in cheek temperature is related to convincing impression. These results confirm the previous studies in which ROIs and factors of detection of deception were studied in isolation (Abdelrahman et al., 2017; Ioannou et al., 2013, 2016; Moliné et al., 2018; Paolini et al., 2016). Our secondary aim was to introduce new methodological approaches that will improve accuracy and reduce the number of false alarms when compared to the previous research of Moliné et al. (2018). These authors found a decrease in the temperature of the nose or an increase in the temperature of the forehead when the participants were lying. Our results confirm these results, adding an increase in the temperature of the cheeks as an index for the detection of deception. Thus, we found that temperature changes of the nose, cheeks, and forehead allow us to detect, with high accuracy and a low number of false alarms, when people are lying about facts, under certain

Table 1. Accuracy and false alarms in Experiment 1 for all possible combinations of the three facial ROIs selected—lower temperature of the tip of the nose, higher temperature of the forehead, and higher temperature of the cheeks.

ROIs	Accuracy	False Alarms
Forehead	77%	27%
Cheeks	80%	33%
Nose	63%	47%
Forehead or Cheeks	94%	57%
Forehead or Nose	90%	60%
Cheeks or Nose	100%	70%
Forehead or Cheeks or Nose	100%	90%
Forehead plus Cheeks	60%	3%
Forehead plus Nose	50%	3%
Cheeks plus Nose	47%	7%
Forehead plus Cheeks plus Nose	37%	0%
Forehead and/or Cheeks and/or Nose: But at least two of the ROIs	83%	13%

conditions which will be described below. Importantly, although a direct statistical comparison between the two studies (Moliné et al., 2018 and the current study) was not performed, the new methodological conditions presented here seem stricter than previous research. For example, there were no significant differences in anxiety between the control condition and the deceptive condition. More importantly, although there was an anxiety effect (i.e. higher levels after the completing the experiment), the STAI anxiety tests (pre and post) did not correlate with the thermal change of any of the ROIs, in either the truthful or the deceptive condition. Furthermore, our results revealed that the anxiety level did not differ significantly between the deceptive and the truthful conditions, which was also supported by a Bayesian Factor calculation in favor of the null hypothesis.

Deceptive participants could be detected with a high average accuracy of 73% and false alarms of 36% when thermal changes in ROIs were computed in isolation. This average accuracy was higher with the dilemma combination between ROIs (96%), although this combination also increased the proportion of false alarms (69%). This trend was especially acute when the three ROIs were combined. The average accuracy for the additive combinations presented the lowest levels of both accuracy and false alarms of all the combinations (around 50% and 3%, respectively). Finally, for the combination taking into account at least two of the three ROIs (i.e. at least two of the three possible thermal changes expected in the deceptive condition), we obtained 83% accuracy and 13% false alarms. In other words, when we combined the forehead thermal change (higher temperature with respect to the baseline) and/or the nose thermal change (lower temperature) and/or the cheek thermal change (higher temperature) to predict who was a liar (two or three changes for the deceptive condition against only one change or no change for the truthful condition) we get the best results. Thus, it seems clear that a good strategy, with high accuracy and a low number of false alarms, is to detect at least two significant thermal changes.

A random decision would give 50% accuracy and 50% false alarms. Any model that produces more than 50% for accuracy and less than 50% for false alarms is a good starting point for detecting deception and truth. The greater the accuracy and the lower the number of false alarms, the better the model. If we consider each thermal change of any ROI in isolation and use this as an index of deception or truth, it appears that we need at least two indexes at the same time to generate results that are better than random decisions for accuracy and for false alarms. Only one index is a necessary but not a sufficient condition to detect deception, particularly if we take into account the fact that the difference between telling the truth and lying is quantitative rather than qualitative (around 70% accuracy but around 30% false alarms). The presence of the three indexes of deception at the same time in the thermogram is a very strict criterion (it detected only 40% of the liars, but this combination never appeared for the truthful participants). At least two of the three indexes as the criterion appears to be a good solution for a starting point (giving higher accuracy and fewer false alarms, and much better than random results). In other words, the mixed model (i.e. combining additive and dilemma models of detection of deception) gives a good reflection of how difficult it is to differentiate a truth from lie. We need to accumulate more than one indicator to obtain a probabilistic decision with guarantees that also takes into account individual differences (because not all liars show all the components of a lie).

3. Experiment 2

To give consistency and to generalize our findings about our trifactorial model based on thermography, a new experiment was proposed. As background, it should be noted that previous research (e.g. Kleinberg et al., 2019) has shown that trying to find the optimal statistical separation between deceptive and truthful statements within a single dataset leads to inflated accuracy estimates. The desirable way to evaluate the predictive accuracy of a classification algorithm is to build the algorithm on one dataset and test it on a new dataset. It is likely that the reported accuracy rate of Experiment 1 could be an overestimate of the true accuracy. In Experiment 2, a new truthful condition was introduced in which participants were interrogated about a fictitious burglary. In this way, we aimed to generalize our results to other situations, particularly those related to criminal acts.

3.1. Method

3.1.1. Participants

Thirty university students (15 women, age range = 18–25 years old) were recruited for Experiment 2. The inclusion criteria were the same as in Experiment 1 (see Section 2.1.1); no participant was excluded. This study was approved by the Research Ethics Committee of the University of Granada, and was carried out in compliance with the principles of the Declaration of Helsinki (World Medical Association, 2013).

3.1.2. Equipment

We employed the same apparatus described for Experiment 1 (see Section 2.1.2).

3.1.3. Procedure and settings

The procedure for this experiment was similar to the one employed for Experiment 1 (see Section 2.1.3) as regards the presentation of the stimuli and the thermographic recordings (truthful and lying conditions). However, here we only used the MATLAB algorithm due to its high reliability. Before the experimental condition, the participants were asked to come to the laboratory to record their personal information. Seven days later, in the deceptive condition, we ran the alibi task again (exactly as in Experiment 1). In the truthful condition, the same participants answered some questions from the director of the laboratory about the (invented) burglary of a handbag belonging to one of the participants (a confederate of the researcher). The participants were brought one by one into the thermographic room, and the researcher asked for their permission to record their answers to questions about the burglary using the thermal camera in order to validate whether their testimony was deceptive or truthful. The open questions were: (1) 'Did you know that last week an object was stolen from this laboratory more or less during the time you were here? Please explain what you did before and after your session'. (2) 'Did you know, hear or see anything strange?' (3) 'Please try to guess what object was stolen: a) A handbag, b) Money, c) A cellphone' (before the interview all the participants had been accidentally informed by a friend, a confederate of the researcher, that a handbag had been stolen; consequently all of them chose the handbag as the stolen item and explained why they knew this). (4) 'Did you steal the handbag?' (5) 'Do you know who did it?' (6) 'Do you think that you have passed the thermographic test, explain why?' Importantly, a state of anxiety of 32.82, as measured by the STAI state scale (Spielberger et al., 1999), was produced after observing these questions, in line with similar states of anxiety in previous studies (Limonero et al., 2015). The order of the truthful and the deceptive conditions were counterbalanced across the participants.

3.2. Results

3.2.1. Thermal change

A significant main effect of lying condition on thermal changes was found, $F(1, 290) = 28.10$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.088$, with higher values for the deceptive (0.24 ± 1.25) than for the truthful condition (−0.24 ± 0.56). The main effect of ROI was significant, too, $F(1, 290) = 23.63$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.246$. Multiple comparisons showed that the values were significantly lower for the nose (−0.89 ± 1.16) than for the left forehead (0.16 ± 0.64), right forehead (0.15 ± 0.73), left cheek (0.24 ± 0.93), and right cheek (0.33 ± 0.94), all *ps* < 0.001; other comparisons did not reach significance (*ps* = 1.000). Furthermore, the interaction lying condition × ROI was also significant (see Figure 4), $F(1, 290) = 14.47$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.166$. Multiple comparisons showed higher values in the deceptive (vs. truthful) condition for the left forehead (0.57 ± 0.60 vs. −0.24 ± 0.38, $p = 0.005$), right forehead (0.54 ± 0.79 vs. −0.23 ± 0.38, $p = 0.010$), left cheek (0.68 ± 1.02 vs. −0.19 ± 0.56, $p = 0.001$), and right cheek (0.77 ± 1.07 vs. −0.12 ± 0.51, $p < 0.001$), but significantly lower values for the nose (−1.34 ± 1.27 vs. −0.43 ± 0.84, $p < 0.001$).

3.2.2. Anxiety

The STAI scores (anxiety level) measured before and after the completion of the task differed significantly, but were not affected by lying condition. The main effect of time

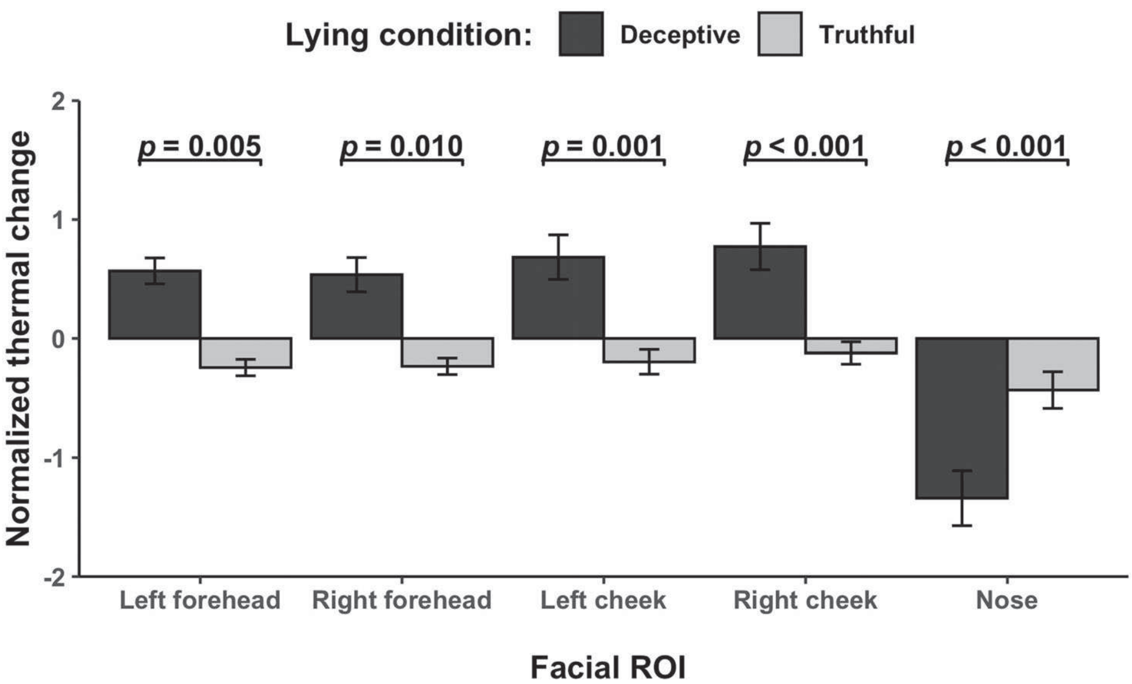

Figure 4. Normalized thermal change for each facial ROI as a function of lying condition (deceptive vs. truthful) in Experiment 2. Error bars represent standard errors of the mean.

was significant, $F(1, 116) = 460.412$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.799$, with the participants presenting lower anxiety levels before performing the task (24.1 ± 2.5) than after they completed it (34.0 ± 2.5). The effect of lying condition was not significant, $F(1, 116) = 0.012$, $p = 0.914$, $BF = 0.19$, $\eta^2_p < 0.001$ (deceptive = 29.1 ± 5.7. truthful = 29.0 ± 5.5), as indicated by both the p value and the BF supporting H_0. The interaction time × lying condition was not significant either, $F(1, 116) = 0.157$, $p = 0.692$, $BF > 100$, $\eta^2_p = 0.001$. Nevertheless, the computed BF supported H_1. Multiple comparisons showed that differences between the 'pre' and 'post' conditions were present within both the truthful condition (24.2 ± 2.4 vs. 33.9 ± 2.7, $p < 0.001$) and the deceptive condition (24.0 ± 2.7 vs. 34.1 ± 2.3, $p < 0.001$).

3.2.3. Cognitive load, subjective arousal and convincing impression

The MANOVA showed that the multivariate effect of lying condition on cognitive load, subjective arousal and convincing impression was significant, $F(1, 298) = 97.754$, $p < 0.001$, $\eta^2_p = 0.498$, Wilks' Lambda = 0.502. Cognitive load was higher in the deceptive condition (4.6 ± 1.7) than in the truthful condition (2.5 ± 1.2), $F(1, 298) = 133.2$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.309$. Subjective arousal was also higher in the deceptive condition (5.1 ± 1.5) than in the truthful condition (3.1 ± 1.6), $F(1, 298) = 85.856$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.224$. On the other hand, higher convincing impression was revealed in the truthful condition (5.6 ± 1.3) than in the deceptive condition (4.7 ± 1.3), $F(1, 298) = 56.333$, $p < 0.001$, $BF > 100$, $\eta^2_p = 0.102$.

3.2.4. Correlations

In Figure 5 we report the correlations between the ROIs and the measured tests for both the truthful and the deceptive conditions. In the truthful condition, significant moderate (for the left) and low (for the right) positive associations were revealed between cognitive load and thermal changes in the sides of the forehead. Furthermore, a significant moderate negative association between subjective arousal and thermal change in the nose was observed. Moreover, significant low (for the left) and moderate (for the right) positive associations were shown between convincing impression and thermal changes in the cheeks.

On the other hand, in the deceptive condition we observed significant high (for the left) and moderate (for the right) positive associations between cognitive load and thermal changes in the sides of the forehead. Moreover, we found a significant moderate negative association between subjective arousal and thermal change in the nose. Furthermore, significant moderate negative associations were revealed between convincing impression and thermal changes in the cheeks. Critically, the anxiety levels (pre and post) were not associated with the thermal change for any of the ROIs in the truthful or deceptive conditions.

3.2.5. Accuracy and false alarms

Considering each ROI described here (see Table 2), the accuracy in detecting the deceptive participants for each ROI calculated separately was 73% for the forehead criterion, 73% for the cheeks criterion and 70% for the nose criterion. The false alarms were 27% for the forehead criterion, 37% for the cheeks criterion and 33% for the nose criterion. The accuracy was higher with the dilemma combination between ROIs, although the false alarms were

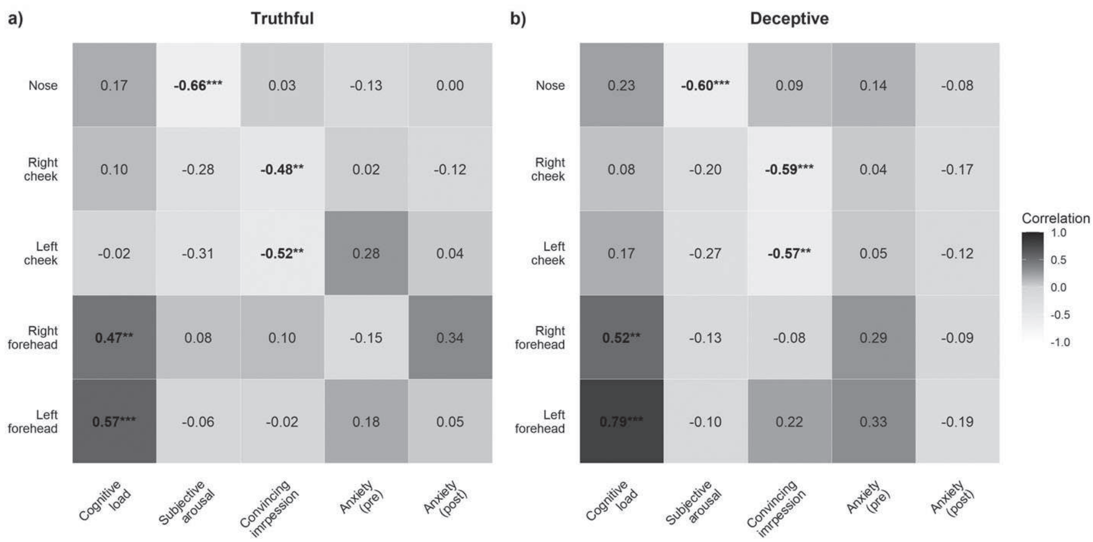

Figure 5. Correlations between the thermal change for each facial ROI and the measured tests within both the truthful condition (panel a) and the deceptive condition (panel b) in Experiment 2. Bold values denote statistical significance ($p < 0.05$*, $p < 0.01$**, $p < 0.001$***).

Table 2. Accuracy and false alarms in Experiment 2 for all the possible combinations of the three facial ROIs selected—lower temperature of the tip of the nose, higher temperature of the forehead, and higher temperature of the cheeks.

ROIs	Accuracy	False Alarms
Forehead	73%	27%
Cheeks	73%	37%
Nose	70%	33%
Forehead or Cheeks	90%	53%
Forehead or Nose	90%	57%
Cheeks or Nose	90%	87%
Forehead or Cheeks or Nose	97%	77%
Forehead plus Cheeks	57%	10%
Forehead plus Nose	50%	3%
Cheeks plus Nose	53%	7%
Forehead plus Cheeks plus Nose	40%	0%
Forehead and/or Cheeks and/or Nose: But at least two of the ROIs	80%	20%

also higher. The criterion with the additive combination between three ROIs showed lower accuracy and fewer false alarms when compared to the other types of combination (e.g. 40% accuracy and 0% for false alarms for forehead plus cheeks plus nose criterion). The results show that the best way to predict who is a liar (having high accuracy and few false alarms) is to consider the criterion of at least two of the ROIs in a mixed model, which gives 80% accuracy and 20% false alarms.

3.3. Discussion

Overall, the results of Experiments 1 were replicated in Experiment 2. Our results support the direct relationship between thermal change and cognitive load, subjective arousal, and convincing impression, adding the increase of the temperature of the cheeks as an index for the detection of deception. Once again, although there was an anxiety effect before and after the session, the STAI anxiety tests (pre and post) did not correlate with the thermal change of any of the ROIs, in either the truthful or the deceptive condition. Furthermore, as in Experiment 1, the anxiety level did not differ between the lying conditions, with the null hypothesis also being supported by the Bayesian Factor. The best combination to detect deception is, once again, the model with 'at least two ROIs detected' (i.e. at least two of the three possible thermal changes expected in the deceptive condition), which showed 80% accuracy and 20% false alarms. In short, the results of Experiment 2 generalize and give consistency to the proposed trifactorial model based on thermography.

4. General discussion

It is quite difficult to distinguish truth from lies, and humans are poor at detecting lies when they base the detection on verbal and non-verbal behavior: their mean accuracy is 54% (DePaulo et al., 2003; Vrij, 2015). Scientific research into the detection of deception must cope with problems of low accuracy and a high number of false alarms. Typically, there is a positive correlation between the accuracy of the detection of deception and false alarms. Here, we focused on a new physiological method for the detection of deception: facial temperature changes, with particular attention to the temperature of the nose (the part of the face most sensitive to changes in temperature and psychological manipulation), the temperature of the forehead (the part of the face least sensitive to changes in temperature and psychological manipulation), and the temperature of the cheeks (which is associated with guilt or shame), social emotions and cognition. We associated these three facial ROIs with specific components of deception: cognitive load, subjective arousal, and convincing impression. In this regard, our results support only the association between thermal changes to the nose and subjective arousal, in line with previous research (Ioannou et al., 2013), and not an association with cognitive load (Abdelrahman et al., 2017).

It is clear now that the difference between telling the truth and lying is a continuum, and that these are not two dichotomous, qualitatively different, categories. We must speak of multidimensional and quantitative differences, or probabilistic differences. If a thermographic model for the detection of deception is to be considered useful in terms of being a starting point for developing a better detection of deception, we should expect high accuracy (clearly not random), and low false alarms (in fact as low as possible). Other models meet the conditions of our starting point to develop a good model for the detection of deception, such as models that require 'only higher temperature in the forehead' or 'higher temperature in forehead and cheeks' (e.g. Moliné et al., 2018). However, the trifactorial model is stricter than the previous approximations and shows high accuracy and a smaller number of false alarms than the previously cited research with open questions. For example, combinations like meeting the forehead criterion or the cheeks criterion have a strong ratio of accuracy to false alarms. However, the criteria requiring 'at least two of the three ROIs to be pictured in the thermogram' (which includes lower nose temperature and/or higher forehead temperature and/or higher cheeks temperature, but with at least two of the thermal changes in the direction expected), presented 83% accuracy (above chance level) and a low 13% false alarms in Experiment 1 and similar rates in Experiment 2.

To summarize, the generalizability and ecological validity of the detection of deception is a big question (Levine, 2018), and laboratory settings and natural settings are very different; there are different natural settings (e.g. criminal scenarios, detecting terrorists in airports, medical simulation of illness, etc.), which have different weights with regard to the three factors cited and different risks of false alarms, with a higher or lower difficulty of obtaining a significant improvement over the random detection of deception. However, the trifactorial model is an advance for the application in natural settings, with high accuracy and few false alarms.

In future studies, we aim to combine our thermographic model for the detection of deception with a task-shifting experimental paradigm (i.e. adding the cognitive cost of the lie in reaction time) for closed questions (Verschuere et al., 2018), and to combine our thermographic model with a cognitive approach (Vrij et al., 2017), and, more generally, with a tactical interview (Ormerod & Dando, 2015), for open questions.

In other work, while we have developed the top-down trifactorial model introduced in this work, we are also developing a bottom-up model based on deep learning. This model will take into account all the pixels of the thermogram. Until now, we have only fed the model with 200 thermograms of deception and 200 thermograms of truth. However, thousands of thermograms (if not more) would be needed. This means that, to follow the developments of the bottom-up model, we will need more time and resources. At the same time, we have begun the process of creating a first inductive model. The idea is to create a state of the art artificial intelligence classifier that can automatically predict, based on a thermal recording, whether a person is lying. The current state of the art image classifiers are based on convolutional neural network architectures. These are a kind of artificial neural network that is trained through 'deep learning' algorithms by feeding it with labeled examples of images or videos. Classifiers of this kind are mostly 'black boxes'. That is, once they have been trained they return a prediction, but it is not explicitly clear what their prediction is based on, apart from being a statistical model based on the training examples.

Our goal is to combine good top-down models, which give us an understanding of the features of the thermal clips that are more predictive of lying, and powerful, automatized inductive models based on massive amounts of data. We expect that this combination will help us improve prediction accuracy.

However, the training of convolutional networks demands a huge amount of training data. This is one of the reasons why we consider it to be so important to create a comprehensive and high-quality database of labeled thermal clips. Such a database would have tremendous value and would be a key asset for a project such as ours.

Data availability

The datasets generated during the current study are available in the Open Science Framework (OSF) repository, accessible at https://osf.io/m65sa/?view_only=3787d23a54d241d7ad4765bb6e30b9c5.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially supported by the National Fund for Scientific and Technological Development (FONDECYT, Chile), project number 1200326, granted by the National Research and Development Agency (ANID, Chile) to Germán Gálvez-García.

ORCID

G. Gálvez-García http://orcid.org/0000-0001-7390-4180

References

Abdelrahman, Y., Velloso, E., Dingler, T., Schmidt, A., & Vetere, F. (2017). Cognitive heat. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *1*(3), 1–20. https://doi.org/10.1145/3130898

Ammer, K., & Formenti, D. (2016). Does the type of skin temperature distribution matter? *Thermology International*, *26*(2), 51–54.

Cho, Y., Bianchi-Berthouze, N., Oliveira, M., Holloway, C., & Julier, S. (2019, September 3–6). *Nose heat: Exploring stress-induced nasal thermal variability through mobile thermal imaging*. 8th international conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, UK (pp. 566–572). IEEE. https://doi.org/10.1109/ACII.2019.8925453

Clay-Warner, J., & Robinson, D. T. (2015). Infrared thermography as a measure of emotion response. *Emotion Review*, *7*(2), 157–162. https://doi.org/10.1177/1754073914554783

Colwell, K., Hiscock-Anisman, C. K., Memon, A., Taylor, L., & Prewett, J. (2007). Assessment criteria indicative of deception (ACID): An integrated system of investigative interviewing and detecting deception. *Journal of Investigative Psychology and Offender Profiling*, *4*(3), 167–180. https://doi.org/10.1002/jip.73

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

Derakhshan, A., Mikaeili, M., Nasrabadi, A. M., & Gedeon, T. (2019). Network physiology of 'fight or flight'response in facial superficial blood vessels. *Physiological Measurement*, *40*(1), 014002. https://doi.org/10.1088/1361-6579/aaf089

Ebisch, S. J., Aureli, T., Bafunno, D., Cardone, D., Romani, G. L., & Merla, A. (2012). Mother and child in synchrony: Thermal facial imprints of autonomic contagion. *Biological Psychology*, *89*(1), 123–129. https://doi.org/10.1016/j.biopsycho.2011.09.018

Engert, V., Merla, A., Grant, J. A., Cardone, D., Tusche, A., & Singer, T. (2014). Exploring the use of thermal infrared imaging in human stress research. *PLoS ONE*, *9*(3), e90782. https://doi.org/10.1371/journal.pone.0090782

Fernández-Cuevas, I., Bouzas Marins, J. C., Arnáiz Lastras, J., Gómez Carmona, P. M., Piñonosa Cano, S., 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. https://doi.org/10.1016/j.infrared.2015.02.007

Gołaszewski, M., Zajac, P., & Widacki, J. (2015). Thermal vision as a method of detection of deception: A review of experiences. *European Polygraph*, *9*(1), 5–24. https://doi.org/10.1515/ep-2015-0001

Harrell Jr, F. E., & Harrell Jr, M. F. E. (2015). Package 'Hmisc'. *CRAN2018*, 235–236. https://doi.org/10.1093/bioinformatics/btv42

Ioannou, S., Ebisch, S., Aureli, T., Bafunno, D., Ioannides, H. A., Cardone, D., Manini, B., Romani, G. L., Gallese, V., Merla, A., & Romanovsky, A. A. (2013). The autonomic signature of guilt in children: A thermal infrared imaging study. *PLoS ONE*, *8*(11), e79440. https://doi.org/10.1371/journal.pone.0079440

Ioannou, S., Gallese, V., & Merla, A. (2014a). Thermal infrared imaging in psychophysiology: Potentialities and limits. *Psychophysiology*, *51*(10), 951–963. https://doi.org/10.1111/psyp.12243

Ioannou, S., Morris, P., Mercer, H., Baker, M., Gallese, V., & Reddy, V. (2014b). Proximity and gaze influences facial temperature: A thermal infrared imaging study. *Frontiers in Psychology*, *5*, 845. https://doi.org/10.3389/fpsyg.2014.00845

Ioannou, S., Morris, P., Terry, S., Baker, M., Gallese, V., & Reddy, V. (2016). Sympathy crying: Insights from infrared thermal imaging on a female sample. *PLoS ONE*, *11*(10), e0162749. https://doi.org/10.1371/journal.pone.0162749

Ito, H., Oiwa, K., & Nozawa, A. (2018, October 21–24). *Face tracking based on temperature distribution of thermal images for real-time psychophysiological states evaluation using facial skin temperature*. International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Bangkok, Thailand (Vol. 3, pp. 6–7). IEEE. https://doi.org/10.1109/ICIIBMS.2018.8549966

Jarlier, S., Grandjean, D., Delplanque, S., N'Diaye, K., Cayeux, I., Velazco, M. I., Sander, D., Vuilleumier, P., & Scherer, K. R. (2011). Thermal analysis of facial muscles contractions. *IEEE Transactions on Affective Computing*, *2*(1), 2–9. https://doi.org/10.1109/T-AFFC.2011.3

Jeffreys, H. (1961). *Theory of probability*. Clarendon Press.

Johnson, J. M., Minson, C. T., & Kellogg Jr, D. L. (2011). Cutaneous vasodilator and vasoconstrictor mechanisms in temperature regulation. *Comprehensive Physiology*, *4*(1), 33–89. https://doi.org/10.1002/cphy.c130015

Kleinberg, B., Arntz, A., & Verschuere, B. (2019). Being accurate about accuracy in verbal deception detection. *PLoS ONE*, *14*(8), e0220228. https://doi.org/10.1371/journal.pone.0220228

Kosonogov, V., De Zorzi, L., Honoré, J., Martínez-Velázquez, E. S., Nandrino, J.-L., Martinez-Selva, J. M., & Sequeira, H. (2017). Facial thermal variations: A new marker of emotional arousal. *PLoS ONE*, *12*(9), e0183592. https://doi.org/10.1371/journal.pone.0183592

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report A-8*. Gainesville.

Levine, T. R. (2018). Ecological validity and deception detection research design. *Communication Methods and Measures*, *12*(1), 45–54. https://doi.org/10.1080/19312458.2017.1411471

Limonero, J. T., Fernández-Castro, J., Soler-Oritja, J., & Álvarez-Moleiro, M. (2015). Emotional intelligence and recovering from induced negative emotional state. *Frontiers in Psychology*, *6*, 816. https://doi.org/10.3389/fpsyg.2015.00816

Marins, J. C. B., Formenti, D., Costa, C. M. A., de Andrade Fernandes, A., & Sillero-Quintana, M. (2015). Circadian and gender differences in skin temperature in militaries by thermography. *Infrared Physics & Technology*, *71*, 322–328. https://doi.org/10.1016/j.infrared.2015.05.008

Meibauer, J. (2018). The linguistics of lying. *Annual Review of Linguistics*, *4*(1), 357–375. https://doi.org/10.1146/annurev-linguistics-011817-045634

Moliné, A., Fernández-Gómez, J., Moya-Pérez, E., Puertollano, M., Gálvez-García, G., Iborra, O., & Gómez-Milán, E. (2018). Skin temperature reveals empathy in moral dilemmas: An experimental thermal infrared imaging study. *Thermology International*, *28*(4), 197–206.

Moliné, A., Gálvez-García, G., Fernández-Gómez, J., De la Fuente, J., Iborra, O., Tornay, F., Mata Martín, J. L., Puertollano, M., & Gómez Milán, E. (2017). The Pinocchio effect and the cold stress test: Lies and thermography. *Psychophysiology*, *54*(11), 1621–1631. https://doi.org/10.1111/psyp.12956

Moreira, D. G., Costello, J. T., Brito, C. J., Adamczyk, J. G., Ammer, K., Bach, A. J. E., Costa, C. M. A., Eglin, C., Fernandes, A. A., Fernández-Cuevas, I., Ferreira, J. J. A., Formenti, D., Fournet, D., Havenith, G., Howell, K., Jung, A., Kenny, G. P., Kolosovas-Machuca, E. S., Maley, M. J., … Sillero-Quintana, M. (2017). Thermographic imaging in sports and exercise medicine: A Delphi study and consensus statement on the measurement of human skin temperature. *Journal of Thermal Biology*, *69*, 155–162. https://doi.org/10.1016/j.jtherbio.2017.07.006

Morey, R. D., Rouder, J. N., & Jamil, T. (2018). BayesFactor: Computation of Bayes Factors for common designs. R package version 0.9. 12-4.2.

Or, C. K., & Duffy, V. G. (2007). Development of a facial skin temperature-based methodology for non-invasive mental workload measurement. *Occupational Ergonomics*, *7*, 83–94.

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), 76–84. https://doi.org/10.1037/xge0000030

Oxburgh, G. E., Myklebust, T., & Grant, T. (2010). The question of question types in police interviews: A review of the literature from a psychological and linguistic perspective. *International Journal of Speech Language and the Law*, *17*(1), 45–66. https://doi.org/10.1558/ijsll.v17i1.45

Pacheco-Unguetti, A. P., Acosta, A., Callejas, A., & Lupiáñez, J. (2010). Attention and anxiety. *Psychological Science*, *21*(2), 298–304. https://doi.org/10.1177/0956797609359624

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), 35174. https://doi.org/10.1038/srep35174

Paolini, D., Alparone, F. R., Cardone, D., van Beest, I., & Merla, A. (2016). "The face of ostracism": The impact of the social categorization on the thermal facial responses of the target and the observer. *Acta Psychologica*, *163*, 65–73. https://doi.org/10.1016/j.actpsy.2015.11.001

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

Pavlidis, I., Levine, J., & Baukol, P. (2000, June 16). *Thermal imaging for anxiety detection*. Proceedings IEEE workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications (Cat. No.PR00640), Hilton Head, SC (pp. 104–109). IEEE. https://doi.org/10.1109/CVBVS.2000.855255

Pollina, D. A., Dollins, A. B., Senter, S. M., Brown, T. E., Pavlidis, I., Levine, J. A., & Ryan, A. H. (2006). Facial skin surface temperature changes during a "concealed information" test. *Annals of Biomedical Engineering*, *34*(7), 1182–1189. https://doi.org/10.1007/s10439-006-9143-3

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. https://doi.org/10.2307/271063

Ring, E. F. J. (2007). The historical development of temperature measurement in medicine. *Infrared Physics & Technology*, *49*(3), 297–301. https://doi.org/10.1016/j.infrared.2006.06.029

Ring, E. F. J., & Ammer, K. (2000). The technique of infrared imaging in medicine. *Thermology International*, *10*(1), 7–14. https://doi.org/10.1088/978-0-7503-1143-4ch1

RStudio Team. (2016). *RStudio: Integrated development for R. RStudio, Inc.* http://www.rstudio.com/

Salazar-López, E., Domínguez, E., Juárez Ramos, V., de la Fuente, J., Meins, A., Iborra, O., Gálvez, G., Rodríguez-Artacho, M. A., & Gómez-Milán, E. (2015). The mental and subjective skin: Emotion, empathy, feelings and thermography. *Consciousness and Cognition*, *34*, 149–162. https://doi.org/10.1016/j.concog.2015.04.003

Shastri, D., Merla, A., Tsiamyrtzis, P., & Pavlidis, I. (2009). Imaging facial Signs of Neurophysiological Responses. *IEEE Transactions on Biomedical Engineering*, *56*(2), 477–484. https://doi.org/10.1109/TBME.2008.2003265

Spielberger, C. D., Gorsuch, R. L., Lushene, R. E., & Cubero, N. S. (1999). *STAI: Cuestionario de ansiedad estado-rasgo*. TEA ediciones.

Verschuere, B., Köbis, N. C., Bereby-Meyer, Y., Rand, D., & Shalvi, S. (2018). Taxing the brain to uncover lying? Meta-analyzing the effect of imposing cognitive load on the reaction-time costs of lying. *Journal of Applied Research in Memory and Cognition*, *7*(3), 462–469. https://doi.org/10.1016/j.jarmac.2018.04.005

Vrij, A. (2015). Verbal lie detection tools: Statement validity analysis, reality monitoring and scientific content analysis. In P. A. Granhag, A. Vrij, & B. Verschuere (Eds.), *Detecting deception: Current challenges and cognitive approaches* (pp. 1–35). Wiley-Blackwell. https://doi.org/10.1002/9781118510001.ch1

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1–2), 39–43. https://doi.org/10.1002/jip.82

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. *Legal and Criminological Psychology*, *22*(1), 1–21. https://doi.org/10.1111/lcrp.12088

Vrij, A., & Verschuere, B. (2013). *Lie detection in a forensic context*. Oxford University Press.

Walløe, L. (2016). Arterio-venous anastomoses in the human skin and their role in temperature control. *Temperature*, *3*(1), 92–103. https://doi.org/10.1080/23328940.2015.1088502

World Medical Association. (2013). WMA Declaration of Helsinki: Ethical principles for medical research involving human subjects. *Journal of the American Medical Association*, *310*(20), 2191–2194. https://doi.org/10.1001/jama.2013.281053