Journal of Experimental Social Psychology 115 (2024) 104658

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

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

Face masks facilitate discrimination of genuine and fake smiles – But people believe the opposite☆

Haotian Zhou [a], Meiying Wang [b], Yu Yang [a,*], Elizabeth A. Majka [c,**]

[a] ShanghaiTech University, China
[b] London Business School, United Kingdom
[c] Elmhurst University, USA

ARTICLE INFO

Keywords:
Mask
Genuine smile
Fake smile
Facial expressions
Deception detection
Cross-cultural difference

ABSTRACT

It seems a foregone conclusion that face mask-wearing hinders the interpretation of facial expressions, increasing the risk of interpersonal miscommunication. This research identifies a notable counter-case to this apparent truism. In multiple experiments, perceivers were more accurate distinguishing between genuine and fake smiles when the mouth region was concealed under a mask versus exposed. Masks improved accuracy by shielding perceivers from the undue influence of non-diagnostic cues hidden behind masks. However, perceivers were unaware of the advantage bestowed by masks, holding, instead, the misbelief that masks severely obscure the distinction between genuine and fake smiles. Furthermore, these patterns proved to be culturally invariant rather than culturally contingent, holding true for both Westerners and Easterners.

1. Introduction

As a result of the COVID-19 pandemic, wearing face masks is ubiquitous. Although most people recognize the health benefits of masks, it has been suggested that they are concerned about masks compromising nonverbal communication (Campagne, 2021; Matuschek et al., 2020; Ramdani et al., 2022). After all, masks conceal a sizable portion of one of our most powerful social communication tools — the face (Ekman & Oster, 1979; Gill et al., 2013; Jack & Schyns, 2015; Todorov et al., 2015). The portion of the face concealed by a mask consists of the mouth and several surrounding areas and is collectively referred to as the *mouth region* (or the *mouth*) in this report for convenience. Comprising more than half of the 20 facial muscles, the mouth region contributes to the production of a wide range of facial expressions, transmitting a wealth of social information (Sendic, 2022). In particular, of the 28 *main* action units (AUs, the basic units of facial movements) codified under the Facial Action Coding System (FACS) (Ekman & Friesen, 1978; RealEye, 2024), 17 are unambiguously instantiated through muscle activation within the mouth region. For reference, consider what we metonymically refer to as the *eye region* (or the *eyes*), the part of a face concealed by a pair of oversized sunglasses. Only about two of the 20 facial muscles reside in the eye region (Sendic, 2022); and correspondingly, just six of the 28 main AUs are grounded in muscles within this region. In fact, the 2002 edition of the FACS Manual has four chapters on lower-face AUs but a single chapter on upper-face AUs (Ekman et al., 2002). Taken together, it seems a foregone conclusion that people will be less accurate in "reading" others' faces if the mouth region is hidden behind a mask. However, this may not always be the case.

In principle, the seeming truism that mask-wearing undermines social perception ought to break down in a circumstance in which the information provided by the mouth region is non-diagnostic. Generally, a cue (e.g., Fido has four legs) is considered nondiagnostic if it is equally likely to be observed under each of the alternative hypotheses under consideration (e.g., Fido is a cat versus Fido is a dog). We posit that such a circumstance may occur when people are tasked with judging the authenticity of the most frequently displayed facial expression—the smile (Calvo et al., 2014). Importantly, just as in cases in which people try to discern whether someone is lying, people may hold faulty beliefs about which information is diagnostic when judging smile authenticity (Hartwig & Granhag, 2015; Zuckerman et al., 1981). As a result, we predict that people will mistakenly believe face masks impair the discrimination of genuine and fake smiles when—in fact—the opposite

☆ This paper has been recommended for acceptance by Dr Rachael Jack.

* Corresponding author at: School of Entrepreneurship and Management, ShanghaiTech University, Shanghai, China.
** Corresponding author at: Department of Psychology, Elmhurst University, Elmhurst, IL, USA.
E-mail addresses: connect2yu@gmail.com (Y. Yang), liz.majka@elmhurst.edu (E.A. Majka).

https://doi.org/10.1016/j.jesp.2024.104658
Received 10 April 2023; Received in revised form 15 June 2024; Accepted 17 June 2024
Available online 26 June 2024

will occur.

1.1. Smiles, genuine or fake

A smile involves the contraction of the zygomatic major muscle that lifts the corners of the mouth (Frank et al., 1993). Although there are different views on the types and function of smiles, many scholars maintain that people express different smiles depending on how they are feeling and in response to contextual influences (e.g., Frank et al., 1993; Martin et al., 2017; Rychlowska et al., 2017). One common approach is to classify smiles as genuine or fake (for a discussion, see Martin et al., 2017).

Proponents of this binary classification assert that genuine smiles (sometimes referred to as real, true, Duchenne, enjoyment, involuntary, spontaneous, or reward smiles) are expressed when people are experiencing positive affect or happiness (Frank & Ekman, 1993; Miles & Johnston, 2007; Niedenthal et al., 2010; Rychlowska et al., 2017; Sheldon et al., 2021a). In addition to using the *zygomatic major* muscle around the mouth, genuine smiles also recruit the *orbicularis oculi* muscle that raises the cheeks, creating the wrinkles around the eyes often referred to as "crow's feet" (Duchenne & de Boulogne, 1990; Frank & Ekman, 1993). By contrast, fake smiles (sometimes referred to as deceptive, false, phony, non-Duchenne, non-enjoyment, voluntary, posed, or social smiles) are expressed in the absence of positive affect or happiness (Ekman & Friesen, 1982; Miles & Johnston, 2007; Niedenthal et al., 2010; Sheldon et al., 2021b) and for a wide range of reasons, such as to conceal negative emotions or be polite (Ekman et al., 1988). Therefore, fake smiles represent a broader category of smiles. Fake smiles use the *zygomatic major* muscle around the mouth, but do not recruit the *orbicularis oculi* muscle (Ekman & Friesen, 1982; Frank & Ekman, 1993; Sheldon et al., 2021b).

Being able to distinguish between genuine and fake smiles is a critical social skill. In a gift-giving context, for example, classifying a smile as genuine or fake could mean the difference, for example, between correctly inferring that your friend is truly happy with your gift (correctly identifying a smile as genuine) and failing to realize that your friend is hoping you included a gift receipt (incorrectly identifying a smile as genuine). People can and do distinguish between genuine and fake smiles (Frank et al., 1993), particularly when socially motivated (Bernstein et al., 2008; Schindler & Trede, 2021). Moreover, they use their perceptions of smile authenticity to inform other social judgments, for example, by judging those displaying genuine smiles (vs. fake) as happier (Miles & Johnston, 2007) and evaluating them more positively on a wide range of interpersonal traits (Frank et al., 1993; Gunnery & Ruben, 2016; Johnston et al., 2010).

1.2. Masks and discerning smile authenticity

Returning to the topic of masked smiles, for a perceiver mainly concerned with discerning the authenticity of a smile, the presence of a mask should constitute no more than a harmless piece of fabric. After all, the diagnostic cues critical to the perceiver's success reside primarily in the region beyond the mask (i.e., the eyes), whereas the occluded region (i.e., the mouth) offers little diagnostic value. Therefore, the perceiver's accuracy should not be affected by whether the smile under consideration is hidden by a mask. Interestingly, this normative claim may fail as a prediction of what empirically transpires. In fact, a case could even be made that the perceiver's accuracy may paradoxically improve when the mouth of a smile is concealed by a mask.

As in other interpersonal judgments (e.g., deception detection), the extent to which people are successful in discriminating between genuine and fake smiles may depend not only on actual diagnostic cues (e.g., movement in the eye region) but also on people's beliefs regarding which cues are diagnostic vis-a-vis non-diagnostic (Hartwig & Granhag, 2015; Zuckerman et al., 1981). A meta-analysis of around 25,000 deception judgments found that the average accuracy rate of human judges was 54%, barely exceeding the chance baseline (Bond & DePaulo, 2006). One particular explanation put forth in the literature for such underwhelming accuracy is that people, including professional lie catchers, such as police officers, usually subscribe to erroneous beliefs about the cues to deception.

Applying this framework to the task of distinguishing genuine from fake smiles—and considering the effect of face masks—two critical questions emerge: 1) What are people's beliefs about cue diagnosticity, particularly the facial region concealed by masks (e.g., mouth)? and 2) Given such beliefs, what is the effect of concealing the mouth (via masking) on judgmental accuracy?

In response to these two interconnected questions, we propose a descriptive model that speaks to the role masks may play in the judgment of genuine and fake smiles. Below we develop and articulate two theoretically derived hypotheses that we comprehensively test in this research: the *maximum confidence-loss* hypothesis and the *ironic performance-gain* hypothesis. These two hypotheses, in tandem, predict that people will mistakenly believe face masks interfere with the discrimination of genuine and fake smiles when–in fact–the opposite will occur.

1.3. Theoretical justifications

1.3.1. The maximum confidence-loss hypothesis

Lay people can discriminate between genuine and fake smiles considerably better than chance, but it is likely that their judgments are based on intuition rather than deliberation (Gigerenzer, 2022; Kruglanski & Gigerenzer, 2011). Existing evidence suggests that people probably lack explicit knowledge of the perceptual cues that distinguish genuine and fake smiles (Frank et al., 1993; Mai et al., 2011). In the absence of explicit knowledge, people tend to overgeneralize, extrapolating information from a familiar, seemingly similar context – a tendency well documented in the judgment and decision-making literature (Cimpian et al., 2010; Leslie et al., 2011; Sutherland et al., 2015; Williams et al., 2013). Since judging the authenticity of a smile is tantamount to making a happy/not happy judgment, when people construct ad hoc beliefs about how concealing the mouth (e.g., via masking) will impact their accuracy in discriminating genuine/fake smiles, they likely generalize from their beliefs about the role of the mouth in emotion recognition in general.

In line with existing literature (Blais et al., 2012), we posit that people believe the mouth region to be more critical than any other face region (including the eyes) for deciphering facial expressions (e.g., happy, surprised, fearful, angry, disgusted, sad). After all, the mouth region, as aforementioned, is the most articulate and innervated part of a face. In fact, when decoding facial expressions, people tend to *under*-utilize facial areas other than the mouth, even if they are informative (Blais et al., 2012). Unreflectively extrapolating this general belief about the importance of the mouth to the particular case of discerning smile authenticity is apt to result in the non-normative sentiment described by the maximum confidence-loss hypothesis.

Maximum confidence-loss hypothesis: Relative to judging exposed smiles, perceivers will lose confidence in their genuine/fake smile discrimination accuracy when the non-diagnostic mouth region is concealed (e.g., by a mask). Furthermore, the confidence loss incurred as a result of concealing the mouth will exceed the confidence loss incurred as a result of concealing other facial areas, even the truly diagnostic eye region.

1.3.2. The ironic performance-gain hypothesis

Given that the mouth region is non-diagnostic with respect to the distinction between genuine and fake smiles, concealing the mouth should presumably be analogous to subtracting zero from an equation, and should therefore be devoid of any causal effect. However, it is well established that an otherwise innocuous nondiagnostic cue can result in judgment errors if perceivers are unaware of its non-diagnosticity and take it into account, either intentionally or inadvertently, in the

judgment process (Camerer et al., 1989; Fischhoff, 1975; Hall et al., 2007; Nisbett et al., 1981; Tversky & Kahneman, 1974). On the flip side, veridical judgments have been shown to correlate with perceivers' ability to isolate and ignore nondiagnostic cues (Bogaard & Meijer, 2018; Ettenson et al., 1987; Jarodzka et al., 2010).

Thus, denying people access to non-diagnostic cues can be beneficial rather than causally neutral, provided that people are incognizant of or even mistaken about the lack of diagnostic value of these cues. By extension, the notion that concealing the mouth region can facilitate the authenticity judgment of a smile is not without merit. This is especially true if people wrongly think of the mouth as the most important source of authenticity markers, as posited by the maximum confidence-loss hypothesis.

Ironic performance-gain hypothesis: Perceivers will be more accurate in discerning the authenticity of a smile when the mouth region of the smile is concealed (e.g., by a mask) versus fully exposed, an effect we refer to as *disclosing-by-masking* effect.

1.4. Culturally contingent or culturally invariant?

Thus far, our theorizing has remained largely agnostic to culture. However, in light of existing literature on cross-cultural differences in emotion recognition (Jack et al., 2009; Yuki et al., 2007), there are reasons to take seriously the possibility of our descriptive model being culturally contingent. That is, our model may not be able to adequately accommodate Eastern experiences.

Members of Eastern (or interdependent) cultures, such as Chinese, Japanese, and Koreans, are known to downplay or even suppress outward expressions of emotions for fear of disrupting social harmony (Ford & Mauss, 2015). However, because muscles within the eye region are usually less amenable to intentional control than those in the mouth region (Matsumoto & Lee, 1993), inner feelings can still be betrayed by involuntary movements around the eyes despite a person's efforts to inhibit them. These observations led Yuki et al. (2007) to propose and evaluate the notion that Easterners should rely on the eyes more than the mouth when decoding others' emotions. Their empirical results demonstrate a cultural bifurcation in how mouths versus eyes are utilized when reading emotions from facial expressions: While members of Western (or independent) cultures give more weight to the mouth region than to the eye region, Easterners "focus more strongly on the eyes than the mouth" (p. 303). This claim, which we refer to as the YMM thesis after the initials of the authors, can have important implications for both constituent hypotheses of our descriptive model, assuming that discerning smile authenticity amounts to inferring if someone is experiencing happiness (or joy) and therefore could be considered a special case of emotion perception.

On the one hand, suppose that Eastern cultures have indeed instilled in their members the belief that the mouth is generally less informative than the eye region for inferring emotions. Given this premise, when decoding facial expressions of emotions, Eastern perceivers should generally be *more* worried about losing access to the eyes than losing access to the mouth. If Easterners apply their beliefs about the relative importance of the eyes versus the mouth to emotion recognition in general to the specific context of discerning the authenticity of smiles, it follows that they may lose confidence the most when the eye region is concealed (instead of when the mouth is concealed). In other words, the maximum confidence-loss hypothesis may *not* characterize the beliefs of Easterners.

On the other hand, suppose that when interpreting facial expressions, Easterners are indeed highly practiced at discounting or even ignoring the mouth because movements there tend to be severely censored per cultural norms. Then, when it comes to discerning smile authenticity, Easterners, compared to their Western counterparts, would benefit markedly less, if at all, from being denied access to the mouth. After all, they presumably would disregard this region even if it were accessible. Therefore, the ironic performance-gain hypothesis may not hold among Easterners.

In short, if we accept both the YMM thesis and the premise that smile authentication exemplifies emotion perception, our descriptive model may only be descriptive of Western populations. Thus, neither the maximum confidence-loss hypothesis nor the ironic performance-gain hypothesis should be expected to generalize to members of Eastern cultures. For convenience, the conjecture that the validity of both constituent hypotheses of our descriptive model is contingent on the culture under consideration is termed the *culture-contingency meta-hypothesis*. It is important to note that our assumption that discerning smile authenticity can be equated to a special case of emotion perception is open to debate. For example, it might entail no more than pattern matching based on low-level visual features.

Potential counterpoints. Interestingly, there is evidence in the existing literature suggesting exceptions to the YMM thesis. That is, there might be situations where perceivers in East and West may not vary, at least not qualitatively, in how they differentially rely on the mouth versus the eyes for interpreting facial expressions. As pertains to the present study, some exceptions call into question the tenability of the culture-contingency meta-hypothesis.

Consider a recent study (Snoek et al., 2023), where the authors devised an innovative procedure to triangulate, for Eastern and Western perceivers separately, AUs that ground the decision to attribute one of the six basic emotional states to a given facial expression. They found that AUs whose presence was critical to making a *happiness* attribution hardly vary between cultures. Specifically, the five AUs that emerged as the most critical to perceiving facial displays as expressions of happiness are nearly identical for members of both cultures. More importantly, all of these AUs consist mainly of muscle movements located within the mouth region. Assuming that discerning smile authenticity could be equated to judging if someone is experiencing happiness, the central role played by the mouth in encoding and decoding happiness in both cultures suggests that Easterners tasked with distinguishing between genuine and fake smiles may not be exempted from either the maximum confidence-loss hypothesis or the ironic performance-gain hypothesis. In other words, the plausibility of our cultural-contingency meta-hypothesis is questionable.

In light of such countervailing evidence, we conducted an ancillary study as a preliminary evaluation of the viability of the culture-contingency meta-hypothesis. Capitalizing on recent advances in computational linguistics (Grand et al., 2022; Pereira et al., 2016), we estimated, separately for English and Chinese, the strength of semantic associations between words referring to the concept of smiles (e.g., "smile" and "笑容") and words referring to different facial regions (e.g., "mouth" and "嘴"). For both Chinese and English, we selected four words denoting the concept of smiles and five words denoting five non-overlapping facial regions, i.e., the eyes, ears, nose, mouth, and jaw. Semantic associations were measured in terms of the geometric proximity (i.e., cosines of the angles) between high-dimensional vector representations (also known as embeddings) of these words. These numerical representations of words are derived from word distribution statistics in large-scale English- or Chinese-specific natural language corpora consisting of billions of words (Pennington et al., 2014). Further details on the methodology followed by this ancillary study can be found in the Supplementary Materials under the heading "Cross-cultural linguistic analysis."

In both languages, all four smiling-related words are more closely associated with the word "mouth" than with the word "eye" (Fig. 1). This pattern suggests that the mental model of a smile for English speakers is similar to that of Chinese speakers, with a greater emphasis on the mouth than the eyes. To the extent that English- and Chinese-speaking communities are representative of Western and Eastern cultures, it follows that the unadaptive tendency to overvalue the mouth but undervalue the eyes when discerning smile authenticity may be shared cross-culturally.

Astute readers may notice that the words that denote smiles in our

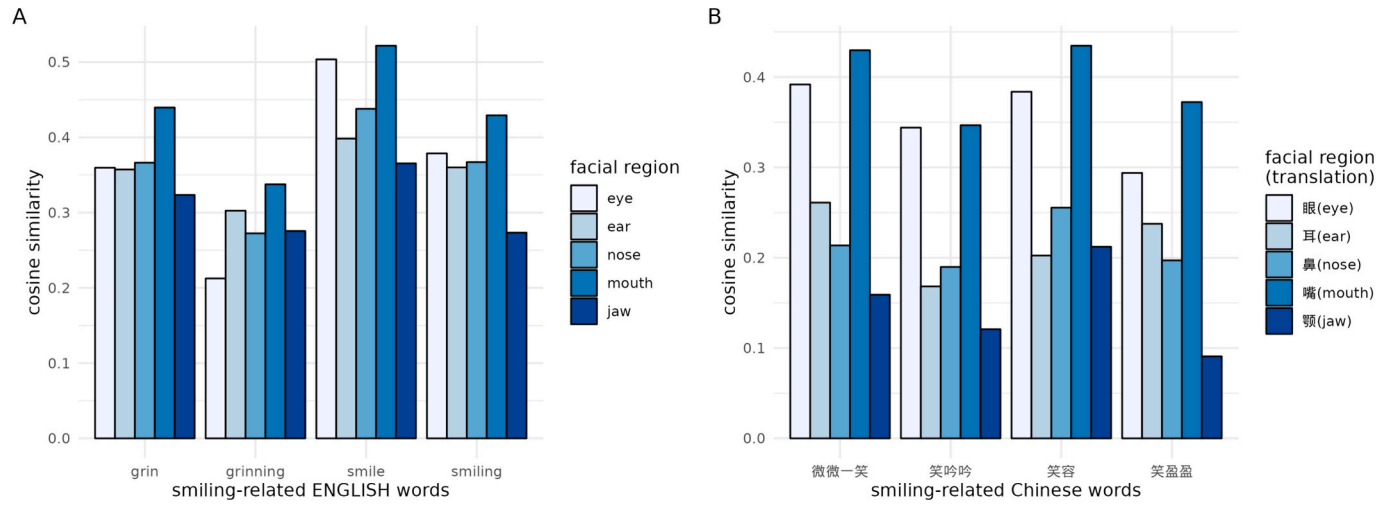

Fig. 1. A bar graph of the geometric proximity between the embeddings of each smile-related word and each word referring to a different facial region. The results for words in English and Chinese are shown separately in Panels A and B. Geometric proximity is measured by the cosine of the angle formed by the vector representations (i.e., embeddings) of the two words under consideration (e.g., grin versus nose).

analysis are generic terms that do not distinguish between genuine and fake smiles. Thus, it is unclear whether and to what extent these findings can inform the discussion about the authenticity of smiles. This is a valid concern, and the reason we did not examine multi-word terms that unambiguously specify authenticity (e.g., "genuine smile" or "真诚的笑容") was purely technological. The pretrained word embedding models from which we obtained the embeddings used in our analysis are designed to learn embeddings for single words almost exclusively. In other words, these models do not "understand" multi-word terms. Consequently, it was impossible to extract from these models embeddings for multi-word terms that incorporate authenticity markers. Although this caveat with our approach should not be taken lightly, the pattern shown in Fig. 1 can still be instructive if, when "smile" is mentioned without any modifier, laypeople are inclined to assume it to denote the genuine type. This genuine-by-default tendency is actually consistent with the truth-default theory in the deception-detection literature, which argues that humans have the innate propensity to assume the honesty of most incoming messages they receive (Levine, 2014). A recent study even provides empirical support for this genuine-by-default tendency as pertains to the perception of smiles (Mui et al., 2020).

Despite the rebuttals from both existing literature (Snoek et al., 2023; also see Jack et al., 2016) and our preliminary analysis, it is still worthwhile to properly put the cultural-contingency meta-hypothesis to empirical test, given the indirectness of these counterpoints.

1.5. Overview

To recap, the present research investigates a potential counter-case to the seeming truism that mask-wearing is detrimental to the interpretation of facial expressions. Specifically, we investigate both the subjective beliefs and objective reality of face masks' effect on distinguishing between genuine and fake smiles. On the belief side, we hypothesize that concealing the mouth region (e.g., with a mask) will undermine perceivers' confidence in their smile discrimination accuracy to a greater extent than concealing any other facial region, including the indispensable eye region (i.e., the maximum confidence-loss hypothesis). On the reality side, we hypothesize that concealing the mouth region (e.g., with a mask) will improve rather than impair perceivers' smile discrimination accuracy (i.e., the ironic performance-gain hypothesis). Furthermore, we *tentatively* postulate that both hypotheses are bounded culturally, neither of which is expected to accommodate the Eastern experiences satisfactorily (i.e., the cultural-contingency meta-hypothesis).

Studies 1 and 2 focus on the maximum confidence-loss hypothesis and the ironic performance-gain hypothesis, respectively. Both studies drew samples from the West and East to evaluate the cultural-contingency meta-hypothesis. Study 3 combines core elements from the two previous studies to ensure the robustness of our results and to shed light on certain more nuanced aspects of the descriptive model we investigate. Studies 4 and 5, respectively, address the underlying mechanism and ecological validity of the disclosing-by-masking effect, the essence of the ironic performance-gain hypothesis.

2. Study 1: Confidence Loss

2.1. Methods and procedures

2.1.1. Purposes and rationales

Study 1 evaluates the maximum confidence-loss hypothesis by testing its two corollary predictions: (1) perceivers will lose confidence in judging the authenticity of a smile if the mouth region is concealed by a mask, even though this region is of little diagnostic value; and (2) the confidence loss incurred as a result of concealing the mouth will exceed the loss incurred as a result of concealing any other facial region, including the truly diagnostic eye region. Both corollaries are tested in a Western sample and an Eastern one to evaluate the cultural-contingency meta-hypothesis, which predicts that Corollary (2) will not hold in the Eastern sample.

2.1.2. Transparency and openness

We report all data exclusions, all manipulations, and all measures in the studies. Data were analyzed using R, version 4.2.2. The design and hypotheses were preregistered; see https://aspredicted.org/ANA_QQW. The sample size was pre-determined as in the pre-registration. Data collection was not continued after data analysis. All exclusion criteria and deviations from the pre-registered plan are reported in the Supplementary. Survey materials, data, code, and pre-registration documents have been made publicly available at OSF and can be accessed at https://osf.io/8pbz5/?view_only=aac8c24f4b684a03b6862a20b43299d7.

2.1.3. Participants

Participants were recruited from the United States and China, two geographic regions that represent Western and Eastern cultures, respectively. 148 self-identified Americans (N_{female} = 85, $Median_{age}$ =

38 years) from Mechanical Turk (MTurk), a crowd-sourcing platform based in America, and 161 self-identified Chinese (N_{female} = 86, $Median_{age}$ = 32 years) from SoJump, a crowd-sourcing platform based in China, completed this study in exchange for monetary compensation. Depending on their nationalities, participants saw the English version or the Chinese version of the same online survey programmed in Qualtrics. The final sample size provided 90% power to detect an effect size of η_p^2 = 0.014 or greater in a one-way ANOVA test with three repeated measures of confidence-change scores and a 5% false-positive rate (assuming correlation among measures as r = 0.5).

2.1.4. Stimuli and measures

The central part of the survey asked participants to *imagine* performing a smile-authentication task where they would need to judge the authenticity of multiple smiles. The introduction to the hypothetical task is reproduced verbatim below:

> We have filmed 20 different people displaying a smile. Some of the people are showing a real (genuine) smile, whereas others are showing a fake smile. Imagine we had you watch these 20 short videos. After watching EACH video, you would have to judge whether the person in that video was showing a real (genuine) or fake smile.

After learning the task, participants forecasted on a 9-point Likert scale how well they would perform under various conditions. The two poles of the scale (i.e., 1 and 9) were respectively labeled "I will be hardly better than random guessing" and "I will be perfect or almost perfect." Each participant forecasted his/her confidence levels for the same set of four visibility conditions: baseline, sans-forehead, sans-eye, and sans-mouth. What each condition entailed was explained both verbally and graphically. The four visibility conditions differed in the facial region of the targets (i.e., people filmed in these videos) that would be concealed during playback. Fig. 2 is the schematic illustration of the four visibility conditions provided to participants.

In the baseline condition, these videos would be shown in their original forms, *without* any part of the targets' faces being concealed. In contrast, an accessory item would be digitally superimposed to conceal a particular part of the targets' faces in each of the other three conditions. Specifically, in the sans-forehead/eye/mouth condition, a bandanna/pair of oversized sunglasses/medical mask would be edited in to occlude the forehead/eye/mouth region during playback. Note that these three FROIs (short for *facial regions of interest*), as delineated in Fig. 2, are more or less mutually exclusive and jointly exhaustive of the entirety of a typical face. For brevity, we employ `\FROI` as a shorthand for the treatment of denying perceivers visual access to a certain FROI, with `\FOREHEAD`, `\EYE`, and `\MOUTH` denoting the treatments administered in the three sans-FROI conditions.

Instead of working with the raw confidence forecasts, we computed for each participant three treatment-specific `confidence-change` scores by subtracting his/her forecast for the baseline condition from the forecasts for each of the three sans-FROI conditions. If an individual believed that a particular treatment (e.g., `\MOUTH`) would be detrimental to performance on the task, their `confidence-change` score specific to that treatment (e.g., `confidence-change\MOUTH`) would be *lower* than zero, indicating a *loss* in confidence. These `confidence-change\FORI` scores presumably measured participants' prior beliefs about the impact of concealing a certain FROI (as opposed to leaving it exposed) on their ability to discern smile authenticity.

Fig. 2. Schematic illustration of all four visibility conditions participants were asked to consider (from left to right: the baseline, sans-forehead, sans-eye, and sans-mouth conditions).

Several additional measures were collected in the same survey. We do not elaborate on these measures, since they were not related to the focal hypotheses of the present research.

2.2. Results and discussion

Fig. 3 displays the summary statistics of treatment-specific `confidence-change` scores, broken down by culture. We examine how the `confidence-change` scores varied as a function of treatment within each cultural group separately.

2.2.1. Western culture (America)

For Americans, the prospect of losing access to the mouth caused a considerable loss of confidence in making veridical authenticity judgments. The mean `confidence-change\MOUTH` score was significantly lower than zero ($M = -4.26$, $SD = 2.53$), $t(147) = -20.43$, $p < .001$, *Cohen's d* $= -1.68$, 95%*CI* $[-1.94, -1.43]$. Pairwise comparisons revealed that the confidence loss specific to `\MOUTH` exceeded not only the loss specific to `\FOREHEAD` ($M = -0.93$, $SD = 1.30$), $t(147) = -16.69$, $p < .001$, *Cohen's d* $= -1.66$, 95%*CI* $[-1.94, -1.39]$, but also the loss specific to `\EYE` ($M = -2.89$, $SD = 1.98$), although the eye region is indispensable for differentiating between genuine and fake smiles, $t(147) = -6.56$, $p < .001$, *Cohen's d* $= -0.68$, 95%*CI* $[-0.90, -0.46]$. Thus, the maximum confidence-loss hypothesis was supported in the Western sample.

2.2.2. Eastern culture (China)

Qualitatively speaking, the `\FORI`-specific `confidence-change` scores measured on Chinese participants show an identical pattern as the scores measured on their American counterparts. The mean `confidence-change\MOUTH` score was significantly lower than zero ($M = -2.88$, $SD = 2.76$), $t(160) = -13.23$, $p < .001$, *Cohen's d* $= -1.04$, 95%*CI* $[-1.24, -0.85]$. Furthermore, it represented a greater confidence loss than both the loss specific to `\FOREHEAD` ($M = -0.34$, $SD = 1.52$), $t(160) = -12.69$, $p < .001$, *Cohen's d* $= -1.11$, 95%*CI* $[-1.32, -0.90]$, and the loss specific to `\EYE` ($M = -2.11$, $SD = 2.42$), $t(160) = -4.42$, $p < .001$, *Cohen's d* $= -0.34$, 95%*CI* $[-0.49, -0.18]$. Thus, the maximum confidence-loss hypothesis was also confirmed in the Eastern sample.

Although the maximum confidence-loss hypothesis was fully compatible with observations on both cultural groups, the cultural-contingency meta-hypothesis could still be partly justified if the other component of our descriptive model, i.e., the ironic performance-gain hypothesis, was culturally contingent.

3. Study 2: Performance Gain

3.1. Methods and procedures

3.1.1. Purposes and rationales

Study 2 evaluates the ironic performance-gain hypothesis by testing the paradoxical disclosing-by-masking effect at its core. The effect is compared between the West and East to find out if the cultural-contingency meta-hypothesis might hold any water. Support for the meta-hypothesis could assume either a strong form or a weak form. Strong support would be obtained if the disclosing-by-masking effect was *only* observed in the West. Weak support would be obtained if the effect was observed in both cultures but was considerably *less* pronounced in the East. Furthermore, this study explores whether people are capable of insight into the actual diagnosticity of the mouth region

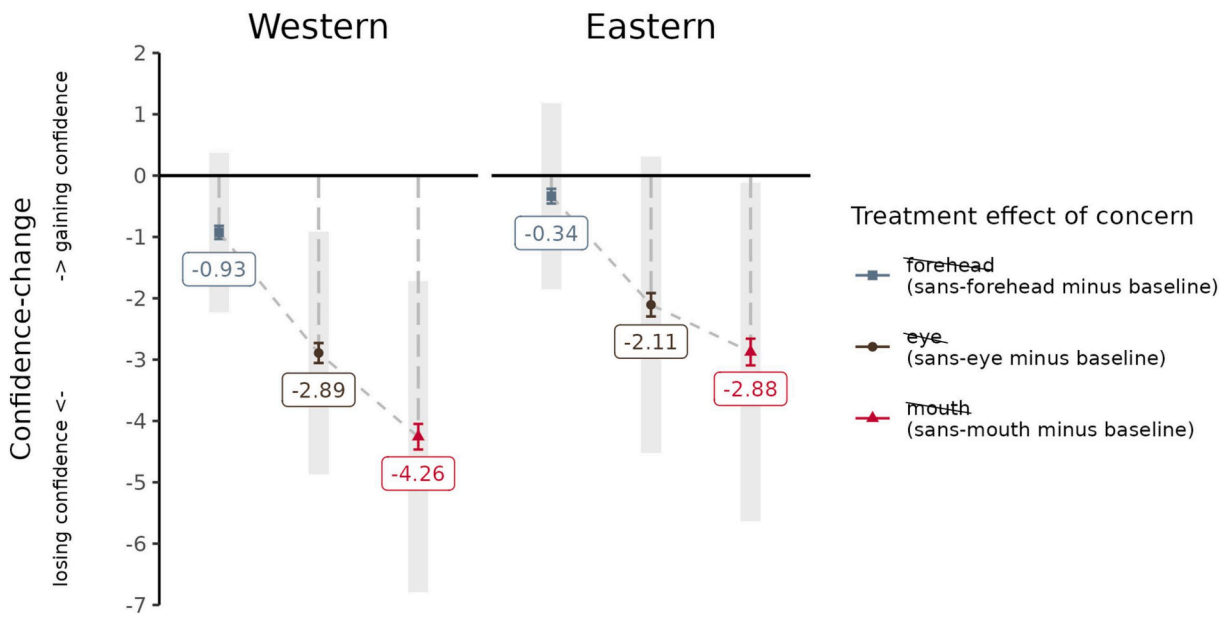

Fig. 3. Summary statistics of `confidence-change` scores specific to different `\FORI` treatments. The two cultural groups are separately graphed. The means are graphically represented by squares/ circles/triangles with their values printed in the adjacent rounded boxes. While the error bars represent the ±1 standard error (SE), the light gray shades behind represent the ±1 standard deviation (SD).

after repeated exposure.

3.1.2. Transparency and openness

We report all data exclusions, all manipulations, and all measures in the studies. Data were analyzed using R, version 4.2.2. The design and hypotheses were preregistered; see https://aspredicted.org/WZV_JXV. The sample size was pre-determined as in the pre-registration. Data collection was not continued after data analysis. All exclusion criteria and deviations from the pre-registered plan are reported in the Supplementary. Survey materials, data, code, and pre-registration documents have been made publicly available at OSF and can be accessed at https://osf.io/8pbz5/?view_only=aac8c24f4b684a03b6862a20b43299d7.

3.1.3. Participants

As in Study 1, participants were recruited from America and China to represent, respectively, Western and Eastern cultures. Specifically, 289 self-identified Americans from MTurk (N_{female} = 147, $Median_{age}$ = 39 years) and 371 self-identified Chinese from SoJump (N_{female} = 170, $Median_{age}$ = 30 years) completed this study in exchange for monetary compensation. The study was administered through an online survey programmed in Qualtrics. Depending on the nationalities of the participants, they either saw the English version or the Chinese version. The sample size provided 90% power to detect an effect size of $\eta_p^2 = 0.019$ or greater for the interaction in a 2(culture) x 3(visibility) between-subject ANOVA test and a 5% false-positive rate.

3.1.4. Stimuli and measures

The survey in the present study consisted of two parts. In Part I, participants *actually* performed the smile-authentication task described in Study 1. Specifically, participants watched 20 short videos (around six seconds each) one at a time. In each video, a different target, with the camera trained on his/her front face, started with a neutral expression and then broke into a smile before returning to the initial neutral state. After watching each video, participants had to make a binary judgment, indicating whether the smile the target displayed was genuine or fake. In half of the videos ($N = 10$, $N_{female} = 5$), the targets faked a smile, while in the remaining half (N = 10, $N_{female} = 2$), the targets smiled genuinely. However, participants were not informed of the relative frequency of genuine smiles versus fake smiles. All 20 stimuli videos were embedded in the Qualtrics survey and displayed in an order that was randomly determined for each individual. Participants could only view the surveys on their own PCs or laptops[1].

These videos have been widely used in previous research (e.g., Bernstein et al., 2008, 2010; Mai et al., 2011; Schindler & Trede, 2021; Young et al., 2015) and were obtained from the BBC Science and Nature website (*BBC - Science & Nature - Human Body and Mind - Spot The Fake Smile,* 2015). The website does not make available information on the age and ethnicity of the individuals filmed in the videos. Thus, we used DeepFace (Serengil & Ozpinar, 2021), an open-source facial attribute analysis framework powered by a state-of-the-art face-recognition neural network to estimate the age and ethnicity from stills extracted from each video. For targets in the genuine smile videos, their estimated age in years ranged from 22 to 38 (M = 29.32 *years*, SD = 4.63 *years*), while those displaying the fake smiles ranged in age from 26 to 39 (M = 29.16 *years*, SD = 4.69 *years*). Fourteen of the 20 targets were identified as white people. Specifically, of the 10 targets expressing a genuine smile, the algorithm identified seven as White and three as Middle Easterners. The composition of the 10 targets expressing a fake smile was comparable according to the algorithm: seven White, one Mid-Easterner, one Asian, and one Black.

Part I employed a between-subject design involving a three-level factor. Specifically, participants in each cultural sample were randomly assigned to watch the videos under one of the three visibility conditions: baseline, sans-eye, or sans-mouth, which matched the visibility conditions of the same names in Study 1. In the baseline condition, these videos were not altered, allowing participants visual access to the entire face of each target. However, in the sans-eye (sans-mouth) condition, a pair of sunglasses (a mask) was digitally superimposed on the targets to conceal their eyes (mouths). The screenshots in Fig. 4 illustrate how the same video would appear in each of the three visibility conditions. For copyright reasons, the faces in these screenshots are blurred.

The accuracy rate of each participant was calculated by dividing the

[1] Please refer to the "Additional information on video stimuli" section in the supplementary material for potential concerns regarding the visual angles of our videos on participants' retinas.

Fig. 4. Screenshots captured at the same time point during the playback of one of the 20 videos in the three visibility conditions (from left to right: the baseline, sans-eye, and sans-mouth conditions). The target is blurred for copyright reasons.

number of correctly judged smiles by 20. An individual who responded randomly would, on average, achieve an accuracy rate of 0.5. Because accuracy rates are essentially probability scores, they are bounded between 0 and 1 and typically exhibit heteroskedasticity, therefore violating the assumptions of statistical methods of the linear regression family, such as the *t*-test and ANOVA. Therefore, we applied a logit function to the accuracy rates, i.e., $log_e\left(\frac{accuracy\ rate}{1-accuracy\ rate}\right)$, to obtain a transformed performance metric called the logarithm of the odds of correct judgment, or, for short, the `log-odds` score.

Two merits of the `log-odds` scores are worth highlighting. First, they preserve the order of accuracy rates because the logit is a strictly increasing function, and therefore higher accurate rates always correspond to higher `log-odds` scores. Second, the `log-odds` scores have a meaningful zero point because a zero `log-odds` score, equivalent to a 0.5 accuracy rate, corresponds to chance-level performance. Therefore, a positive (negative) `log-odds` score would indicate that an individual was more (less) accurate than chance performance.

Upon completing the authentication task in Part I, the survey automatically progressed to Part II where participants were asked to provide three confidence ratings. First, participants reported their confidence levels by rating how well they thought they had performed on a 9-point Likert scale adapted from Study 1. The two endpoints of the scale were, respectively, labeled "1: I was hardly better than random guessing" and "9: I was perfect or almost perfect". Subsequently, participants were briefed on the other two visibility conditions they could have been assigned to but were not (i.e., the counterfactual conditions). After learning the specifics of each counterfactual condition, participants were invited to speculate, on the same 9-point scale, how well they would have done on the task had they been assigned to the counterfactual condition under consideration. Thus, of the three confidence ratings collected from each participant in Part II, one was for the factual condition the person actually experienced and two for the two counterfactual conditions they could have experienced.

As in Study 1, rather than working directly with the raw confidence ratings, we derived two treatment-specific `confidence-change` scores for each participant by subtracting his/her confidence ratings for the baseline condition from the rating for either the sans-eye condition or the sans-mouth condition. Note that the derivation of the `confidence-change` scores was indifferent to whether a confidence rating was for the factual condition or for the counterfactual condition.

Similar to the `confidence-change` scores measured in Study 1, these `confidence-change` scores presumably reflected participants' beliefs of how either concealment treatment (i.e., `\EYE` or `\MOUTH`) would affect their ability to discern smile authenticity. Except that the post-task beliefs measured in this study were supposedly informed by participants' first-hand experiences with authenticating smiles under the visibility condition they were assigned to.

3.2. Results

3.2.1. Part I: task performance

Fig. 5 displays the culture-specific summary statistics of the `log-odds` scores achieved by participants assigned to each visibility condition in Part I of the survey. We report within-culture analyses of how our visibility manipulations affected performance before turning our attention to cross-cultural comparisons.

Western culture. Recall that the diagnostic markers of the authenticity of a smile reside mainly in the eye region. Unsurprisingly, participants in the sans-eye condition ($M = 0.53$, $SD = 0.52$) performed significantly worse than their baseline counterparts ($M = 0.82$, $SD = 0.59$), $t(286) = -2.97$, $p = .003$, *Cohen's d* $= -0.43$, 95%*CI* $[-0.72, -0.14]$. On the contrary, as predicted by the ironic performance-gain hypothesis, participants in the sans-mouth condition ($M = 1.35$, $SD = 0.87$) significantly outperformed their baseline counterparts, $t(286) = 5.35$, $p < .001$, *Cohen's d* $= 0.77$, 95%*CI* $[0.48, 1.06]$. In the Supplementary under the heading "High-power Replication of the Disclosing-by-Masking Effect", we report an ancillary study that successfully replicated the disclosing-by-masking effect with a high-power design ($N = 551$).

Eastern culture. The pattern observed in the American sample was replicated in the Chinese sample. Relative to the baseline condition ($M = 0.82$, $SD = 0.63$), performance was significantly inferior in the sans-eye condition ($M = 0.55$, $SD = 0.49$), $t(368) = -3.29$, $p = .001$, *Cohen's d* $= -0.42$, 95%*CI* $[-0.67, -0.17]$, but significantly superior in the sans-mouth condition ($M = 1.25$, $SD = 0.75$), $t(368) = 5.39$, $p < .001$, *Cohen's d* $= 0.69$, 95%*CI* $[0.43, 0.94]$. Thus, Easterners were not exempted from the disclosing-by-masking effect.

Inter-culture comparisons. Because the strong form of support for the cultural-contingency meta-hypothesis was no longer tenable, we sought the weak form of support by testing whether the disclosing-by-masking effect was *less* pronounced among Easterners than Westerners. A 2 (culture)-by-3 (visibility) ANOVA on *log*-`odds` scores found that only the main effect of visibility was significant, $F(2, 654) = 75.28$, $p < .001$, $\eta_p^2 = .187$. The nonsignificant two-way interaction, $F(2, 654) = 0.48$, $p = .621$, $\eta_p^2 = .001$, suggested that the performance-enhancing effect of the `\MOUTH` treatment was more or less comparable between the East (95% CI of *Cohen's d* $= [0.41, 0.92]$) and the West (95% CI of *Cohen's d* $= [0.51, 1.09]$).

Part I of the present study, together with Study 1, demonstrated the cross-cultural validity of both component hypotheses of our descriptive model, thus making a strong case against the cultural-contingency meta-hypothesis. In the General Discussion section, we address the apparent inconsistency between our findings and Yuki et al. (2007), whose observations purportedly affirm the YMM thesis from which our disconfirmed cultural-contingency meta-hypothesis is derived.

3.2.2. Part II: post-task beliefs

Fig. 6 displays the summary statistics of the two treatment-specific `confidence-change`scores (i.e., `confidence-change\EYE` versus

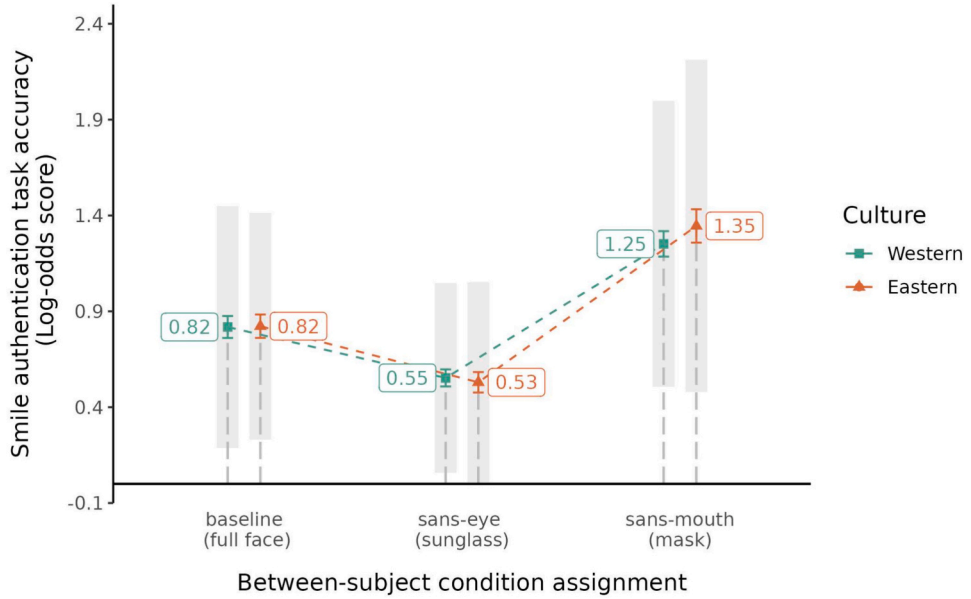

Fig. 5. Condition-wise summary statistics of smile authentication task accuracy (i.e., the `log-odds` scores) for the two cultural groups. The means are represented by either green squares (i.e., the Western sample) or orange triangles (i.e., the Eastern sample) with their numerical values printed in the adjacent rounded boxes. While the error bars represent ±1 standard error (SE), the light-gray shades behind represent ±1 standard deviation (SD). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

`confidence-change\MOUTH`) calculated for individuals of a given culture assigned to each of the three visibility conditions in Part I. These `confidence-change` scores were analyzed within each cultural group separately.

Western culture. A 3 (visibility assignment in Part I)-by-2 (treatment-specificity: `\MOUTH`-specific vs. `\EYE`-specific) mixed ANOVA on the `confidence-change` scores revealed a significant two-way interaction, $F(2, 286) = 57.35$, $p < .001$, $\eta_p^2 = .286$. Dissecting this interaction effect revealed two noteworthy patterns. On the one hand, participants with access to the mouth in Part I (i.e., those assigned to the baseline or sans-eye condition) seemed to gain little insight from their experiences. The pattern of their supposedly experience-informed `confidence-change` scores (i.e., `confidence-change\MOUTH< confidence-change\EYE` < 0) echoes the pattern observed in Study 1 in which task participation was hypothetical. It appeared that the ample opportunities to scrutinize the mouth had little remedial effect on the prior misbeliefs about the diagnostic value of this region. After the task, these participants still considered the `\MOUTH` treatment not only a handicap ($M = -3.26$, $SD = 2.28$, $t(189) = -19.66$, $p < .001$, *Cohen's d* $= -1.43$, 95%*CI* $[-1.63, -1.23]$) but also a more severe handicap than `\EYE` ($M = -1.54$, $SD = 1.74$, $t(189) = -9.07$, $p < .001$, *Cohen's d* $= -0.66$, 95%*CI* $[-0.82, -0.50]$).

On the other hand, participants *without* access to the mouth region in Part I (i.e., those assigned to the sans-mouth condition) appeared to have gained a small dose of insight into the diagnostic value of the mouth from their task experiences. Their `confidence-change` scores formed a pattern (i.e., `confidence-change\EYE < confidence-change\MOUTH`< 0) that was more congruent with the reality that eyes are more critical than mouths for distinguishing between genuine and fake smiles. Although these participants still mistook `\MOUTH` for a handicap ($M = -1.33$, $SD = 1.46$, $t(286) = -6.62$, $p < .002$), they were correct to rank `\EYE` as a worse handicap ($M = -2.48$, $SD = 2.09$, $t(286) = 4.65$, $p < .001$, *Cohen's d* $= 0.60$, 95%*CI* $[0.34, 0.86]$).

Eastern culture. In the Chinese sample, a 3 (visibility assignment in Part I)-by-2 (treatment-specificity) mixed ANOVA on the `confidence-change` scores also detected a significant two-way interaction, $F(2, 368) = 37.53$, $p < .001$, $\eta_p^2 = .169$, which invited the same interpretations as in the American sample. On the one hand, participants with *access* to the mouth region (i.e., the baseline or sans-eye condition) exhibited a pattern they would have exhibited if they had not actually performed the task (i.e., `confidence-change\MOUTH<confidence- change\EYE` < 0), *ts* ≤ -1.98, *ps* $\leq .049$, *Cohen's ds* ≤ -0.20. On the other hand, participants *without* access to the mouth region (i.e., the sans-mouth condition) displayed a pattern suggestive of partial insight acquired through experiencing the task firsthand (i.e., `confidence-change\EYE <confidence-change\MOUTH`< 0), $t(368) = 2.86$, $p = .005$, *Cohen's d* $= 0.29$, 95%*CI* $[0.09, 0.49]$.

We defer an exploration of the nature of the partial learning that seemed to transpire exclusively among the sans-mouth participants in both cultural groups to the General Discussion.

4. Study 3: Belief-Reality Correspondence

4.1. Methods and procedures

4.1.1. Purposes and rationales

The present study combines the previous two studies into a single experiment. Specifically, its procedure begins with a close adaptation of Study 1 followed by an *exact* replication of Study 2. Therefore, it measures people's beliefs about the diagnosticity of the mouth and eyes both before and after they actually perform the smile-authentication task. This design allows us to further probe the interplay between beliefs and actual performance from two new angles. First, we can analyze the correspondence between *pre*-task beliefs and performance to ascertain the degree to which people might have some privileged access to their own judgment prowess with respect to discerning smile authenticity. Perhaps laypeople are not as clueless as our findings so far seem to suggest. Second, we can analyze the correspondence between pre- and post-task beliefs to more precisely evaluate the impact of performing the authentication task on lay beliefs.

4.1.2. Transparency and openness

We report all data exclusions, all manipulations, and all measures in the studies. Data were analyzed using R, version 4.2.2. The design and hypotheses were preregistered; see https://aspredicted.org/7NJ_21H.

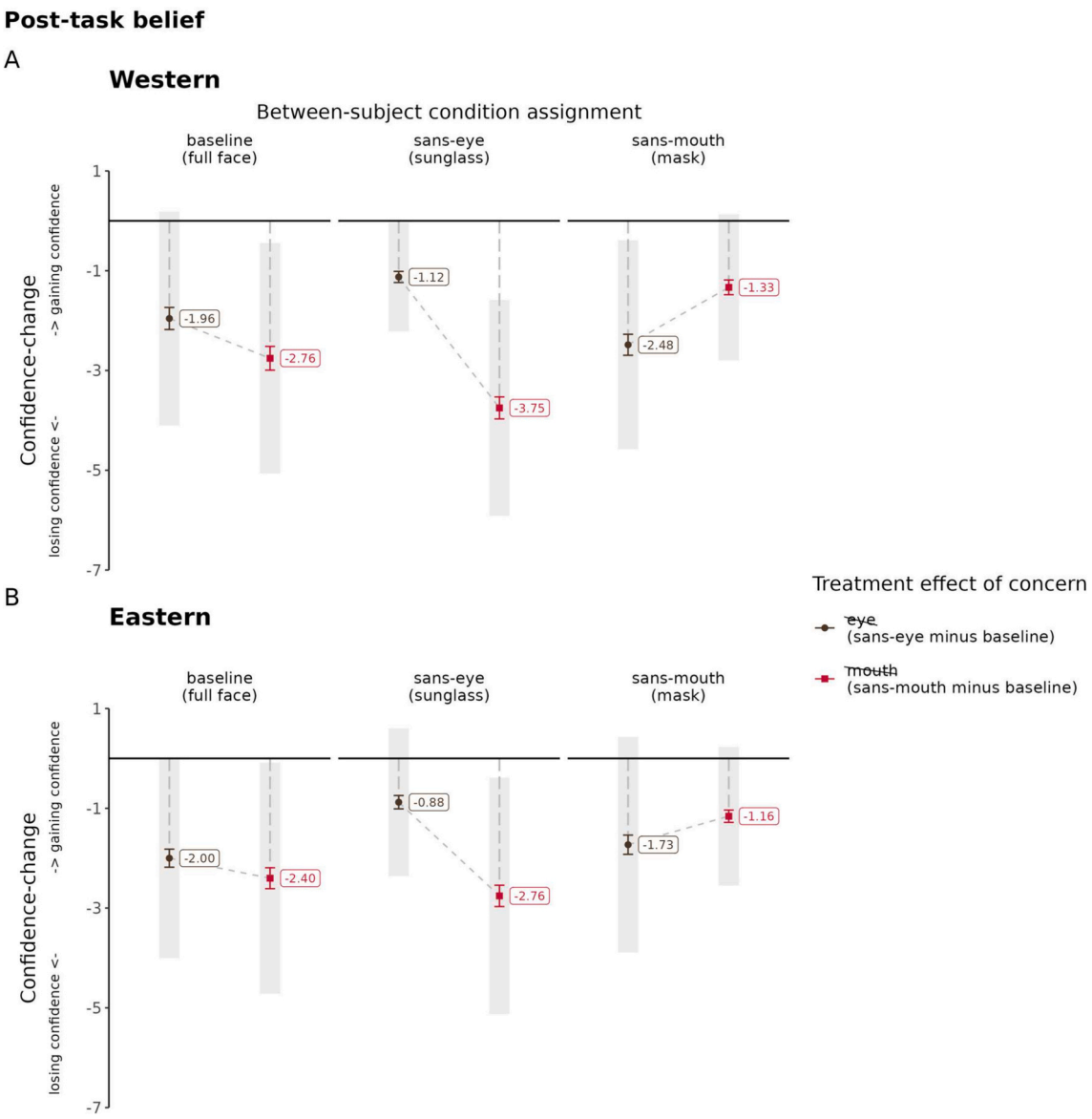

Fig. 6. The summary statistics of `confidence-change` scores specific to different `\FROI` treatments, broken down by visibility condition assignments in Part I. Results for the two cultural samples are separately graphed in Panels (A) and (B). For participants assigned to a given visibility condition in Part I, the mean `confidence-change\EYE` scores and mean `confidence-change\MOUTH` scores are respectively represented by a brown circle and a red square with their values printed in the adjacent rounded boxes. While the error bars represent the ±1 SE, the light-gray shades behind represent ±1 SD. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The sample size was pre-determined as in the pre-registration. Data collection was not continued after data analysis. Exclusion criteria for this study are reported in the *Exclusions* section in the Supplementary. Survey materials, data, code, and pre-registration documents have been made publicly available at OSF and can be accessed at https://osf.io/8pbz5/?view_only=aac8c24f4b684a03b6862a20b43299d7.

4.1.3. Participants

Since no empirical support for cultural contingency–either belief-wise or performance-wise–had been found, an outcome foreshadowed by some recent works from the field of social psychophysics (Jack et al., 2016; Snoek et al., 2023) as well as our linguistic analysis, the culture factor was not incorporated in either this or any of the subsequent studies, all of which sampled exclusively from the West for expediency reasons.

A total of 254 Americans ($N_{\text{female}} = 141$, $Median_{\text{age}} = 37.5$ years) from Prolific (a crowd-sourcing platform similar to Mechanical Turk but purportedly providing higher quality data) completed the study in exchange for monetary compensation. This sample size provided 73% power to detect an effect size of *Cohen's d* = 0.40 (i.e., the smallest effect size observed so far) or greater in an independent *t*-test on performance between two visibility conditions.

4.1.4. Stimuli and measures

The study was administered through an online survey programmed in Qualtrics. The survey consisted of three main parts, which measured, respectively: (I) forecasts of task performance (i.e., pre-task confidence ratings), (II) actual performance on the smile-authentication task, and (II) retrospective appraisals of task performance (i.e., post-task confidence ratings).

Part I of the survey followed the same protocol in Study 1, except for two minor modifications. First, individual participants were asked to forecast their confidence levels for three instead of four visibility conditions: baseline, sans-eye, and sans-mouth. The sans-forehead

(bandanna-wearing) condition was discarded because we did not plan to implement the \FOREHEAD treatment in Part II. Second, the graphic illustrations for the three visibility conditions were no longer based on cartoon drawings as in Fig. 2. Instead, participants were shown three pictures that are the *unblurred* versions of the screenshots in Fig. 4. The new illustration based on actual stimuli was intended to give participants a better idea of what was entailed by different treatments (i.e., \MOUTH versus \EYE).

Parts II and III, together, constituted an exact replication of Study 2. In Part II, participants were randomly assigned to perform the smile-authentication task under one of the three visibility conditions: baseline, sans-mouth, or sans-eye. In Part III, they provided one *factual* confidence rating along with two *counterfactual* ratings in light of their experience with the task. The factual ratings presumably reflected how well they thought they had performed under the visibility condition they were assigned to in Part II, whereas the counterfactual ratings were speculations of how well they would have performed if they had been assigned to either of the two visibility conditions they did not get to experience firsthand.

As in the previous two studies, rather than working directly with the raw confidence ratings, we derived treatment-specific confidence-change scores for each participant by subtracting their confidence ratings for the baseline condition from the rating for either sans-FROI condition. For each participant, two *pairs* of treatment-specific confidence-change scores were derived, respectively, from their three pre-task confidence ratings (measured in Part I) and three post-task confidence ratings (measured in Part III). Although both pairs of confidence-change scores reflected participants' beliefs about the respective impacts of the two focal treatments (i.e., \EYE and \MOUTH) on their ability to discern smile authenticity, the pair based on post-task confidence ratings were supposedly informed by task experiences, and therefore could potentially be more enlightened.

4.2. Results

4.2.1. Part I: pre-task beliefs

Part I of the present study successfully replicated the pattern predicted by the maximum confidence-loss hypothesis. First, the prospect of losing access to the mouth made participants lose confidence in their ability to distinguish genuine smiles from fake ones. The mean confidence-change\MOUTH score ($M = -4.15$, $SD = 2.60$) was significantly lower than zero (indicating a loss of confidence), $t(253) = -25.43$, $p < .001$, *Cohen's d* $= -1.60$, 95%*CI* $[-1.78, -1.41]$. More importantly, the confidence loss specific to \MOUTH significantly exceeded the loss specific to \EYE ($M = -2.42$, $SD = 2.01$), $t = -6.56$, $p < .001$, *Cohen's d* $= -0.68$, 95%*CI* $[-0.90, -0.46]$, although the diagnostic cues actually reside in the eyes instead of the mouth.

4.2.2. Part II: task performance

In Part II, participants' condition-wise accuracy levels (measured by log-odds scores) in the authentication task follow the same pattern observed in Study 2. Specifically, participants assigned to the sans-eye condition ($M = 0.60$, $SD = 0.56$) performed significantly *worse* than their baseline counterparts ($M = 0.88$, $SD = 0.69$), $t(251) = -2.62$, $p = .009$, *Cohen's d* $= -0.40$, 95%*CI* $[-0.71, -0.10]$, affirming the diagnostic value of the eye region. In contrast, participants assigned to the sans-mouth condition ($M = 1.15$, $SD = 0.81$) performed significantly *better* than their baseline counterparts, $t(251) = 2.56$, $p = .011$, *Cohen's d* $= 0.40$, 95%*CI* $[0.09, 0.70]$, thus replicating the disclosing-by-masking effect predicted by the ironic performance-gain hypothesis. The mismatch between reality and pre-task belief, especially with respect to the effect of the \MOUTH treatment, is readily apparent.

4.2.3. Part I vis-a-vis II: predictive validity of pre-task beliefs

As mentioned above, one benefit of combining Studies 1 and 2 is that we could investigate how well people's prior beliefs were predictive of their performance. Such an undertaking would reveal whether, at an individual level, people's prior beliefs possess a certain degree of predictive validity. After all, participants might have had privileged access to their own idiosyncrasies with respect to discerning smile authenticity. In the following, we examine the predictive validity of prior beliefs about the effect of \MOUTH (measured by pre-task confidence-change\MOUTH scores) and prior beliefs about the effect of \EYE (measured by pre-task confidence-change\EYE scores), separately.

Beliefs about the \MOUTH treatment. If there was a kernel of truth to participants' pre-task beliefs about the impact of losing access to the mouth on their authenticity judgments, then the effect of \MOUTH on performance should be moderated by individuals' pre-task confidence-change\MOUTH scores. Specifically, if participants were capable of self-insight in this context, those who felt *less* threatened by the prospect of losing access to the mouth (i.e., those with less extreme confidence-change\MOUTH scores) should be affected *less* by the \MOUTH treatment.

To test this prediction, we retained only participants assigned to either the sans-mouth or the baseline condition in Part II, and then modeled their data with multiple regression to predict their log-odds scores with three terms: a condition assignment indicator (i.e., sans-mouth versus baseline), pre-task confidence-change\MOUTH scores and the two-way interaction between the first two terms. The condition indicator emerged as the only significant predictor, $\beta = 0.27$, $SE = 0.12$, $t = 2.34$, $p = .021$. Neither confidence-change\MOUTH scores ($\beta = 0.04$, $SE = 0.03$, $t = 1.12$, $p = .264$) nor, more importantly, the two-way interaction term ($\beta = -0.07$, $SE = 0.05$, $t = -1.51$, $p = .133$) were related to performance. The *unqualified* positive effect of the condition indicator suggests that participants benefited equally from the \MOUTH treatment regardless of how much they felt threatened by the prospect of losing access to the mouth.

Beliefs about the \EYE treatment. Analogously, if there was a kernel of truth to participants' pre-task beliefs about the impact of losing access to the eyes on their ability to authenticate smiles, then the performance effect of \EYE should be moderated by individuals' pre-task confidence-change\EYE scores. Specifically, if participants were capable of self-insight in this context, those who felt *less* threatened by the prospect of losing access to the eye (i.e., those with less extreme confidence-change\EYE scores) should be affected *less* by the \EYE treatment.

To test this prediction, we conducted an analogous regression analysis on a different subset of participants (i.e., those assigned to the sans-eye or the baseline condition in Part II), predicting their log-odds scores with a condition assignment indicator (i.e., baseline versus sans-mouth), pre-task confidence-change\EYE scores as well as their interaction. Only the coefficient for the condition indicator ($\beta = -0.27$, $SE = 0.10$) was statistically significant, $t = -2.75$, $p = .007$. Neither confidence-change\EYE ($\beta = -0.04$, $SE = 0.04$, $t = -1.03$, $p = .307$) nor, more importantly, the interaction term ($\beta = 0.01$, $SE = 0.05$, $t = 0.23$, $p = .819$) was significant. The *unqualified* negative effect of the condition indicator suggests that participants were disadvantaged to the same extent by the inaccessibility of the eye region, regardless of how much they had felt threatened by the prospect of losing access to the eyes.

In sum, the results of the pair of parallel analyses described above indicate that participants' pre-task beliefs lacked self-insight. For each of the two treatments considered here (i.e., \EYE or \MOUTH), people's prior beliefs about its impact failed to track its actual impact on performance.

4.2.4. Part III: post-task beliefs

Fig. 7(B) displays the summary statistics of the post-task confidence-change\EYE versus confidence-change\MOUTH

scores of participants assigned to each of the three visibility conditions in Part II. The pattern closely resembles that reported in Study 2 (Fig. 6), attesting to the robustness of our results. A 3 (visibility assignment in Part II)-by-2 (treatment-specificity: \MOUTH- vs. \EYE-specific) mixed ANOVA on these post-task confidence-change scores revealed a significant two-way interaction, $F(2, 250) = 25.17$, $p < .001$, $\eta_p^2 = .168$.

On the one hand, participants *with* access to the mouth when performing the task (i.e., those assigned to the baseline or sans-eye condition in Part II) regarded the effect of the \MOUTH treatment as detrimental rather than beneficial (in the baseline group: $M = -3.41$, $SD = 1.94$, 95%$CI_{mean} = [-3.84, -3.00]$, *Cohen's d* $= -1.78$; in the sans-eye group: $M = -4.12$, $SD = 2.27$, 95%$CI_{mean} = [-4.60, -3.63]$, *Cohen's d* $= -1.81$). Moreover, they considered the \MOUTH treatment a worse handicap than the \EYE treatment (in the baseline group: $t(162) = -3.63$, $p < .001$, *Cohen's d* $= -0.57$, 95%$CI = [-0.88, -0.25]$; in the sans-eye group: $t(151) = -8.31$, $p < .001$, *Cohen's d* $= -1.27$, 95%$CI = [-1.59, -0.94]$). In other words, even after repeated exposure to the mouth region, the beliefs held by participants were still aligned with the pattern predicted by the maximum confidence-loss hypothesis (i.e., confidence-change\MOUTH < confidence-change\EYE < 0).

On the other hand, participants *without* access to the mouth during the task (i.e., those assigned to the sans-mouth condition in Part II) appeared to have gained a small dose of insight into the diagnosticity (or lack thereof) of the mouth. Like their counterparts in Study 2, these participants still mistook the \MOUTH treatment for a handicap ($M = -1.65$, $SD = 1.99$, *t*-test with zero: $t(84) = -7.62$, $p < .001$). However, contrary to the maximum confidence-loss hypothesis, they no longer considered \MOUTH a worse handicap than \EYE ($M = -1.86$, $SD = 2.29$, $t(84) = -7.47$, $p < .001$). The difference between confidence-change\MOUTH scores and confidence-change\EYE scores (i.e., -1.65 versus -1.86) was not significant, $t(165) = 0.64$, $p = .522$, *Cohen's d* $= 0.10$, 95%CI $[-0.20, 0.40]$.

It is worth noting that after their vastly different task experiences during Part II, participants in all three conditions uniformly persisted in believing that the \MOUTH treatment was detrimental to discerning smile authenticity. In each condition, the mean post-task confidence-change\MOUTH score was significantly lower than zero ($Ms \leq -1.13$), $ts \leq -5.99$, $ps < .001$.

4.2.5. Part I vis-a-vis III: correspondence between pre- and post-task beliefs

The other merit of combining Studies 1 and 2 is that participants' beliefs were measured both before (Part I) and after (Part III) the authentication task (Part II), allowing us to probe how performing the task under various visibility conditions would affect lay beliefs. In our analyses, beliefs specific to the \MOUTH treatment (measured by confidence-change\MOUTH scores) were examined separately from beliefs specific to the \EYE treatment (measured by confidence-change\EYE scores).

Pre- versus post-task beliefs about the \MOUTH treatment. We modeled data observed on all participants with multiple regression to estimate the extent to which variations in experience (i.e., performing the task under different conditions in Part II) modulated the alignment between pre- and post-task beliefs. Specifically, we regressed post-task confidence-change\MOUTH scores onto mean-centered pre-task confidence-change\MOUTH scores, treatment assignments in Part II, and their two-way interaction.

Among participants assigned to each of the three visibility conditions in Part II, the post-task confidence-change\MOUTH scores were significantly predicted by the pre-task confidence-change\MOUTH scores: $\beta_{baseline} = 0.39$, $SE = 0.09$, $t = 4.38$, $p < .001$; $\beta_{sans-eye} = 0.35$, $SE = 0.07$, $t = 5.03$, $p < .001$; and $\beta_{sans-mouth} = 0.20$, $SE = 0.08$, $t = 2.38$, $p = .018$. On the one hand, the strength of the association between the pre- and post-task scores in the sans-eye condition hardly differed from the baseline condition ($\beta = 0.35$ versus 0.39), $t = -0.30$, $p = .762$. On the other hand, the association was noticeably weaker in the sans-mouth condition than in the baseline condition ($\beta = 0.20$ versus 0.39), although the difference was not significant, $t = -1.37$, $p = .132$.

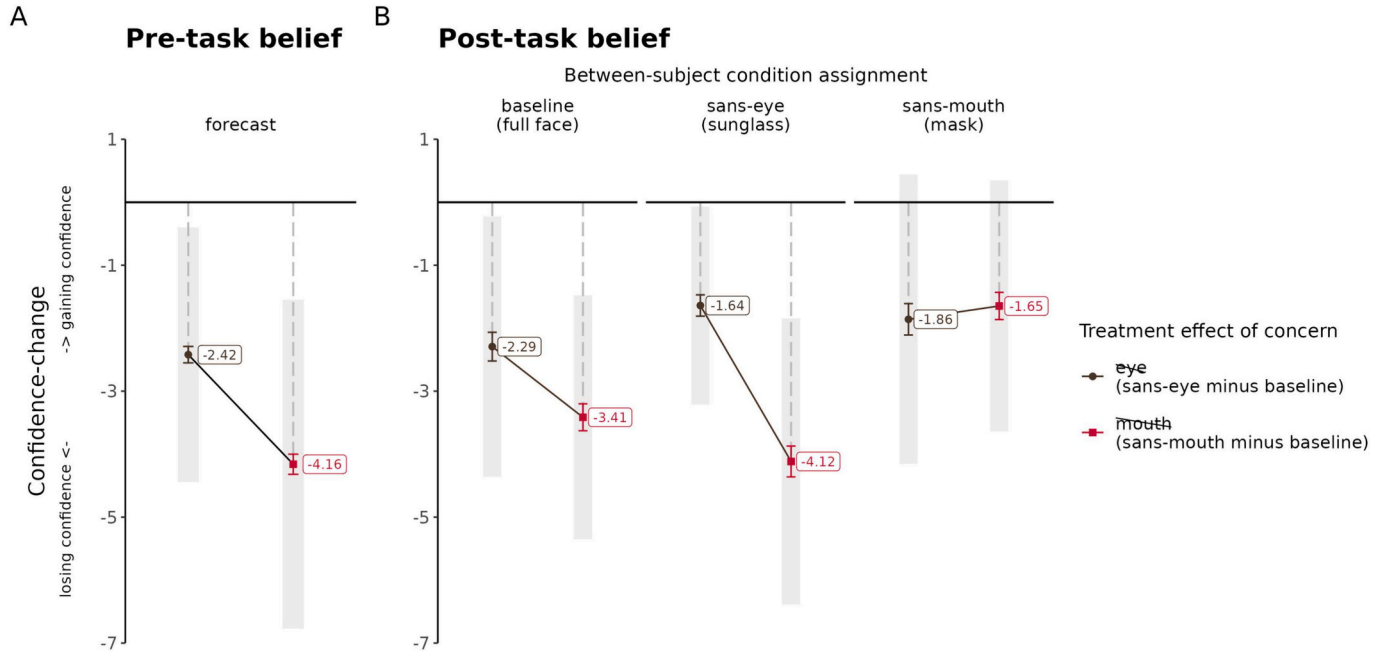

Fig. 7. The summary statistics of both pre-task and post-task confidence-change scores specific to different \FROI treatments. The pre-task statistics, Panel (A), were computed for the entire sample, while the post-task ones, Panel (B), were computed separately for participants assigned to each visibility condition in Part II. The mean confidence-change\EYE score and mean confidence-change\MOUTH score are respectively represented by a brown circle and a red square with their values printed in the adjacent rounded boxes. While the error bars represent the ±1 SE, the light-gray shades behind represent ±1 SD. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Pre- versus post-task beliefs about EYE. A similar multiple regression model was fitted to predict post-task confidence-change\EYE scores with mean-centered pre-task confidence-change\EYE scores, treatment assignments as well as their two-way interaction. Within each of the three experimental groups, the post-task confidence-change\EYE scores were significantly predicted by the pre-task confidence-change\EYE scores: $\beta_{baseline} = 0.60$, $SE = 0.10$, $t = 5.90$, $p < .001$; $\beta_{sans-eye} = 0.25$, $SE = 0.09$, $t = 2.70$, $p = .007$; and $\beta_{sans-mouth} = 0.50$, $SE = 0.10$, $t = 2.38$, $p = .018$. On the one hand, the strength of the association between the pre- and post-task scores in the sans-mouth condition hardly differed from the baseline condition ($\beta = 0.50$ versus 0.60), $t = -0.74$, $p = .462$. On the other hand, the association was noticeably and significantly weaker in the sans-eye condition than in the baseline condition ($\beta = 0.25$ versus 0.60), $t = -2.52$, $p = .012$).

In sum, taking the baseline condition as a reference point, it seems that people's beliefs about the diagnostic value of a given FROI tended to undergo a greater revision when their task experiences had *denied* them access to that FROI than when their task experiences allowed them access *only* to that FROI. Specifically, relative to the baseline condition, the correspondence between pre- and post-task confidence-change\MOUTH (confidence-change\EYE) scores was attenuated considerably after performing the task with *no* access to the mouth (eyes) but was barely affected after performing the task with access only to the mouth (eyes). We return to this interesting pattern in the General Discussion.

5. Study 4: Underlying Mechanism

5.1. Methods and procedures

5.1.1. Purposes and rationales

The present study seeks to elucidate the underlying mechanism of the disclosing-by-masking effect. As mentioned, our account for the effect is that masks shield perceivers from being unduly influenced by the nondiagnostic cues in the mouth region. A testable implication of this account is that the performance benefits conferred by the \MOUTH treatment can be emulated by disabusing perceivers of their faulty beliefs about the mouth's diagnostic value. After all, if perceivers realize how little diagnostic value is offered by the mouth, they likely would tune it out when authenticating smiles, thereby freeing themselves from being misled by the specious cues within this region.

Alternatively, it is not implausible that participants who performed the authentication task under the sans-mouth condition might have felt particularly challenged by being denied access to the ostensibly most critical information source, leading to increased motivation and, therefore, better performance. A testable implication of this competing account is that the performance benefits of the \MOUTH treatment could be emulated by strengthening perceivers' motivation to improve their accuracy.

The strategy we opt for to adjudicate between these two competing accounts is to assess to what extent the performance gain produced by the \MOUTH treatment can be reproduced by disabusing misbelief versus strengthening motivation. To this end, we devise two alternative treatments to compare to the \MOUTH treatment. One alternative, referred to as the *informing* treatment, is intended to dispel misbelief and consists of providing a scientifically valid message on the location of truly diagnostic cues to smile authenticity. The other alternative, referred to as the *incentivizing* treatment, is intended to enhance motivation and consists of promising a monetary reward for outstanding performance. We predict that the *informing* treatment would result in a performance gain comparable to \MOUTH, whereas the *incentivizing* treatment would be ineffective.

5.1.2. Transparency and openness

We report all data exclusions, all manipulations, and all measures in the studies. Data were analyzed using R, version 4.2.2. The design and hypotheses were preregistered; see https://aspredicted.org/XPK_KF1. The sample size was pre-determined as in the pre-registration. Data collection was not continued after data analysis. All exclusion criteria and deviations from the pre-registered plan are reported in the Supplementary. Survey materials, data, code, and pre-registration documents are publicly available at OSF and can be accessed at https://osf.io/8pbz5/?view_only=aac8c24f4b684a03b6862a20b43299d7.

5.1.3. Participants

Two hundred and seventy-nine self-identified Americans from MTurk ($N_{female} = 154$, $Median_{age} = 38$ years) completed this study in exchange for monetary compensation. The study was administered in the form of an online survey programmed in Qualtrics. The sample size provided 90% power to detect an effect size of $\eta_p^2 = 0.049$ or greater in a one-way ANOVA test with four groups and a 5% false-positive rate.

5.1.4. Stimuli and measures

The survey consists of two main parts. Part I administered the same smile-authentication task as in previous studies and individual performance was measured in log-odds scores as before. Participants were randomly assigned to four between-subject conditions: baseline, sans-mouth, plus-information, and plus-incentive. The specifics of the baseline and sans-mouth conditions exactly matched their namesakes in Studies 2 and 3. The remaining two conditions implemented the informing and motivating treatments introduced earlier while simultaneously granting participants full access to the targets' faces as in the baseline condition. In the plus-information condition, participants were instructed, at the start of the task, to read a passage summarizing Ekman's research on markers of genuine versus fake smiles (Ekman et al., 1990; Ekman & Friesen, 1982). The passage is reproduced verbatim below.

Research has shown that:

- In a genuine smile, muscles around the eyes would contract, and this leads to the following changes in appearance: the lower eyelid moves up; crow's feet may appear at the outer corner of the eyes, and the eyebrows move down slightly.
- A fake smile, although featuring *the same* muscle movements *around the mouth* as a genuine one, *barely involves* any changes from muscle contractions *around the eyes*.

In the plus-incentive condition, participants were promised, just before the task, an additional cash bonus for outstanding performance. They were told:

> *Note that you will be able to get an additional $0.60 in bonus based on your accuracy level in the "spot fake smiles" task.* Specifically, if your accuracy ranks among the Top 20% of all MTurk workers who take part in this study, you will be paid an extra $0.60 cents in addition to the $0.60 base payment.

Upon completing the task, the survey automatically progressed to Part II, where all participants first reported their confidence in their performances on a 9-point Likert scale, whose two poles were respectively labeled "hardly better than random guesses" and "perfect or almost perfect." Afterward, participants were asked to speculate, on the same 9-point scale, how well they would have performed if a particular parameter of the task had been different from what was actually the case. The parameter in question was determined by the defining characteristic of each condition and thus varied between the four conditions.

Specifically, sans-mouth participants speculated on the counterfactual in which the mask concealing the mouth had been removed; plus-information participants speculated on the counterfactual in which

they had *not* been exposed to Ekman's research on the facial anatomy of genuine versus fake smiles; plus-incentive participants speculated on the counterfactual in which they had *not* been financially incentivized to do well in the task; and lastly, baseline participants speculated on the counterfactual in which the targets' mouths have been concealed by digitally superimposed masks. Thus, two confidence ratings were obtained from each participant, one for the factual condition they had experienced and the other for a counterfactual condition that was essentially the complement of the factual condition.

For each participant, we calculated a single `confidence-change` score by subtracting their post-task confidence rating for the *control* condition from the rating for the *treatment* condition. For participants in the sans-mouth, plus-information, or plus-incentive condition, the treatment condition denotes the factual conditions they experienced firsthand, and the control condition denotes the counterfactual conditions they simulated mentally. However, for participants in the baseline condition, the designations were reversed, with treatment denoting the counterfactual condition and control denoting the factual condition.

As in earlier studies, the `confidence-change` scores in the sans-mouth and baseline conditions presumably reflected participants' beliefs about the impact of the `\MOUTH` treatment on their ability to discern smile authenticity. Similarly, the `confidence-change` scores in the plus-information (plus-incentive) condition presumably measured participants' beliefs about the impact of the informing (motivating) treatment.

5.2. Results

5.2.1. Part I: task performance

Fig. 8 displays the summary statistics of the `log-odds` scores for each condition. A one-way ANOVA detected a significant main effect, $F(3, 275) = 12.25$, $p < .001$, $\eta_p^2 = 0.118$. Follow-up pairwise comparisons revealed several notable findings.

First, the disclosing-by-masking effect was once again replicated. Participants in the sans-mouth condition ($M = 1.37$, $SD = 0.78$) performed significantly better than their baseline counterparts ($M = 0.83$, $SD = 0.71$), $t(275) = 4.36$, $p < .001$, *Cohen's d* $= 0.73$, 95%*CI* $[0.28, 1.17]$. Second, the informing treatment markedly improved performance. Participants in the plus-information condition ($M = 1.45$, $SD = 0.79$) were significantly more accurate than their baseline counterparts ($M = 0.83$, $SD = 0.71$), $t(275) = 4.84$, $p < .001$, *Cohen'sd* $= 0.83$, 95%*CI* $[0.38, 1.29]$. Third, imparting scientific knowledge benefited authenticity judgment to the same extent as concealing the mouth. Performance in the plus-information condition was statistically indistinguishable from the sans-mouth condition, $t(275) = 0.63$, $p = .531$, *Cohen'sd* $= 0.11$, 95%*CI* $[-0.22, 0.43]$. Last but not least, the motivating treatment barely made a dent. Participants in the plus-incentive condition ($M = 0.92$, $SD = 0.70$) were not better off compared to their baseline counterparts ($M = 0.83$, $SD = 0.71$), $t(275) = 0.67$, $p = .506$, *Cohen's d* $= 0.12$, 95%*CI* $[-0.23, 0.45]$. Together, these results lend strong support to our misbelief-based account of the disclosing-by-masking effect.

5.2.2. Part II: post-task beliefs

Fig. 9 displays the summary statistics of the `confidence-change` scores for participants assigned to each condition in Part I. Analyzing these measures of subjective beliefs revealed two clusters of notable findings. On the one hand, neither the baseline nor the sans-mouth participants were able to intuit the beneficial effect of the `\MOUTH` treatment. In both conditions, the mean `confidence-change` scores were significantly lower than zero, $Ms \leq -1.64$), $ts(275) \leq -8.99$, $ps < .001$, *Cohen's ds* ≤ -1.02. As in the earlier studies, repeated practice failed to rectify the faulty prior belief participants had about the diagnostic value of the mouth region.

On the other hand, participants assigned to either the plus-information or plus-incentive condition appeared to have a well-informed perspective on the efficacy of the treatments they received. The plus-information participants readily acknowledged the benefit of the science-based passage provided to them. The significantly positive mean `confidence-change` score ($M = 1.60$, $SD = 1.69$, $t(275) = 8.28$, $p < .001$, *Cohen's d* $= 1.00$, 95%*CI* $[0.75, 1.26]$) suggests that these participants believed that their accuracy would have been much worse if

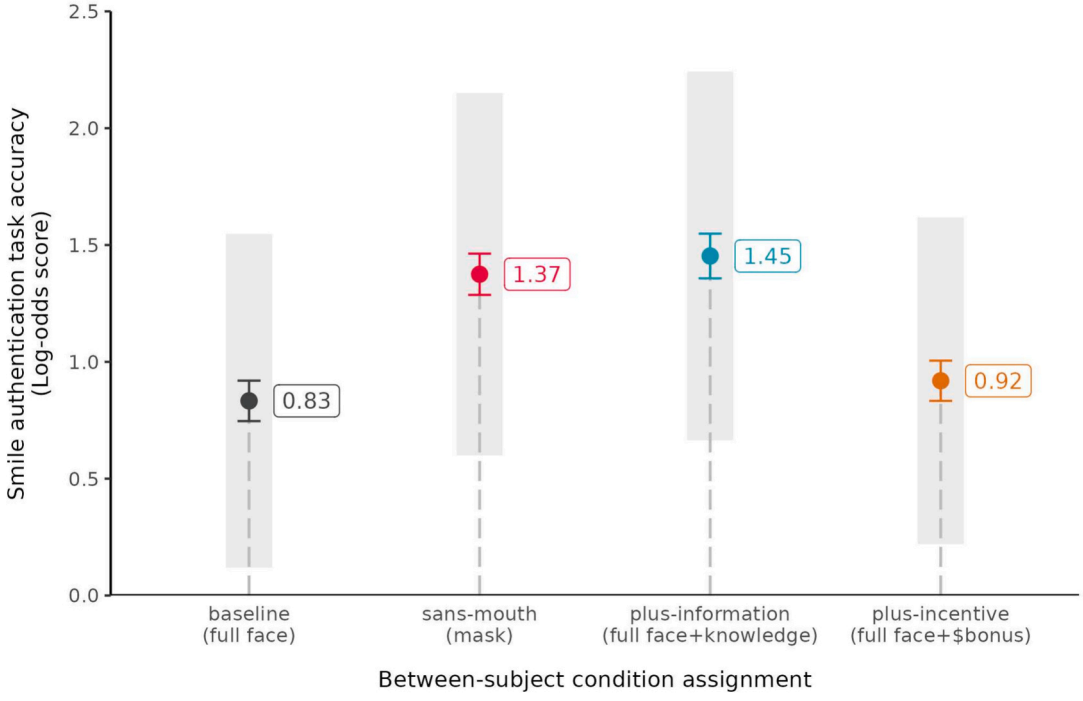

Fig. 8. Summary statistics of the log-odds scores by experimental conditions. The means are visually represented by filled circles with their numerical values displayed in the adjacent rounded boxes. While the error bars represent ±1 SE, the light-gray shades behind represent ±1 SD.

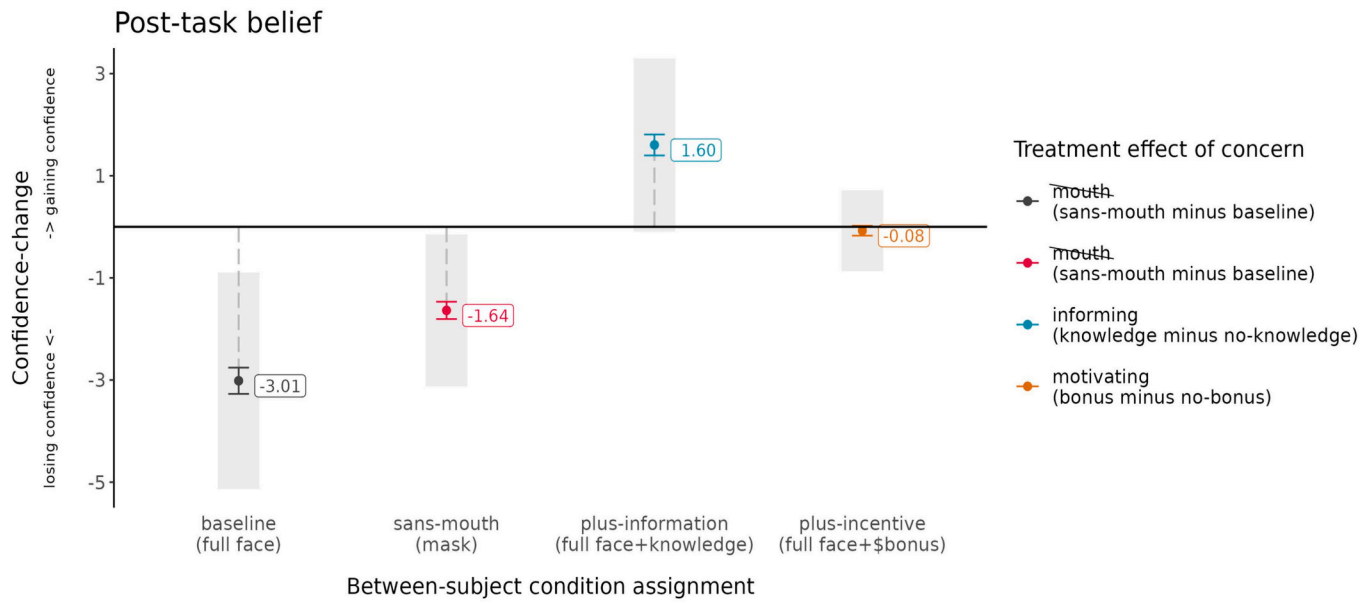

Fig. 9. Summary statistics of `confidence-change` scores by condition assignment. The treatments to which `confidence-change` scores are specific vary across the four conditions. The mean `confidence-change` scores for participants in different conditions are visually represented by filled circles with their numerical values shown in the adjacent rounded boxes. The error bars represent ± 1 SE, and the light-gray shades behind represent ± 1 SD.

they had not learned about Ekman's research and had had to rely on their prior beliefs for guidance. In contrast, the plus-incentive participants resolutely rejected the notion that the financial incentive was of any help. The near-zero mean `confidence-change` score ($M = -0.08$, $SD = 0.79$, $t(275) = -0.39$, $p = .700$, *Cohen's d* $= -0.05$, 95%*CI* $[-0.29, 0.20]$) suggests that these participants dismissed the role of motivation in doing well on this task.

6. Study 5: Covert Authenticity Judgment

6.1. Methods and procedures

6.1.1. Purposes and rationales

Assessing the authenticity of a smile is rarely the be-all and end-all of social perceptions. Often, the authenticity of a smile is among the many concurrent signals that a perceiver would draw on to construct some more complex social inferences to guide their interaction with others. For example, Bernstein et al. (2008) observed that participants implicitly took into account the authenticity of a stranger's smile to infer the risk of entering a collaborative relationship with that stranger. Similarly, Rychlowska et al. (2017) found that computer-generated genuine smiles (corresponding to reward smiles in the authors' framework) were perceived as communicating a stronger signal of social connectedness than computer-generated fake smiles (corresponding to either affiliative or dominance smiles in the authors' framework). The present study investigates the possibility that mask-wearing could be beneficial even in situations where authenticity judgment is not explicitly required but would nevertheless inform important interpersonal judgments.

6.1.2. Transparency and openness

We report all data exclusions, all manipulations, and all measures in the studies. Data were analyzed using R, version 4.2.2. Data collection was not continued after data analysis. Due to a coordination oversight, this study was launched before the preregistration was filed. Exclusion criteria for this study are reported in the *Exclusions* section in the Supplementary. Survey materials, data, code, and relevant documents have been made publicly available at OSF and can be accessed at https://osf.io/8pbz5/?view_only=aac8c24f4b684a03b6862a20b43299d7.

6.1.3. Participants

The present study comprised two phases: the pretest and the main test. Fifty-five ($N_{female} = 25$, $Median_{age} = 36$ years) and 238 ($N_{female} = 102$, $Median_{age} = 40$ years) self-identified Americans from Mturk, respectively, participated in the two phases for monetary compensations. There was no overlap between the two samples. Both the pretest and the main test were administered in the form of an online survey programmed in Qualtrics. The sample size provided 90% power to detect an effect size of $\eta_p^2 = 0.042$ or greater in a 2-by-3 mixed ANOVA and a 5% false-positive rate.

6.1.4. Stimuli and measures

Both the pretest and the main test involved a warmth evaluation task in which participants evaluated the social warmth of twenty strangers. We chose to focus on the perception of social warmth because recent literature suggests that warmth is the dimension people are most concerned with when meeting someone for the first time (Lin et al., 2021).

Pretest. We extracted the first frames of the same 20 BBC smile videos employed in previous studies. These still frames captured the targets in a neutral state. The 20 resulting images of neutral faces were displayed to the participants one at a time in random order. The 20 targets were sorted into two authenticity groups based on whether they were filmed smiling genuinely or feigning a smile: authentic targets versus inauthentic targets.

Participants were allowed to view each image for as long as they wanted before reporting how each of the two adjectives, i.e., *friendly* and *likable*, was characteristic of the target in the image on a 9-point Likert scale: "1: Not at all characteristic" and "9: Very much characteristic." The two ratings for the same target were averaged to form a social warmth rating for that target. Then, for each participant, two `pooled-warmth` scores were obtained by averaging their social warmth ratings for targets in each of the two authenticity groups.

Main test. Participants watched the 20 BBC smile videos one by one. At the end of each video, they rated the target on the same two attributes as their counterparts in the pretest, i.e., likability and friendliness. As in the pretest, the two ratings were averaged to form a social warmth rating for each target. Then, two `pooled-warmth` scores were calculated for each participant by averaging their social warmth ratings for targets in each of the two authenticity groups. Participants were randomly

assigned to watch the videos in one of the three between-subject conditions: full-face, mask-cued, or mask-uncued. Thus, the main test implemented a 2 (target authenticity)-by-3 (experimental manipulation) mixed design.

In the full-face and mask-cued conditions, participants watched, respectively, the original and "mask-wearing" versions of the videos. In both conditions, the instructions explicitly stated the fact that people filmed in these videos were smiling. However, the instructions *never* broached the issue of the genuineness of the targets' smiles. If the disclosing-by-masking effect extended beyond circumstances that require overt authenticity judgment, the perceived warmth of inauthentic targets would trail behind that of authentic targets to a greater extent in the mask-cued condition than in the full-face condition.

The mask-uncued condition was mostly the same as the mask-cued condition except that instructions did not even bring up the targets' facial expressions, further obscuring the relevance of smile authenticity. Thus, the mask-uncued condition constituted a more conservative test of our proposal that the disclosing-by-masking effect would still manifest in situations where discrimination between genuine and fake smiles is not explicitly required, but can nonetheless be useful.

Note that in the full-face condition, smiles would be the most prominent feature of these videos, making the verbal cue about what the targets were expressing redundant. As a result, it was unnecessary to distinguish between full-face-cued and full-face-uncued conditions.

Fig. 10 shows the summary statistics of the `pooled-warmth` scores separately for authentic and inauthentic targets, broken down by participant grouping, i.e., the pretest or one of the three between-subject conditions in the main test.

6.1.5. Phase I: pretest

In the pretest, participants viewed neutral photos of the 20 targets in randomized orders. A paired *t*-test found no statistical difference in mean `pooled-warmth` scores between inauthentic targets ($M = 5.19$, $SD = 0.99$) and authentic ones ($M = 5.25$, $SD = 1.15$), $t(54) = 0.71$, $p = .482$, *Cohen's d* $= 0.10$, 95%*CI* $[-0.17, 0.36]$. Thus, without smiles, the two groups of targets did not seem to systematically differ in static facial features (e.g., morphology and complexion) that might influence perceived warmth.

6.1.6. Phase II: main test

Once smiles were involved, a different pattern emerged. In all three conditions, participants perceived inauthentic targets ($Ms \leq 5.62$, $SD_{full-face} = 1.22$, $SD_{mask-uncued} = 1.15$, $SD_{mask-cued} = 1.24$) as lacking warmth relative to authentic ones ($Ms \geq 6.39$, $SD_{full-face} = 1.19$, $SD_{mask-uncued} = 1.18$, $SD_{mask-cued} = 1.28$), $ts(235) \geq 7.62$, $ps \leq .001$, *Cohen's ds* ≥ 0.639. Thus, even though authenticity judgment was not explicitly asked for, participants spontaneously accounted for the authenticity of the smile when gauging a stranger's interpersonal warmth.

A 2 (target authenticity)-by-3 (experimental condition) mixed ANOVA on the `pooled-warmth` scores revealed a significant two-way interaction, $F(2, 235) = 20.00$, $p < .001$, $\eta_p^2 = .145$. Dissecting this interactive effect revealed two key patterns. First, as predicted, concealing the mouth amplified the discrepancy in perceived warmth between the two authenticity groups. With masks *on* (i.e., the mask-cued or mask-uncued conditions), inauthentic targets lagged even more behind authentic targets in warmth ($M_{\text{inauthentic}} - M_{\text{authentic}} \leq -1.43$, $SE \leq 0.11$) than with masks *off* (i.e., the full-face condition, $M_{\text{inauthentic}} - M_{\text{authentic}} = -0.77$, $SE = 0.10$), $ts \leq -4.67$, $ps \leq .001$, *Cohen's ds* ≤ -0.725. Second, the social penalty (i.e., being judged as lacking warmth) incurred by inauthentic targets was hardly ameliorated when the instructions refrained from mentioning smiles at all. In the mask-uncued condition, the gap in the `pooled-warmth` scores between the two authenticity groups ($M_{\text{inauthentic}} - M_{\text{authentic}} = -1.43$, $SE = 0.10$) was *not* significantly reduced compared to the mask-cued condition ($M_{\text{inauthentic}} - M_{\text{authentic}} = -1.66$, $SE = 0.11$), $t(235) = 1.58$, $p = .115$, *Cohen's d* $= 0.25$, 95%*CI* $[-0.06, 0.57]$.

These findings suggest that concealing mouths would facilitate the discrimination between fake and genuine smiles not only when authenticity judgment is explicitly required, as in the smile-authentication task in preceding studies, but also when the judgment process is triggered automatically or even unconsciously, as in the social-warmth evaluation task in this study.

7. General discussion

Collectively, the results across multiple studies of this research demonstrate a conspicuous misalignment between subjective appraisals and the objective reality of the effect of face mask-wearing on the

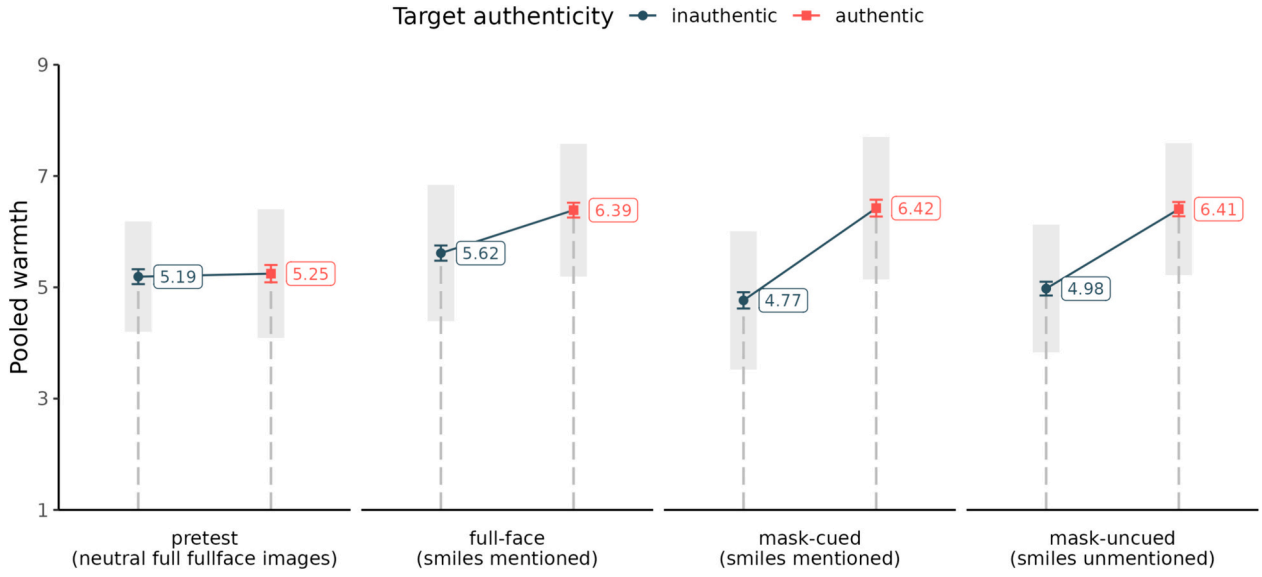

Fig. 10. Summary statistics of the `pooled-warmth` judgments for inauthentic and authentic targets in the pretest as well as in the three experimental conditions of the main test. The means are represented by filled circles/squares with their numerical values printed in the adjacent rounded boxes. The error bars represent ±1 SE, and the light gray shades behind represent ±1 SD.

accurate classification of genuine versus fake smiles. On the *subjective* side, both Westerners and Easterners regarded masks, which block visual access to the normatively *non*diagnostic mouth region, as a severe handicap. As a result, they lost confidence in their ability to discern the authenticity of a smile partially concealed by a mask. The confidence loss inflicted by concealing the mouth eclipsed the loss inflicted by concealing (via sunglasses) the normatively diagnostic eye region (i.e., the maximum confidence-loss hypothesis). On the *objective* side, denying visual access to the mouth improved rather than impaired the accuracy of smile authentication for both Westerners and Easterners (i.e., the ironic performance-gain hypothesis). For reference, denying visual access to the eyes did undermine authentication accuracy, as should be the case. That perceivers were advantaged rather than disadvantaged as a result of losing access to the mouth demonstrated the fallibility of the seeming truism that mask-wearing always interferes with inferring mental states from facial expressions.

7.1. Parallels with deception detection

Both authenticating smiles and identifying lies exemplify interpersonal perception under adversarial conditions. It is not surprising that many parallels can be found between observations reported in this research and notable discoveries documented in the literature on deception detection. First, many cues believed by people, including purported professional lie detectors like police detectives and immigration officers, to be diagnostic of lying (e.g., gaze aversion), are, in fact, unrelated to deception (Colwell et al., 2006; Stromwall & Granhag, 2003). In contrast, many cues that are indicative of lies are dismissed by people as irrelevant (e.g., the absence of nonverbal hand gestures known as illustrators). Moreover, there are cues that covary with deception in one direction (e.g., lying is actually accompanied by less frequent foot movements) but are believed to covary in the opposite direction (e.g., lying is believed to be accompanied by more frequent foot movements). Overall, there is very little correspondence between people's beliefs about cues to deception and reality. According to an earlier study (Zuckerman et al., 1981), lay perceptions of the diagnostic values of various behavioral cues were barely correlated, $r = 0.11$, with the actual diagnostic values of these cues. Last, but not least, there is some preliminary evidence that denying people access to nondiagnostic cues to deception could improve the chance of catching liars. For example, Zuckerman et al. (1981) found that denying people access to the face, which, contrary to what people would like to think, provides little to no diagnostic value, could facilitate their attempts at detecting deception because doing so would force people to focus on the more informative channels such as speech and the body. Interestingly, a later meta-analysis did not replicate this paradoxical phenomenon analogous to our disclosing-by-making effect (Bond & DePaulo, 2006).

There is another aspect of our research that echoes the deception detection literature. When we look at the correlation between participants' `log-odds` scores and their post-task *raw* confidence ratings for the visibility conditions under which they performed the authentication task (i.e., the factual conditions), we find little evidence of confidence tracking judgment accuracy. In the 13 independent samples observed in our research: six samples (= 3 visibility conditions ×2 cultures) in Study 2, three samples (= 3 visibility conditions) in Study 3, and four samples (= 4 experimental conditions) in Study 4, the absolute values of the accuracy-confidence correlation range from 0.02 at the lowest to 0.29 at the highest. Aside from Western participants in the sans-mouth condition of Study 2 ($r = 0.25$, $t(97) = 2.52$, $p = .014$) and participants in the sans-mouth condition of Study 4 ($r = 0.29$, $t(75) = 2.63$, $p = .010$), the confidence-accuracy correlations did not differ significantly from zero in any of the remaining 11 samples ($ps \geq .071$). The lack of self-insight betrayed by these results is in line with the conclusion of a meta-analysis by DePaulo et al. (1997): People's accuracy at judging deception is *un*correlated with their confidence in these judgments. The authors attribute the statistical independence between confidence and accuracy to the empirical fact that people hold erroneous beliefs about cues to deception. This explanation is readily applicable to the analogous accuracy-confidence disconnection noted here.

7.2. The implicit-explicit divergence

In this research, participants' beliefs about the diagnostic utility of eyes versus mouth as cues to smile authenticity were operationalized in terms of `confidence-change` scores, which were derived from participants' self-reported confidence in performing well in the smile-authentication task under different circumstances. Since participants did not explicitly opine about how smile authenticity relates to movements within each of the two FROIs, these `confidence-change` scores can be conceptualized as measuring *implicit* beliefs about cues to smile authenticity. Given the frequently documented dissociation between implicit and explicit beliefs in various domains of social cognition (Gregg et al., 2006; Rydell et al., 2007; Wilson et al., 2000), including deception detection (Hartwig & Granhag, 2015), it is natural to consider whether people's explicit beliefs about the diagnosticity of eyes versus mouths align with or diverge from their implicit beliefs as suggested by the `confidence-change` scores.

To address this question, we conducted an exploratory study using an open-ended approach that allowed participants to freely report whatever visual cues they considered diagnostic of smile authenticity. Specifically, we asked 360 participants recruited from Prolific ($N_{female} = 166$, $Median_{age} = 40$ years) to nominate at least two but no more than three visual cues that they believed would effectively separate genuine and fake smiles. Participants were instructed to provide a short sentence describing each cue they thought of. The key advantage of this open-ended approach over a closed-ended alternative is that participants were not confined to pre-selected options, which could bias the process of mental retrieval. Furthermore, many have argued that open-ended questions constitute a more ecologically valid response format because text rather than numerical ratings are the natural medium through which people communicate their mental states (Hitsuwari et al., 2024; Kjell et al., 2022).

In total, we obtained a total of 948 (= 360 + 360 + 228) text responses across all participants. A coding scheme was implemented to assign the visual cue featured in each response to one of the five mutually exclusive categories that encode the possible locations of the cue in question: F_g, F_m, F_o, B_{-F}, and *NV*. F_g denotes *f*acial regions typically covered by sunglasses (e.g., eyes, brows); F_m denotes *f*acial regions typically covered by a *m*ask (e.g., mouth, chin, lips); F_o denotes *f*acial regions *o*ther than those covered by either sunglasses or masks (e.g., forehead, jowl); B_{-F} denotes parts of the *b*ody after excluding the face (e.g., neck, shoulder); *N* denotes *n*on-codable responses, which either describe some non-visual aspects (e.g., the causal antecedents of genuine vs fake smiles) or are simply non-sequiturs.

All responses were independently coded twice, once manually by one of the authors and once automatically by an open-source zero-shot classifier (Laurer et al., 2023), which is a fine-tuned version of DeBERTa-v3, a popular large-language model (LLM) pretrained by Microsoft. The reason for involving the AI coder was to provide a robustness check on the results of manual coding. After all, the human coder, being a member of the research team, was keenly aware of the purpose of the present research and could, therefore, be biased in their interpretation of the text responses. It is worth noting that the original responses were slightly modified before being fed to the LLM for classification. Specifically, we replaced any instance of the word "smile(s)" with the pronoun "one(s)" in each response to prevent the semantic of "smile(s)" from unduly influencing the model's understanding of each response.

The three histograms making up the top (bottom) triptych in Fig. 11 display the distribution of the responses assigned manually (automatically) to each of the five categories defined by our coding scheme, grouped by whether they were the first, second, or possibly third

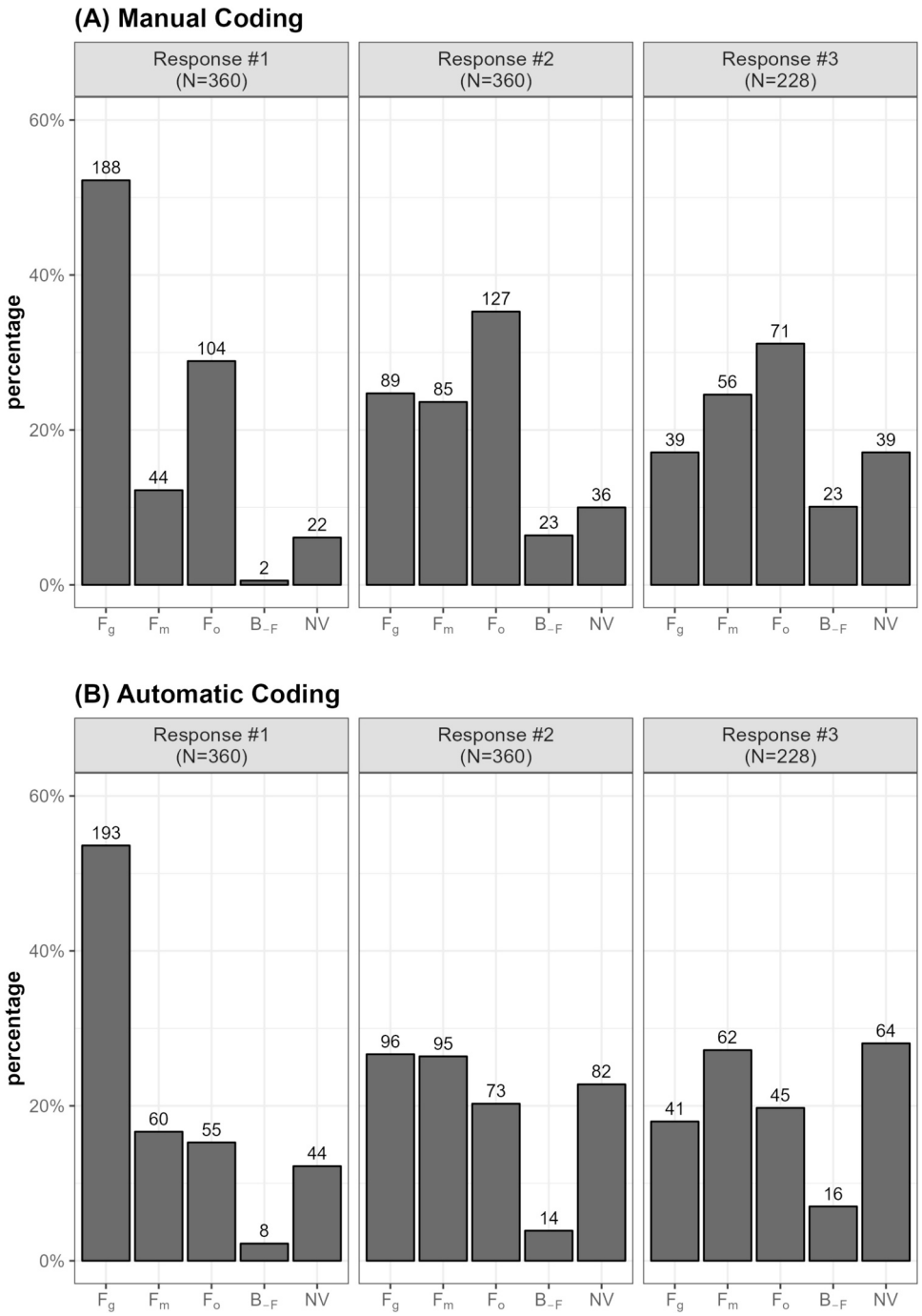

Fig. 11. Distributions of participant responses over the five predefined coding categories, as assigned by a human coder (top triptych) and an AI coder (bottom triptych). The three constituent histograms of the top (bottom) triptych summarize the manual (automatic) categorization of the first, second, and third, optional responses provided by each participant. The slanted number on the top of each bar denotes the raw count as opposed to the percentage of responses assigned to the corresponding category.

responses enumerated by the participants. Recall that not everyone (228 out of 360 did) generated a third response because it was optional. There appears to be a high degree of agreement between manual coding and automatic coding.

The most surprising pattern in Fig. 11 is that participants were more likely to think of the eyes as opposed to the mouth when explicitly instructed to reflect on the visual distinctions between genuine and fake smiles. The chance of some facet of the eye region being the first visual cue that crossed people's minds exceeded 50% (the F_g bar in the first histogram of either triptych in Fig. 11). In contrast, the corresponding chance for some facet of the mouth region was only around 15% (the F_m bar in the first histogram of either triptych in Fig. 11). Furthermore, collapsing across the set of responses provided by a given participant, we found that nearly four-fifths of the participants–78.33% (79.44%) per manual (automatic) coding–nominated some cues related to the eye region, whereas just half of them–41.67% (50.00%) per manual (automatic) coding–nominated some cues related to the mouth region at all. Together, it seems that people's explicit belief about the diagnosticity of the eyes versus the mouth as cues to smile authenticity aligned closely with reality, which was not at all the case with their implicit belief measured by `confidence-change` scores in the main studies.

What could account for this intriguing explicit-implicit divergence?

First, it should be noted that mental availability does not necessarily correspond to perceived importance. Our open-ended survey did not ask participants to evaluate the diagnosticity of the visual cues they nominated. Therefore, we cannot rule out the possibility that although the eyes were more mentally accessible than the mouth, participants nevertheless considered the latter more critical to discriminating between genuine and fake smiles. In fact, past research has shown that it is not uncommon for open-ended questions to fail to capture important beliefs people hold but do not recall at the moment they are reading the question (Marksteiner et al., 2012). In short, despite the ostensible divergence suggested by our exploratory study, people's explicit and implicit beliefs largely agree with each other.

Alternatively, the literature on the dual attitudes model provides a different perspective (Gregg et al., 2006; Rydell et al., 2007; Wilson et al., 2000). This theory contends that implicit and explicit attitudes tend to result from different mental processes and are informationally sequestered. Therefore, it is not uncommon for people to simultaneously hold implicit and explicit attitudes that conflict with each other. Implicit attitudes are typically formed through early experiences and emotional associations, whereas explicit attitudes are typically shaped by rational thoughts and social norms. Consequently, while implicit attitudes are more resistant to change, explicit attitudes tend to be open to influence from recent data or experiences. It is possible that due to the mask mandate enacted during the COVID pandemic, people were forced to adapt to both recognizing and conveying joy or happiness via cues in facial parts unaffected by masks (e.g., smizing or smiling with the eyes). These experiences, in turn, may have led to the insight that the eyes can reliably signal positive affect. Considering how implicit and explicit attitudes are differentially sensitive to new information, it is likely that the impact of this insight is largely limited to the explicit level, therefore potentially explaining the apparent inconsistency between the exploratory study and the main studies. In a sense, this interpretation is consistent with an observation in Study 4 where participants in the plus-information condition readily acknowledged that their performance had benefited from reading the passage on Ekman's research. For the passage to resonate with them to such an extent, these participants probably had already held some belief consistent with the passage's message. Nonetheless, additional empirical research is needed to adjudicate between the two competing accounts considered here.

7.3. The "unexceptional" East

Another notable finding of the present research is the *null* result that Eastern perceivers proved to be *no exception* to either component hypothesis of our descriptive model. Drawing on previous cross-cultural literature, especially the theorization formulated in Yuki et al. (2007), i.e., the YMM thesis, we *tentatively* expected the Eastern experience to defy our descriptive model on two fronts (i.e., the cultural-contingency meta-hypothesis): (1) concealing the mouth would deflate Easterners' confidence to a lesser degree than concealing the eyes, and (2) concealing the mouth would improve performance considerably less, if at all, among Easterners than among Westerners. However, we also expressed reservations about the viability of this meta-hypothesis in light of the innovative work (Jack et al., 2016; Snoek et al., 2023) from the emerging field of social psychophysics (Jack & Schyns, 2017) and our ancillary analysis of semantic associations based on large-scale natural language data, both of which indirectly suggest that Easterners would, like their Western counterpart, treat the mouth as a more informative cue to smile authenticity than the eye.

Although our null findings may initially appear disruptive, the absence of cross-cultural differences within the narrow context of discerning smile authenticity does not necessarily challenge, let alone invalidate, the broader claim of the YMM thesis that Easterners and Westerners differentially rely on the eyes versus the mouth to decode facial expressions of emotions. First, for our null findings to have any bearing on the validity of the YMM thesis, it is imperative that distinguishing between genuine and fake smiles constitutes an instance of emotion perception task. Although it could be argued that smile authentication amounts to recognizing if a target is in a happy/joyful state, thereby representing a special case of emotion perception, it is quite possible that participants approached smile authentication simply as a low-level pattern-matching task without regard for the affective meaning of these patterns. In such a case, our null results would merely suggest that the reference visual patterns invoked by the participants to guide their authenticity judgments and confidence assessment in our studies did not vary qualitatively by culture. Nonetheless, this interpretation of our null results raises questions about the nature and ontogeny of these culturally invariant reference patterns, which can be more thoroughly explored in future research.

Second, even if emotion perception does underlie authenticity judgment, as we have speculated, it is important to emphasize that the cultural universality documented in our research may not extend to other emotion perception tasks. The cultural difference posited by the YMM thesis can very well prove to be the general rule when a wider range of these tasks is taken into account. In other words, our null findings might just be a curious exception to the rule. Indeed, the YMM thesis has received ample corroboration from independent studies involving more prototypical emotion perception tasks (e.g., classifying facial expressions into one of the six basic emotion categories; see Jack et al., 2009; Liu et al., 2022). It then follows that the seeming contradiction between the cultural convergence we observed and the cultural divergence observed by Yuki et al. does not constitute grounds for doubting the credibility of the latter. Future inquiries seeking to explain how and why authenticity judgments deviate from other emotion perception tasks could provide important insight into the interplay between culture and emotions.

7.4. Learning from experiences?

In Studies 2 and 3, we noted a consistent pattern with respect to participants' *post*-task beliefs. It seems that practicing authenticity judgment *without* access to the mouth (i.e., under the sans-mouth condition) constituted a more edifying experience than doing so *with* access to this region (i.e., under the baseline or the sans-eye condition). Although sans-mouth participants still incorrectly regarded the `\MOUTH` treatment as detrimental to discerning smile authenticity, they no longer abode by the maximum confidence-loss hypothesis. Instead, they now considered the `\MOUTH` treatment a *lesser* handicap than `\EYE`, a belief that was more in line with the normative model. In contrast, despite the ample opportunity to scrutinize the mouth, both baseline and sans-eye participants persisted in regarding `\MOUTH` as more detrimental than `\EYE`, a view they would have held even if they had not performed the task.

How can we make sense of the paradox that those who did not see the mouth appeared to attain a better understanding of the diagnostic value of this region than those who did? After all, the sans-mouth condition precluded the possibility of participants noticing the *lack* of systematic variation in mouth movements between genuine and fake smiles. The paradox could be resolved if this apparent learning was proven to have resulted from some *non-learning* process.

A telltale clue can be found in a set of analyses we performed in Study 3 to evaluate the correspondence between pre-task and post-task beliefs about each of the `\FROI` treatments. Compared to the baseline condition, the correspondence between pre- and post-task `confidence-change\MOUTH` scores (in terms of bivariate correlation) was noticeably weakened when the task was performed under the condition that denied participants access to the mouth ($r_{\text{baseline}} = 0.48$ versus $r_{\text{sans-mouth}} = 0.25$, $p = 0.091$), but the correspondence was largely unchanged when the task was performed under the condition that granted access only to the mouth ($r_{\text{baseline}} = 0.48$ versus $r_{\text{sans-eye}} = 0.46$, $p = 0.86$). It is tempting to interpret these results as additional evidence that

participants assigned to the sans-mouth condition gleaned some insight from their experiences, but such an interpretation is doubtful because confidence-change scores specific to the \EYE treatment exhibit an analogous pattern. Compared to the baseline condition, the correspondence between pre- and post-task confidence-change\EYE scores was noticeably weakened when the task was performed under the condition that denied participants access to the eyes ($r_{baseline} = 0.56$ versus $r_{sans\text{-}eye} = 0.33$, $p = 0.060$), but the correspondence barely changed when the task was performed under the condition that granted access only to the mouth ($r_{baseline} = 0.56$ versus $r_{sans\text{-}mouth} = 0.44$, $p = 0.29$). Yet, as mentioned above, participants in the sans-eye condition did not seem to gain any more insight from their task experiences than participants in the baseline condition. It seems unlikely that these two strikingly similar patterns were the result of two distinct cognitive processes, one involving learning and the other not.

We contend that the *apparent* learning that uniquely transpired in the sans-mouth condition was actually an instance of a well-documented attentional phenomenon, i.e., the devaluation-by-inhibition effect (Raymond et al., 2003; Gollwitzer, Martiny-Huenger, & Oettingen, 2014 for reviews). Briefly, people tend to devalue a stimulus if they have deliberately refrained from attending to it earlier. For participants assigned to one of the sans-FROI conditions, they probably were withholding attention from the concealed region during the smile-authentication task, which set into motion the devaluation-by-inhibition process. As a result, they later came to regard the concealed region as having a lower diagnostic value. This speculation was supported by a series of auxiliary analyses detailed in the Supplementary Materials, under the heading "Learning by Not Seeing: Real or Illusory".

The finding that participants seemed incapable of learning from repeated practices at discerning smile authenticity echos what the deception detection literature abundantly documents. Many studies have tested the intuitively appealing notion that people more practiced at catching lies (e.g., law enforcement personnel or managers in organizations) should hold more accurate beliefs about cues to deception than nonprofessionals such as college students (Granhag et al., 2004; Hart et al., 2006; Vrij & Semin, 1996). Yet, they consistently found that professionals endorsed the same types of misbeliefs as laypeople (e.g., liars avert eye contact, see Colwell et al., 2006; Stromwall & Granhag, 2003).

A natural question to ask is why practices had little remedial effect on incorrect beliefs about diagnostic values of different cues to smile authenticity. The literature on concept formation and identification as well as judgment and decision-making, has established that when a task is concerned with distinguishing among two or more categories of stimuli, of which authenticating smiles and deception detection are both prime examples, repeatedly practicing the task rarely enables learning unless immediate and correct feedback is available (Homa & Cultice, 1984; Ashby et al., 1999; Bröker et al., 2022, see Hogarth, 2001 for a general review of feedback in learning). However, this crucial feature was not incorporated in the smile-authentication task deployed in the present research. Indeed, the literature on deception detection has invoked the unreliability of the feedback structure inherent in most lie-deception settings to account for the perpetuation of erroneous beliefs about cues to deception among even experienced professionals (Strömwall et al., 2004; Vrij & Semin, 1996).

7.5. When might mouths matter?

One could question the generalizability of the disclosing-by-masking effect beyond the narrow setting of the smile-authentication task. After all, the task instruction we provided to participants constrained the number of plausible hypotheses about a target's mental state to just two: (1) the target was experiencing happiness and (2) the target was pretending to be happy. However, in more realistic circumstances, a perceived expression must first be categorized as a smile before the question of its authenticity could arise. Recent research showed that mask-wearing increased the chance of a smile being confused for some different expression (Grundmann et al., 2021; Kastendieck et al., 2023). Therefore, in the real world, the mouth could be critical, as it helps perceivers narrow down the alternative hypotheses they need to consider when decoding an apparent smile. Indeed, Blais et al. (2012) present evidence that visual information within the mouth region is most critical to discriminating between facial expressions that encoded the six basic emotional states.

At first glance, Study 5 in our research seems to be at odds with the supremacy of the mouth demonstrated by Blais et al. (2012). Recall that in the mask-uncued condition of Study 5, participants were *not* told that the mask-wearing targets in the videos were smiling and, therefore, presumably faced a larger platter of plausible hypotheses about the targets' mental states than their counterparts in the mask-cued condition, who were told such. Yet, when judging the social warmth of the targets, uncued participants were just as sensitive to the authenticity of the targets' smiles as their cued counterparts. Therefore, even when perceivers were *not* certain they were perceiving a smile, the mouth region did not appear to be missed. A plausible account for reconciling the apparent expendability of the mouth suggested by Study 5 with existing evidence for the supremacy of this region (2012) is that our participants were able to infer movements within the concealed mouth region from muscle contractions in the exposed neighboring areas. After all, when the muscles in the mouth region contract to form a smile, be it fake or genuine, the muscles in the neighboring areas must move in a complementary manner to accommodate the former.

8. Conclusion

Human beings' uncanny ability to notice the analogy between a novel event and past experiences based on incomplete information and extrapolate what is true of the latter to the former, though usually adaptive, can lead to grave errors when the resemblance between the present and past is only superficial. Current research provides a topical case study of the perils of this type of hasty overgeneralization. Although the mouth region is often critical to the basic-level categorization of a facial expression (e.g., is the person smiling, frowning, or pouting, see Blais et al., 2012; Smith et al., 2005), it is mostly superfluous for distinguishing between genuine and fake smiles. However, people, regardless of their cultural background, erroneously assumed otherwise. As a result, they mistook mask-wearing for a serious threat to making veridical authenticity judgments, while unknowingly benefiting from having the mouth region concealed by masks.

Author note

This work was supported by ShanghaiTech Faculty Research Grant and National Science Foundation S-STEM award (DUE-1929882). Materials, datasets, codes, and pre-registration documents can be accessed at https://osf.io/8pbz5/?view_only=aac8c24f4b684a03b6862a20b43299d7. We would like to thank the following research assistants for their help on this project: Annette Kwasniewski, Carla Santacruz, Elizabeth Schulteis, Itzel Rayo, Jason Gill, Kiwaun Wallace, Ming Ding, Yuliya Zanevych, and Yixiao Shen.

Open practices

Materials, data, codes, and pre-registration documents for the experiments are available at https://osf.io/8pbz5/?view_only=aac8c24f4b684a03b6862a20b43299d7.

CRediT authorship contribution statement

Haotian Zhou: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project

administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Meiying Wang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Formal analysis, Data curation. **Yu Yang:** Validation, Supervision, Project administration, Funding acquisition. **Elizabeth A. Majka:** Writing – review & editing, Validation, Resources, Data curation, 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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2024.104658.

References

Ashby, F. G., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics, 61*(6), 1178–1199. https://doi.org/10.3758/BF03207622

BBC - Science *& Nature - Human Body and Mind - Spot The Fake Smile*. http://www-cs-faculty.stanford.edu/~uno/abcde.html, (2015, February).

Bernstein, M. J., Sacco, D. F., Brown, C. M., Young, S. G., & Claypool, H. M. (2010). A preference for genuine smiles following social exclusion. *Journal of Experimental Social Psychology, 46*(1), 196–199. https://doi.org/10.1016/j.jesp.2009.08.010

Bernstein, M. J., Young, S. G., Brown, C. M., Sacco, D. F., & Claypool, H. M. (2008). Adaptive responses to social exclusion: Social rejection improves detection of real and fake smiles. *Psychological Science, 19*(10), 981–983. https://doi.org/10.1111/j.1467-9280.2008.02187.x

Blais, C., Roy, C., Fiset, D., Arguin, M., & Gosselin, F. (2012). The eyes are not the window to basic emotions. *Neuropsychologia, 50*(12), 2830–2838. https://doi.org/10.1016/j.neuropsychologia.2012.08.010

Bogaard, G., & Meijer, E. H. (2018). Self-reported beliefs about verbal cues correlate with deception-detection performance. *Applied Cognitive Psychology, 32*(1), 129–137. https://doi.org/10.1002/acp.3378

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

Bröker, F., Love, B. C., & Dayan, P. (2022). When unsupervised training benefits category learning. *Cognition, 221*, Article 104984. https://doi.org/10.1016/j.cognition.2021.104984

Calvo, M. G., Gutiérrez-García, A., Fernández-Martín, A., & Nummenmaa, L. (2014). Recognition of facial expressions of emotion is related to their frequency in everyday life. *Journal of Nonverbal Behavior, 38*(4), 549–567. https://doi.org/10.1007/s10919-014-0191-3

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy, 97*(5), 1232–1254. https://doi.org/10.1086/261651

Campagne, D. (2021). The problem with communication stress from face masks. *Journal of Affective Disorders Reports., 3*, Article 100069. https://doi.org/10.1016/j.jadr.2020.100069

Cimpian, A., Brandone, A. C., & Gelman, S. A. (2010). Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science, 34*(8), 1452–1482. https://doi.org/10.1111/j.1551-6709.2010.01126.x

Colwell, L. H., Miller, H. A., Miller, R. S., Lyons, J., & Phillip, M. (2006). US police officers' knowledge regarding behaviors indicative of deception: Implications for eradicating erroneous beliefs through training. *Psychology, Crime & Law, 12*(5), 489–503. https://doi.org/10.1080/10683160500254839

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. https://doi.org/10.1207/s15327957pspr0104_5

Duchenne, G.-B., & de Boulogne, G.-B. D. (1990). *The mechanism of human facial expression*. Cambridge University Press.

Ekman, P., Davidson, R. J., & Friesen, W. V. (1990). The duchenne smile: Emotional expression and brain physiology: II. *Journal of Personality and Social Psychology, 58*(2), 342.

Ekman, P., & Friesen, W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. *Consulting Psychologists, 3*.

Ekman, P., & Friesen, W. V. (1982). Felt, false, and miserable smiles. *Journal of Nonverbal Behavior, 6*(4), 238–252.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *The facial action coding system: A technique for the measurement of facial movement*. San Francisco, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & O'Sullivan, M. (1988). Smiles when lying. *Journal of Personality and Social Psychology, 54*(3), 414–420. https://doi.org/10.1037/0022-3514.54.3.414

Ekman, P., & Oster, H. (1979). Facial expressions of emotion. *Annual Review of Psychology, 30*(1), 527–554. https://doi.org/10.1146/annurev.ps.30.020179.002523

Ettenson, R., Shanteau, J., & Krogstad, J. (1987). Expert judgment: Is more information better? *Psychological Reports, 60*(1), 227–238. https://doi.org/10.2466/pr0.1987.60.1.227

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1*, 288–299. https://doi.org/10.1037/0096-1523.1.3.288

Ford, B. Q., & Mauss, I. B. (2015). Culture and emotion regulation. *Current Opinion in Psychology, 3*, 1–5. https://doi.org/10.1016/j.copsyc.2014.12.004

Frank, M. G., & Ekman, P. (1993). Not all smiles are created equal: The differences between enjoyment and nonenjoyment smiles. *6*(1), 9–26. https://doi.org/10.1515/humr.1993.6.1.9

Frank, M. G., Ekman, P., & Friesen, W. V. (1993). Behavioral markers and recognizability of the smile of enjoyment. *Journal of Personality and Social Psychology, 64*(1), 83. https://doi.org/10.1037/0022-3514.64.1.83

Gigerenzer, G. (2022). *How to stay smart in a smart world: Why human intelligence still beats algorithms*. MIT Press.

Gill, D., Garrod, O., Jack, R., & Schyns, P. (2013). Beyond facial morphology: Social impressions from dynamic face gestures. *Journal of Vision, 13*(9), 1270. https://doi.org/10.1167/13.9.1270

Gollwitzer, P. M., Martiny-Huenger, T., & Oettingen, G. (2014). Chapter two - affective consequences of intentional action control. In A. J. Elliot (Ed.), *Vol. 1. Advances in motivation Science* (pp. 49–83). Elsevier. https://doi.org/10.1016/bs.adms.2014.08.002.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour, 6*(7), 975–987. https://doi.org/10.1038/s41562-022-01316-8

Granhag, P. A., Andersson, L. O., Strömwall, L. A., & Hartwig, M. (2004). Imprisoned knowledge: Criminals' beliefs about deception. *Legal and Criminological Psychology, 9*(1), 103–119. https://doi.org/10.1348/135532504322776889

Gregg, A., Seibt, B., & Banaji, M. (2006). Easier done than undone: Asymmetry in the malleability of implicit preferences. *Journal of Personality and Social Psychology, 90*(1), 1–20. https://doi.org/10.1037/0022-3514.90.1.1

Grundmann, F., Epstude, K., & Scheibe, S. (2021). Face masks reduce emotion-recognition accuracy and perceived closeness. *PLoS One, 16*(4), Article e0249792. https://doi.org/10.1371/journal.pone.0249792

Gunnery, S. D., & Ruben, M. A. (2016). Perceptions of Duchenne and non-Duchenne smiles: A meta-analysis. *Cognition and Emotion, 30*(3), 501–515. https://doi.org/10.1080/02699931.2015.1018817

Hall, C. C., Ariss, L., & Todorov, A. (2007). The illusion of knowledge: When more information reduces accuracy and increases confidence. *Organizational Behavior and Human Decision Processes, 103*(2), 277–290. https://doi.org/10.1016/j.obhdp.2007.01.003

Hart, C., Hudson, L., Fillmore, D., & Griffith, J. (2006). Managerial beliefs about the behavioral cues of deception. *Individual Differences Research, 4*, 176–184.

Hartwig, M., & Granhag, P. A. (2015). Exploring the nature and origin of beliefs about deception: Implicit and explicit knowledge among lay people and presumed experts. In *Detecting deception: Current challenges and cognitive approaches* (pp. 125–154). John Wiley & Sons.

Hitsuwari, J., Okano, H., & Nomura, M. (2024). Predicting attitudes toward ambiguity using natural language processing on free descriptions for open-ended question measurements. *Scientific Reports, 14*(1), 8276. https://doi.org/10.1038/s41598-024-59118-z

Hogarth, R. M. (2001). *Educating intuition*. University of Chicago Press.

Homa, D., & Cultice, J. C. (1984). Role of feedback, category size, and stimulus distortion on the acquisition and utilization of ill-defined categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 83–94. https://doi.org/10.1037/0278-7393.10.1.83

Jack, R. E., Blais, C., Scheepers, C., Schyns, P. G., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current Biology, 19*(18), 1543–1548. https://doi.org/10.1016/j.cub.2009.07.051

Jack, R. E., & Schyns, P. G. (2015). The human face as a dynamic tool for social communication. *Current Biology, 25*(14), R621–R634. https://doi.org/10.1016/j.cub.2015.05.052

Jack, R. E., & Schyns, P. G. (2017). Toward a social psychophysics of face communication. *Annual Review of Psychology, 68*, 269–297. https://doi.org/10.1146/annurev-psych-010416-044242

Jack, R. E., Sun, W., Delis, I., Garrod, O. G. B., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General, 145*(6), 708–730. https://doi.org/10.1037/xge0000162

Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction, 20*(2), 146–154.

Johnston, L., Miles, L., & Macrae, C. N. (2010). Why are you smiling at me? Social functions of enjoyment and non-enjoyment smiles. *British Journal of Social Psychology, 49*(1), 107–127. https://doi.org/10.1348/014466609X412476

Kastendieck, T., Dippel, N., Asbrand, J., & Hess, U. (2023). Influence of child and adult faces with face masks on emotion perception and facial mimicry. *Scientific Reports, 13*(1), 14848. https://doi.org/10.1038/s41598-023-40007-w

Kjell, O., Sikström, S., Kjell, K., & Schwartz, H. (2022). Natural language analyzed with AI-based transformers predict traditional subjective well-being measures approaching the theoretical upper limits in accuracy. *Scientific Reports, 12*(1), 3918. https://doi.org/10.1038/s41598-022-07520-w

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97–109. https://doi.org/10.1037/a0020762

Laurer, M., Atteveldt, W., Casas, A., & Welbers, K. (2023). *Building efficient universal classifiers with natural language inference. arXiv:2312.17543.* https://doi.org/10.48550/arXiv.2312.17543

Leslie, S.-J., Khemlani, S., & Glucksberg, S. (2011). Do all ducks lay eggs? The generic overgeneralization effect. *Journal of Memory and Language, 65*(1), 15–31. https://doi.org/10.1016/j.jml.2010.12.005

Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language And Social Psychology., 33*(4), 378–392. https://doi.org/10.1177/0261927X14535916

Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of English trait words. *Nature Communications, 12*(1, Suppl). https://doi.org/10.1038/s41467-021-25500-y. S2.

Liu, M., Duan, Y., Ince, R. A. A., Chen, C., Garrod, O. G. B., Schyns, P. G., & Jack, R. E. (2022). Facial expressions elicit multiplexed perceptions of emotion categories and dimensions. *Current Biology, 32*(1), 200–209.e6. https://doi.org/10.1016/j.cub.2021.10.035

Mai, X., Ge, Y., Tao, L., Tang, H., Liu, C., & Luo, Y.-J. (2011). Eyes are windows to the chinese soul: Evidence from the detection of real and fake smiles. *PLoS One, 6*(5), Article e19903. https://doi.org/10.1371/journal.pone.0019903

Marksteiner, T., Reinhard, M. A., Dickhäuser, O., & Sporer, S. L. (2012). How do teachers perceive cheating students? Beliefs about cues to deception and detection accuracy in the educational field. *European journal of psychology of education, 27*, 329–350.

Martin, J., Rychlowska, M., Wood, A., & Niedenthal, P. (2017). Smiles as multipurpose social signals. *Trends in Cognitive Sciences, 21*(11), 864–877. https://doi.org/10.1016/j.tics.2017.08.007

Matsumoto, D., & Lee, M. (1993). Consciousness, volition, and the neuropsychology of facial expressions of emotion. *Consciousness and Cognition, 2*(3), 237–254. https://doi.org/10.1006/ccog.1993.1022

Matuschek, C., Moll, F., Fangerau, H., Fischer, J., Zänker, K., Griensven, M., … Haussmann, J. (2020). Face masks: Benefits and risks during the COVID-19 crisis. *European Journal of Medical Research, 25*, 32. https://doi.org/10.1186/s40001-020-00430-5

Miles, L., & Johnston, L. (2007). Detecting happiness: Perceiver sensitivity to enjoyment and non-enjoyment smiles. *Journal of Nonverbal Behavior, 31*(4), 259–275. https://doi.org/10.1007/s10919-007-0036-4

Mui, P., Gan, Y., Goudbeek, M., & Swerts, M. (2020). Contextualising smiles: Is perception of smile genuineness influenced by situation and culture? *Perception, 49*(3), 357–366. https://doi.org/10.1177/0301006620904510

Niedenthal, P. M., Mermillod, M., Maringer, M., & Hess, U. (2010). The Simulation of Smiles (SIMS) model: Embodied simulation and the meaning of facial expression. *Behavioral and Brain Sciences, 33*(6), 417–433. https://doi.org/10.1017/S0140525X10000865

Nisbett, R. E., Zukier, H., & Lemley, R. E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology, 13*(2), 248–277. https://doi.org/10.1016/0010-0285(81)90010-4

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics https://doi.org/10.3115/v1/D14-1162.

Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology, 33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907

Ramdani, C., Ogier, M., & Coutrot, A. (2022). Communicating and reading emotion with masked faces in the Covid era: A short review of the literature. *Psychiatry Research, 316*, Article 114755. https://doi.org/10.1016/j.psychres.2022.114755

Raymond, J. E., Fenske, M. J., & Tavassoli, N. T. (2003). Selective attention determines emotional responses to novel visual stimuli. *Psychological Science, 14*(6), 537–542. https://doi.org/10.1046/j.0956-7976.2003.psci_1462.x

RealEye. (2024). Facial Action Coding System—FACS. https://www.realeye.io/features/online-webcam-facial-coding/facial-action-coding-system-facs.

Rychlowska, M., Jack, R. E., Garrod, O. G. B., Schyns, P. G., Martin, J. D., & Niedenthal, P. M. (2017). Functional smiles: Tools for love, sympathy, and war. *Psychological Science, 28*(9), 1259–1270. Retrieved December 27, 2023, from https://eprints.gla.ac.uk/142573/.

Rydell, R., McConnell, A., Strain, L., Claypool, H., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology, 37*(5), 867–878. https://doi.org/10.1002/ejsp.393

Schindler, S., & Trede, M. (2021). Does social exclusion improve detection of real and fake smiles? A replication study. *Frontiers in Psychology, 12*.

Sendic, G. (2022, July). Facial muscles. https://www.kenhub.com/en/library/anatomy/the-facial-muscles.

Serengil, S. I., & Ozpinar, A. (2021). Hyperextended lightface: A facial attribute analysis framework. *International Conference on Engineering and Emerging Technologies (ICEET), 1–4*. https://doi.org/10.1109/ICEET53442.2021.9659697

Sheldon, K. M., Corcoran, M., & Sheldon, M. (2021a). Duchenne smiles as honest signals of chronic positive mood. *Perspectives on Psychological Science, 16*(3), 654–666. https://doi.org/10.1177/1745691620959831

Sheldon, K. M., Goffredi, R., & Corcoran, M. (2021b). The glow still shows: Effects of facial masking on perceptions of duchenne versus social smiles. *Perception, 50*(8), 720–727. https://doi.org/10.1177/03010066211027052

Smith, M. L., Cottrell, G. W., Gosselin, F., & Schyns, P. G. (2005). Transmitting and decoding facial expressions. *Psychological Science, 16*(3), 184–189. https://doi.org/10.1111/j.0956-7976.2005.00801.x

Snoek, L., Jack, R. E., Schyns, P. G., Garrod, O. G. B., Mittenbühler, M., Chen, C., … Scholte, H. S. (2023). Testing, explaining, and exploring models of facial expressions of emotions. *Science Advances, 9*(6), eabq8421. https://doi.org/10.1126/sciadv.abq8421

Stromwall, L., & Granhag, P. A. (2003). How to detect deception? Arresting the beliefs of police officers, prosecutors and judges. *Psychology, Crime & Law, 9*(1), 19–36. Retrieved January 8, 2024, from https://heinonline.org/HOL/P?h=hein.journals/pcyceadl9&i=19.

Strömwall, L. A., Granhag, P. A., & Hartwig, M. (2004). Practitioners' beliefs about deception. In L. A. Strömwall, & P. A. Granhag (Eds.), *The detection of deception in forensic contexts* (pp. 229–250). Cambridge University Press. https://doi.org/10.1017/CBO9780511490071.010.

Sutherland, S. L., Cimpian, A., Leslie, S.-J., & Gelman, S. A. (2015). Memory errors reveal a bias to spontaneously generalize to categories. *Cognitive Science, 39*(5), 1021–1046. https://doi.org/10.1111/cogs.12189

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology, 66*(1), 519–545. https://doi.org/10.1146/annurev-psych-113011-143831

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vrij, A., & Semin, G. R. (1996). Lie experts' beliefs about nonverbal indicators of deception. *Journal of Nonverbal Behavior, 20*(1), 65–80. https://doi.org/10.1007/BF02248715

Williams, J. J., Lombrozo, T., & Rehder, B. (2013). The hazards of explanation: Overgeneralization in the face of exceptions. *Journal of Experimental Psychology: General, 142*(4), 1006–1014. https://doi.org/10.1037/a0030996

Wilson, T., Lindsey, S., & Schooler, T. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101–126. https://doi.org/10.1037/0033-295X.107.1.101

Young, S. G., Slepian, M. L., & Sacco, D. F. (2015). Sensitivity to perceived facial trustworthiness is increased by activating self-protection motives. *Social Psychological and Personality Science, 6*(6), 607–613. https://doi.org/10.1177/1948550615573329

Yuki, M., Maddux, W. W., & Masuda, T. (2007). Are the windows to the soul the same in the east and west? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. *Journal of Experimental Social Psychology, 43*(2), 303–311. https://doi.org/10.1016/j.jesp.2006.02.004

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. In L. Berkowitz (Ed.), *Vol. 14. Advances in experimental social psychology* (pp. 1–59). Academic Press. https://doi.org/10.1016/S0065-2601(08)60369-X.