Face Recognition Patterns in Cross-Race Effect

Why They All Look Alike?!

Tarana Sharma
Grade 12 | Westmount Charter School

Abstract

Face recognition is a natural form of biometric authentication that is integral to social dynamics as people often engage in face-to-face communications, which regularly requires the ability to verify each other's identities through quick and reliable recall techniques. Cross-race effect is a psychological phenomenon which suggests that people tend to recognize faces belonging to their own race more easily and better than faces from another race. Multiple theories have been posited to explain why this phenomenon arises, but there is no universally accepted model as yet. However, there is a controversial, yet common belief that Cross-Race Effect is a direct outcome of innate racial bias.

In an effort to dispute that idea of racial preconception, this project questioned Cross-Race Effect through the lens of pattern recognition. Following a research review methodology, it investigated foundational and current research that explored intuitive reliance on patterns during face recognition. It was hypothesized that – **If people are quizzed on face recognition tests, then they will exhibit regularities in the challenges experienced because our cognitive ability to encode faces employs systematic pattern recognition processes**. Following critical and in-depth evaluation of research evidence, it was established that face recognition categorically relies on a norm-based

coding of facial features. Additional findings tied face recognition to neural activity thereby establishing neurophysiological correlates for this cognitive process. The hypothesis was strongly supported with a clear indication that our ability to recognize and differentiate between faces relies on systematic coding of visual patterns, which are based on common structural particulars of facial geometry. Future implications suggested experimental research around association of face recognition with experience, aging and neuroplasticity for improved substantiation of the mechanics of face recognition. This research could assist in the development of assistive technologies for the rehabilitation of face recognition in brain conditions such as autism spectrum disorder (ASD).

Keywords: Face recognition, cross-race effect, norm-based coding

Table of Contents

Acknowledgments

I would like to acknowledge my teachers, Ms. Sharissa Dyke, Ms. Allison Pinnock and Mr. Monaghan for the interesting discussions that shaped the completion of my project. I would also like to acknowledge my school science fair coordinators Ms. Lai, Ms. Cruickshank and Mr. Ferg for their support. Finally, I would like to thank my friends and family for their encouragement every step of the way.

Introduction

Face recognition refers to the ascertainment and verification of the identity of a person through facial features. It is essentially a cognitive process that involves the capture, analysis and comparison of information about eyes, nose, mouth, hair, etc. Interestingly, studies in behavioural psychology and cognitive neuroscience have established that deep insights about the mental ability of face recognition can be gathered even through low technology measurements of behavioural responses that quantify – time and accuracy. In the absence of a meta level theory of such cognitive abilities, the research was crucial for an understanding of the elemental nature of perception. More recently, research in neuroscience has corroborated these findings with EEG measurements of active centers in the brain's frontal lobe and right Temporoparietal Junction (Deen et al., 2017; 2020).

Behavioural analysts claim that cross-race effect implicitly results from racial bias (Lebrecht et al., 2009), which is also known as own-race bias or other-race bias. However, popular literature as well as academic research exploring cognitive processes points to an underlying intuitive reliance on patterns (Hawkins & Blakeslee, 2004; Mckone, 2012). The focus of this project was

guided by these two perspectives to determine evidence of the use of pattern recognition in face recognition.

This report presents the framework of the project, detailed background research, a summary of the analysis and discussion of findings, conclusions, implications for further research, and a references list and acknowledgements.

Project Framework

Problem

Common belief claims that Cross-Race Effect is an outcome of innate racial preconceptions

Objective

1. Is our cognitive ability to recognize and differentiate between faces related to intelligence?
2. Is Cross-Race Effect purely an outcome of racial preconception?
3. In Face Recognition , why is in-between category differentiation better than within-category differentiation?

Hypothesis

~~**If** people are quizzed on face recognition tests, **then** they will exhibit regularities in the challenges experienced **because** our cognitive ability to encode faces employs systematic pattern recognition processes~~

Face Recognition intuitively relies on systematic pattern recognition processes

Methodology & Findings

Participants

Materials

Procedure

Initially, the project design intended to investigate foundational research in behavioural psychology that explored intuitive reliance on patterns during face recognition through thorough yest simplistic behavioural experiments. Critical analysis revealed that the investigation was eye-opening yet minimal in scope, which led to inclusion of more recent studies from cognitive neuroscience in order to ensure a more robust body of research evidence.

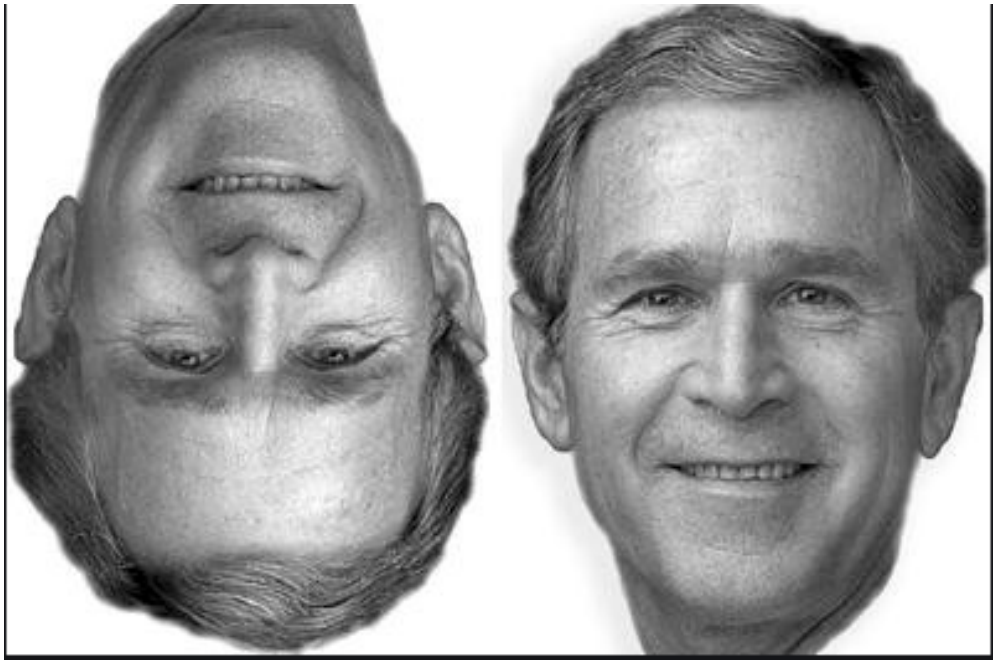
Through extensive reading and detailed analysis, a systematic understanding was developed which elucidated the basis of face recognition patterns in cross-race effect through processes mainly including:

1. Face inversion effect
2. Whole-part effect
3. Composite effect
4. Norm-based coding

1. Robert Yin, 1969

Face Recognition works differently from object recognition

- The **face inversion effect**: a greater decrease in memory for upside-down compared to upright stimuli for faces than other stimuli. Disproportionate inversion effect
- However, the level of decrease in memory for houses or other non-human objects is not as drastic; this could perhaps be because the brain stores a more complex representation of faces than non-human objects which in turn results in less flexibility in pattern recognition!!!



2. Tanaka and Farah, 1993; Young et al., 1987

Face Recognition is holistic

- **Whole-part effect**: subjects are better able to discriminate parts in the context of the whole face than when presented alone. Face representations are holistic, not decomposed into parts!
- **Composite effect**: when the top and bottom halves of two separate faces are put together, it is comparatively easier to identify the faces when they are misaligned than when they are smoothly aligned

Source: Young, Andrew W, Deborah Hellawell, and Dennis C Hay. "Configurational Information in Face Perception." *Perception (London)* 42.11 (2013): 1166–1178. Web.

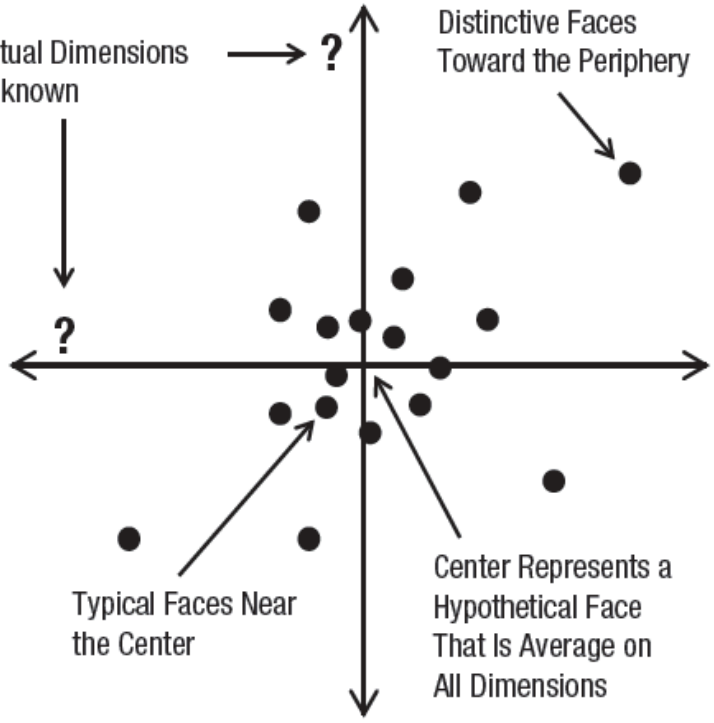
3. Elinor Mckone, 2012

Face Recognition uses norm-based coding

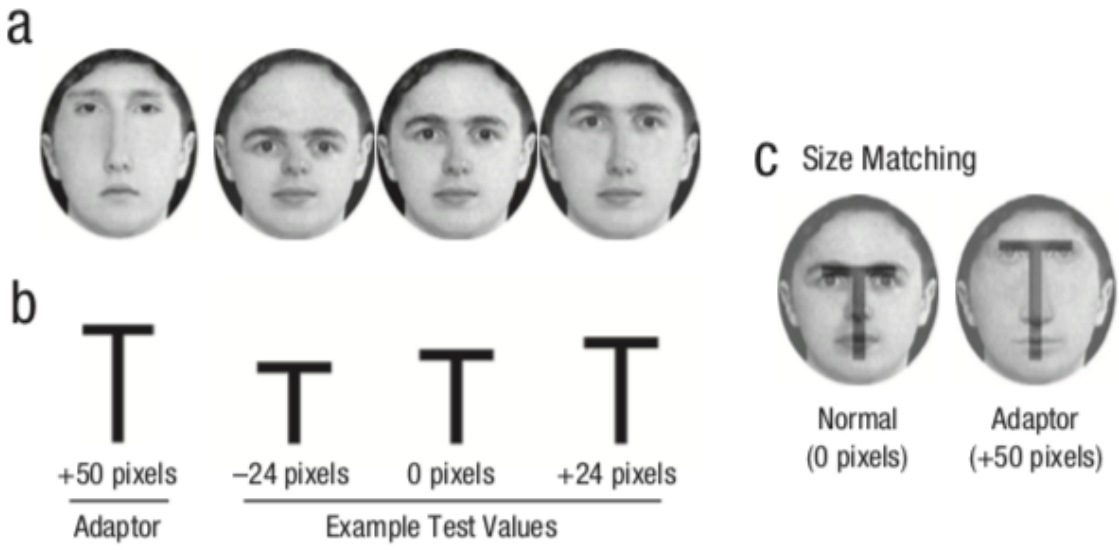
- **Face-space coding of face identity**: we code how a face differs from the average face or or face norm; an analog would be memorizing faces by placing them on a 2-D face space domain
- Note that in reality the brain is using a multi-dimensional space (not just 2-D or 3-D) to encode and respond to different facial attributes
- The face space coding is based on two types of responses: broadband-opponent coding and linear response functions.
- Broadband coding is set up such that one set of neurons respond to the maximal value of an attribute while another group of neurons respond to the minimal value of the same attribute
- The linear opponent coding function is a physiologically confirmed phenomenon which involves face selective neurons that have linear ramp functions to respond to several face attributes
- Elinor Mckone et al proved this with a series of experiments in which participants were shown faces with the distance between eyes and mouth stretched in two phases. In the first phase, the participants were shown stretched faces and asked to determine whether the face was normal while in the second phase, the

participants were pre-conditioned by showing an adaptor stretched face before showing them test images

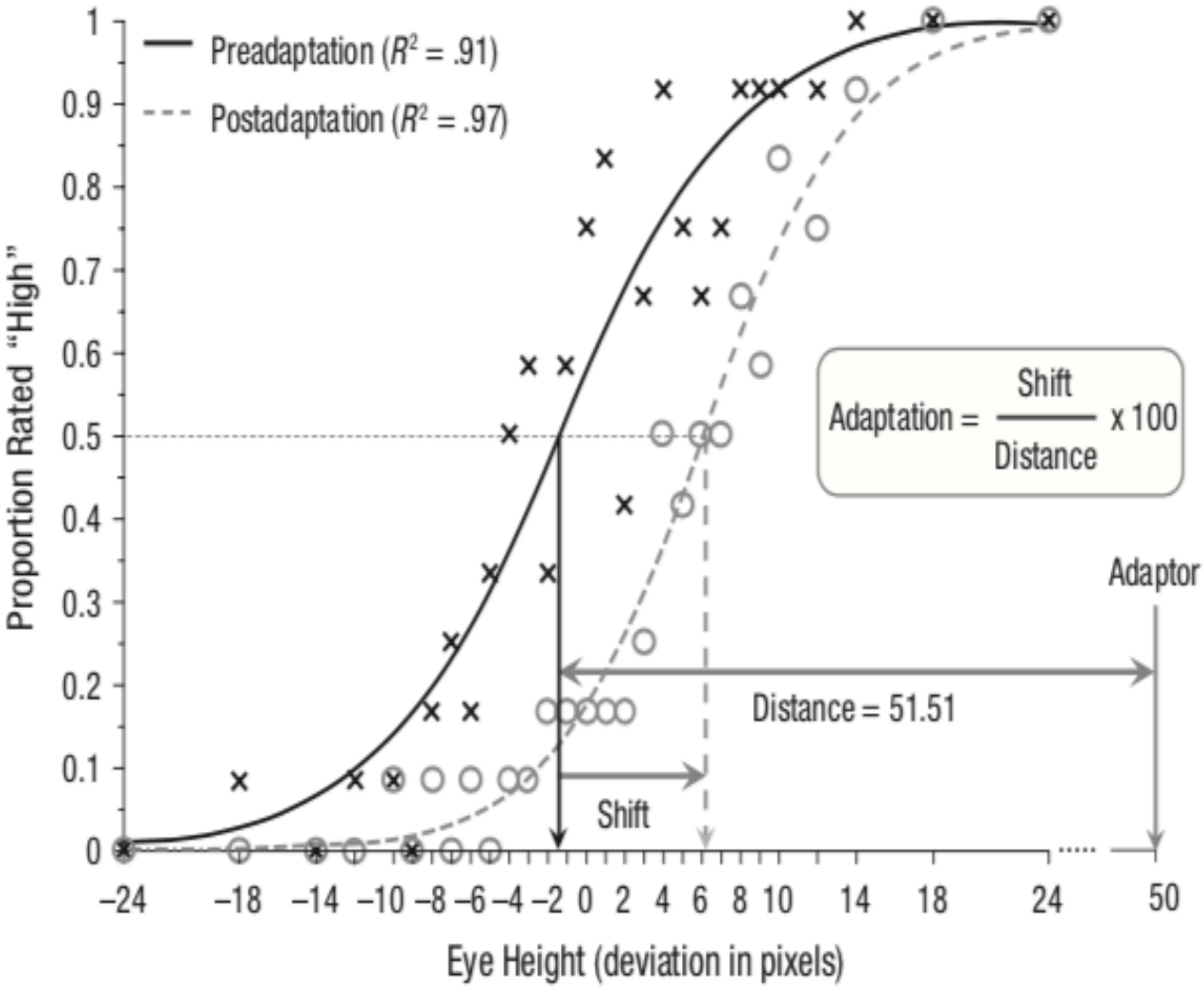
- Two key findings from this study were that people adapt to facial feature recognition once they have been conditioned to a new norm and secondly, that the facial feature recognition ability varies between individuals
- Everyone has their own face norm that changes over time. Every time you look at somebody for a while, your norm adjusts in the direction of that person.
- For example, the **other race effect: they all look alike**. Whoever they are, if I haven't spent as much time around them as I have around my people, then they all look alike! This is not racism; it is a fact about perception. We are just much better at discriminating faces of races that we're familiar with than faces of races that we are less familiar with. Since we have built a norm in face space for the people we have encountered, it is difficult to adapt that space to a race of people we have not been exposed to before.



Source: Dennett, Hugh & Mckone, Elinor & Edwards, Mark & Susilo, Tirta. (2012). Face Aftereffects Predict Individual Differences in Face Recognition Ability. Psychological science. 23. 10.1177/0956797612446350.



Source: Dennett, Hugh & Mckone, Elinor & Edwards, Mark & Susilo, Tirta. (2012). Face Aftereffects Predict Individual Differences in Face Recognition Ability. Psychological science. 23. 10.1177/0956797612446350.



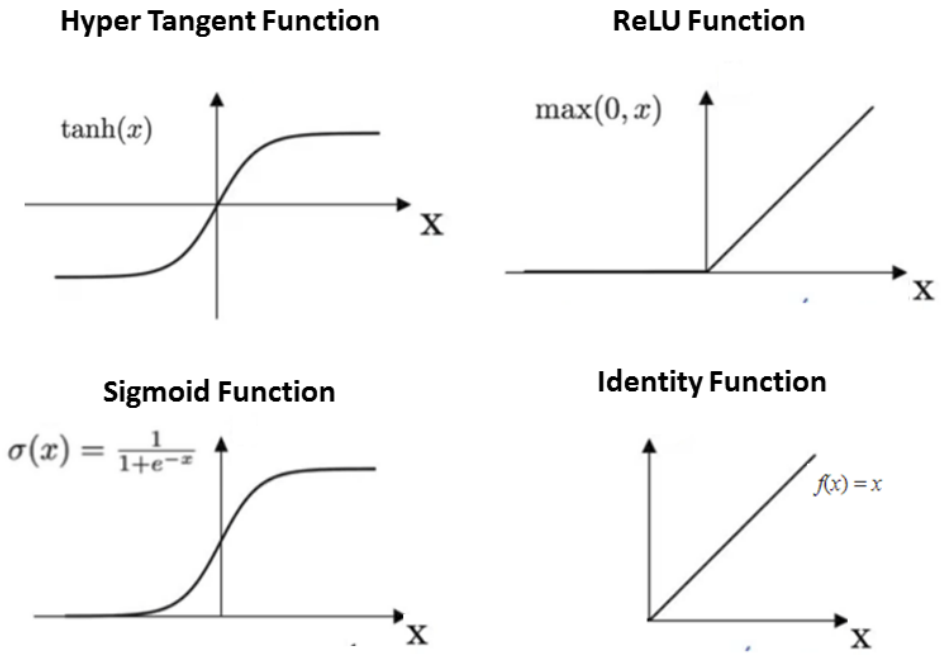
Source: Dennett, Hugh & Mckone, Elinor & Edwards, Mark & Susilo, Tirta. (2012). Face Aftereffects Predict Individual Differences in Face Recognition Ability. Psychological science. 23. 10.1177/0956797612446350.

4. Hawkins & Blakeslee, 2004

1. On Intelligence by Hawkins & Blakeslee, 2004

    - Jeff Hawkins' book on creating machine intelligence reinforces these ideas by noting that two of the key components to creating a complete artificial brain are memory and recall
    - Moreover, he notes that unlike our concept of classical computer, **the brain (more precisely, the neocortex):**
        - **Stores sequences of patterns**
        - **Recalls patterns auto associatively**
        - **Stores patterns in an invariant form**
        - **Stores patterns in a hierarchy**
    - Hawkin's maxims align well with the experimental observations on facial recognition from Robert Yin, Tanaka and Farah, Young et al and Mckone et al
    - Similar to the general memory and recall processes of the brain, face recognition must be based on patterns stored in a sequential, hierarchical and time invariant manner which are recalled auto associatively and not as a single complete entity

- The response to external stimuli can be approximated using common mathematical functions such as hypertangent, sigmoid, identity or linear ramp and ReLu functions



- These mathematical functions approximate the behavior of broadband-opponent and linear opponent functions of neurons and face detecting cells
- Additionally, these functions represent the norm based coding aspect and bias in neural processing explored by Mckone
- Unsurprisingly, these functions also form the basis for artificial intelligence and machine learning algorithms being used in current consumer technology which range from learning human behavior to predicting the future based on past and present stimuli

5. Deen, Kanwisher, Saxe et al., 2017; 2020

Neurophysiological Validation

- Ben Deen, Nancy Kanwisher and Dr. Rebecca Saxe validated the neuronal basis of face recognition by using EEG measurements
- Her experiments demonstrated that the frontal lobe of the brain and the rTPJ experience high activity when the brain is focused on the facial expressions while trying to interpret the mind of another person
- This confirms that the behavioral model proposed by Mckone and Hawkins has correlational basis in the human brain
- Therefore, by association we can conclude that face recognition is a result of neuronal activity; experimentally establishing the complete neural pathways involved in face recognition will require extensive EEG measurements on several living human bodies; as such it may be impossible

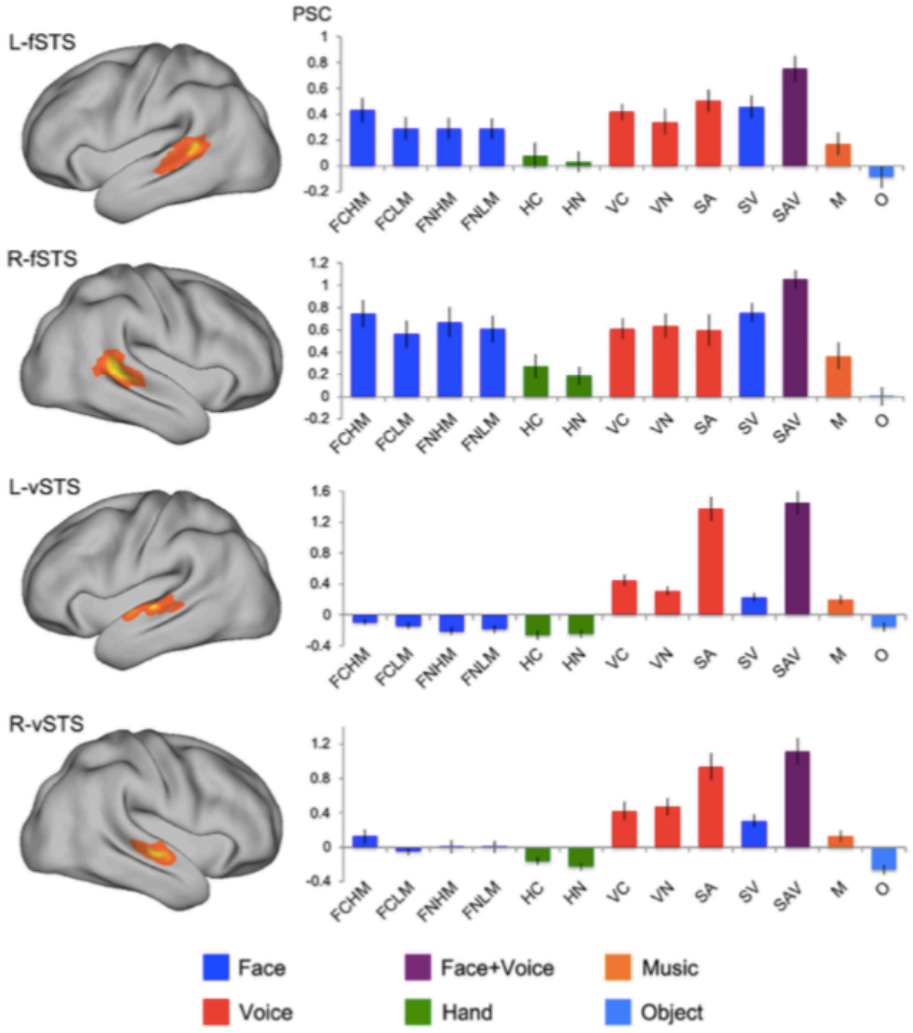


Fig. 3. Face-responsive posterior STS (fSTS) respond strongly to all face movements and vocal sounds, while voice-responsive middle STS (vSTS) responds selectively to speech sounds. Regions were defined using a faces > objects contrast (fSTS) and a voices > music contrast (vSTS). Left: heat maps of region-of-interest locations across participants. Right: responses of these regions (in percent signal change, PSC) across the thirteen experimental conditions, extracted from data independent from those used to define the regions. Condition labels: FC = communicative face movement, FN = noncommunicative face movement, HM = high mouth motion, LM = low mouth motion, HC = communicative hand movement, HN = noncommunicative hand movement, VC = communicative vocal sound, VN = noncommunicative vocal sound, SA = audio speech, SV = visual speech, SAV = audiovisual speech, *M* = music, *O* = objects.

NOTES: **Face Recognition varies across subjects, is heritable, not correlated with IQ**

- Is heritable: identical twins > fraternal twins
- not correlated with IQ: Must be a separate system/part of the brain than cognition with no reason to be correlated domains

6. List of Scientific Concepts

Explore the key points, problems and issues related to your subject matter. Ensure that your information is accurate and complete for your level of knowledge and understanding. Relevant graphs or tables from other workers' research may help to summarize your concepts. Remember to get permission to us other people's graphs, pictures etc., or at least to give the proper credits (CYSF, 2009).

Cross race effect
Face recognition
Pattern recognition
Face inversion effect
Whole-part effect
Composite effect
Norm-based effect

NOTE: See Research Methodology for definitions

Analysis & Conclusion

The final outcome of your investigation as supported by the research; relate your conclusion directly to your initial thesis/hypothesis (CYSF, 2009).

It is essentially a cognitive process that involves the capture, analysis and comparison of information about eyes, nose, mouth, hair, etc. https://www.jumio.com/facial-recognition-vs-facial-authentication/

Essentially, lessons from these foundational behavioural experiments, studies in cognitive neuroscience and **subsequent** applied research indicate three aspects about our ability to recognize and differentiate between faces

- It is different from object recognition
- It takes into account holistic face representation
- It uses norm-based coding systems

On a parallel note, the book **On Intelligence** reinforces the experimental observations of Yin, Young, Tanaka, McKone and their teams. Author and machine intelligence expert Hawkins notes that the key components for creating a **complete artificial brain** are **memory** and **recall**. He sums up findings in four maxims that the neocortex region of the human brain

- Stores sequences of patterns
- Stores patterns in an invariant form
- Stores patterns in a hierarchy
- Recalls patterns auto associatively

These rules are **seamlessly** applicable to the cognitive ability of Face Recognition because our face identification system

- Detects faces
- Captures data on facial features and
- Analyzes facial geometry and then
- Matches the input against stored databases of known faces

Therefore, insufficient databases can **directly** limit Face Recognition capabilities. Even a person with **high** IQ could either simply mistake a

person for another or completely fail to recognize a person because of limited awareness. All this data proves that face recognition is not an exclusive outcome of intelligence. Additionally, evidence of norm-based coding of faces refutes racial bias as **the cause** underlying inability to differentiate between faces. **Ambiguity** in recognizing faces is a **genuine** issue stemming from lack of our **limited** observation and **scarce** exposure.

On the whole, these findings **support** the initial prediction that our ability to recognize and differentiate between faces **categorically relies on pattern recognition processes**. Face recognition **engages** systematic encoding and decoding processes, which draw heavily from experience of visual patterns that are based on common structural particulars of facial geometry.

That explains **why** between-category differentiation is better than within-category differentiation! People may not experience confusion in delineating Asian faces from African faces from Caucasian faces but find it difficult to differentiate among Asian faces, among African faces or among Caucasian faces depending on their unfamiliarity quotient.

Going back to my early experiences with cross-race effect: By exclaiming “You look so alike!” our parents may not have **intended** but **definitely** telegraphed that they **did not notice our uniqueness**. We laughed it off, but I cannot deny that the interchangeability can be a minimizing experience. As research has revealed that lack of exposure underlies difficulty in face recognition, it is important that the role of exposure in the development of face recognition is explored.

Facial Recognition systems can be improved by training algorithms on a **larger** and **more** diverse set of faces and by using additional facial parameters which are independent of skin color to **compensate** for the limitations of the human brain. Minimizing and eventually eliminating systemic bias in surveillance and security systems is crucial for **equitable** law enforcement.

Ultimately, our ability to work around Covid-19 mandated face masks to recognize people suggests that it is **completely** possible to tackle limitations of Face Recognition patterns that lead to Cross-Race Effect!

The hypothesis of this project presented an expectation that if people are quizzed on face recognition tests, then they will exhibit regularities in the challenges experienced because our cognitive ability to encode faces employs systematic pattern recognition processes. Analysis of findings from the research consistently supported the hypothesis through evidence that our intuitive ability to encode faces draws on systematic patterning through behavioural phenomena such as the face inversion effect, the whole part effect, and the composite effect.

In a clear answer to the question posed by the project, evidence that challenges faced by subjects on face recognition tasks exhibit patterns clearly indicated that people rely on pattern recognition when they engage in the recognition of faces. This goes to indicate that cross-race effect is associated with face recognition patterns.

Following critical evaluation of the body of research, it can be summarized that face recognition must be based on patterns stored in a sequential, hierarchical and time invariant manner which are recalled auto associatively and not as a single complete entity. As such, the facial recognition mechanisms are based on

conditioning the neural system using a range of external stimuli. This leads to a fairly flexible and non-linear mechanism that can tackle a wide range of variation in patterns and external stimuli. The non-linearity spans multi-dimensional face space coding and is not limited to 2D or 3D maps of faces. However, it is also evident that when presented with facial features beyond the pre-conditioned range, people have difficulty recognizing faces accurately in the extended range. Equally fascinating is the fact the neural cells engaged in facial recognition can be re-conditioned fairly quickly to a different range of external stimuli which has been experimentally validated. It is furthermore evident that this capability varies across individuals depending on the perceptive capability of the elemental neurons and the collective neural system of every individual. This research strongly confirms our hypothesis that people use patterns to recognize faces and that there are systematic differences in facial recognition processes which account for the cross-race effect.

Limitations & Implications

Discuss how you could take your research further, or what experiments you could undertake to support your conclusion. Include an explanation of why

people would be interested in knowing your results and how they can be used

(CYSF, 2009).

Limitations & Implications

What is important here is that this effect or bias has **urgent implications** for the expanding use of **Facial Recognition technology** as a commonly used biometric to authenticate identities. I mean, we depend on it without much thought to unlock our phones, use social media, etc. and law enforcement uses it as a surveillance tool to identify criminal suspects, witnesses and other people of interest.

In fact, high level results from **NISTIR 8280** - a landmark US Federal study released in December of 2019 showed that Facial Recognition systems **misidentified** people of **color** more often than **white** people. In a press release, Patrick Grother – primary author of the report was quoted “While it is usually incorrect to make statements across algorithms, we found empirical evidence for the existence of demographic differentials in the majority of the face recognition algorithms, we studied.”

The shocking finding has intensified focus on mitigation of inherent racial bias in surveillance tools.

The study does not discriminate between Facial Recognition and Facial

Identification.

The findings from this research review demonstrate that individual differences

can be investigated to understand aspects of cognitive neuroscience because

we can learn about the cognitive ability of face recognition using simple

measurements of behavioral responses based on accuracy and time.

Although the sample size and trials in this practice project were too small to establish validity of the findings, the distinct trends are encouraging for extended research using sizeable sample sets and multiple trials. It would be interesting to explore the following pertinent questions through experimental studies:

- What is the role of experience in face recognition?
- Does the age of the perceiver and age of the subject of identification influence face recognition?

These efforts can add unique information to the existing pool of knowledge about human behavior and cognition improve our understanding of the role of spontaneous collateral interactions.

References

Calgary Youth Science Fair. (2009). Elements of a non-experimental/research project

Deen, B., Richardson, H., Dilks, D. et al. (2017). Organization of high-level visual cortex in human infants. Nat Commun 8, 13995. Retrieved on December 2, 2020 from https://doi.org/10.1038/ncomms13995

Deen, B., Saxe, R & Kanwisher, N. (2020) Processing communicative facial and vocal cues in the superior temporal sulcus. NeuroImage, Volume 221, 117191, Retrieved on December 2, 2020 from https://doi.org/10.1016/j.neuroimage.2020.117191

Kanwisher, Nancy. 2020. Nancysbraintalks. MIT. Retrieved on December 15, 2020 from http://nancysbraintalks.mit.edu

Lebrecht S, Pierce LJ, Tarr MJ, Tanaka JW (2009) Perceptual Other-Race Training Reduces Implicit Racial Bias. PLoS ONE 4(1): e4215. Retrieved on August 13, 2020 from https://doi.org/10.1371/journal.pone.0004215

McKone E, Crookes K, Jeffery L & Dilks D D. (2012) A critical review of the development of face recognition: Experience is less important than previously

believed, Cognitive Neuropsychology, 29:1-2, 174-212, Retrieved on November 4, 2020 from DOI: 10.1080/02643294.2012.660138

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 46A*(2), 225–245. Retrieved on October 26, 2020 from https://doi.org/10.1080/14640749308401045

Taylor, John. (2005). On Intelligence, Jeff Hawkins, Sandra Blakeslee Times Books (2004). Artificial Intelligence. 169. 192-195. 10.1016/j.artint.2005.10.011

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*(1), 141–145. Retrieved on September 1, 2020 from https://doi.org/10.1037/h0027474

Young AW, Hellawell D, Hay DC. Configurational information in face perception. Perception. 1987;16(6):747-59. Retrieved on October 10, 2020 from doi: 10.1068/p160747. PMID: 3454432

Appendix: Research Publications

Journal of Experimental Psychology
1969, Vol. 81, No. 1, 141–145

LOOKING AT UPSIDE-DOWN FACES[1]

ROBERT K. YIN[2]

Massachusetts Institute of Technology

Memory for faces was compared with memory for other classes of familiar and complex objects which, like faces, are also customarily seen only in one orientation (mono-oriented). Performance was tested when the inspection and test series were presented in the same orientation, either both upright or both inverted, or when the two series were presented in opposite orientations. The results show that while all mono-oriented objects tend to be more difficult to remember when upside-down, faces are disproportionately affected. These findings suggest that the difficulty in looking at upside-down faces involves two factors: (*a*) a general factor of familiarity with mono-oriented objects; and (*b*) a special factor related only to faces.

It is a well-known fact that pictures of human faces, when viewed upside-down, are extremely difficult to recognize (Arnheim, 1954, p. 86; Attneave, 1967, p. 26; Köhler, 1940, p. 60). Köhler not only noted this, but also speculated that the difficulty was attributable to the loss of "facial expression" in the inverted picture. More recently, investigators have examined this phenomenon in several ways. Brooks and Goldstein (1963) showed that recognition of inverted faces is worse than that of upright faces when children are asked to identify snapshots of their classmates. That memory for inverted faces is poorer than memory for upright faces among adults has been shown in a paired-associate task (Goldstein, 1965) and a recognition task (Hochberg & Galper, 1967).

These studies have not indicated the extent to which the difficulty in viewing an upside-down face is related specifically to the face. An alternative hypothesis would be that any set of objects customarily seen in one orientation, i.e., mono-oriented, might be more difficult to recognize when inverted. Some evidence for this hypothesis was reported by Henle (1942), who showed that alphabetic letters were correctly perceived more frequently than their mirror reversals by *S*s familiar with the letters, and by Ghent (1960), who found that young children are markedly dependent on familiar orientation for recognizing realistic figures. In addition, Dallett, Wilcox, and D'Andrea (1968) reported that memory for upright magazine pictures was better than that for the same pictures when presented upside-down. The investigators did not indicate, however, the extent of homogeneity among the pictures or the degree to which the pictures were of objects that are customarily mono-oriented.

The present experiments were designed to test whether a general impairment on mono-oriented objects when inverted could account for the difficulty with viewing upside-down faces. More specifically, performance on upright and inverted tasks for faces was compared with that for other classes of everyday objects having a priori properties similar to faces in being mono-oriented, familiar, complex, and not easily verbalized, i.e., objects that are not distinguished from each other by the use of simple labels.

To test performance, a forced-choice recognition memory task was used. In this task, *S*s were shown individual pictures (an inspection series) and then presented with pairs of pictures (a test series). In the test series they indicated the one of the pair they thought they had seen in the inspection series. Three experiments were conducted. In the first and third, the

[1] This study was supported by a grant from the John A. Hartford Foundation, Inc. (New York, N. Y.) to H.-L. Teuber and a predoctoral award to the author from the National Science Foundation. The author gratefully acknowledges the advice and encouragement of H.-L. Teuber throughout all phases of this work.

[2] Requests for reprints should be sent to Robert K. Yin, Department of Psychology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

orientation of the materials in both the inspection and test series was the same (both upright or both inverted). In the second, the orientations were opposite (inspected upright and tested upside-down or inspected upside-down and tested upright).

Experiment I

Method

Subjects.—There were 26 paid volunteers, 13 men and 13 women, ranging from 18 to 31 yr. of age (mean age = 21.7 yr.). These were undergraduate and graduate students attending summer schools in the Boston area and represented a wide variety of geographical origins and academic interests.

Materials.—There were 64 different pictures, all black and white, within each of four types of materials: faces, houses, airplanes, and men in motion. All pictures were pasted on a 3 × 5 in. card for presentation.

The faces were studio pictures of adult males, chosen to be similar with respect to general age, expression, and lack of outstanding distinguishing features, such as glasses, beards, or unique marks. All poses were full face, and the pictures were trimmed just under the chin to eliminate as much clothing as possible. The houses were generally of the same architecture, but were not as uniform as the faces in orientation of view or size of picture. In addition, since all were actual photographs, the pictures included trees and other natural features, although they were trimmed to minimize the presence of distinguishing features, such as fences, front stoops, and roof markings.

Neither the airplanes nor the men in motion were real photographs, but were caricatures. The planes were sideview silhouettes of all types and models (military, commercial, and private) of planes found in the world today. The men-in-motion pictures consisted of the same cartoon stick figure engaged in various everyday movements and postures, with no other objects present in any of the pictures.

Procedure.—Each *S* looked at an inspection series of 40 pictures, presented singly and turned by *E* at a rate of 3 sec. per picture. Then a test series, consisting of 24 pairs of pictures, was presented. Each pair contained 1 old picture (an exact duplicate of a picture in the inspection series) and a new picture (one not previously shown), and *S* had merely to indicate which picture in each pair was the old one. The *S* proceeded at his own rate in the test series. Since only 24 pairs were in the test series, there were 16 pictures in the inspection series which did not recur in the test series.

Each inspection and test series constituted a block and was a mixed list, containing two different types of materials, 20 of each in the inspection series and 12 pairs of each in the test series. The order of presentation of the 40 inspection series pictures was randomized, with the two exceptions that neither of the two materials was shown for more than four consecutive cards and that there were always at least 2 of the nonrecurring pictures, one of each type of material, at either end of the series. The order of the 24 test series pairs was dictated by the order of pictures in the inspection series, so that there was a constant lag between each inspection picture and its occurrence in the test series.

All *S*s went through four such blocks of inspection and test series, viewing two blocks rightside-up (both series upright) and two upside-down (both series inverted). Thus each *S* performed in all experimental conditions, viewing the four materials in two presentations. The order of presentation among the blocks was balanced in the following manner: (*a*) Each *S* was shown all four materials (two blocks) first; half of the *S*s saw these two blocks upside-down first, the other half rightside-up first; (*b*) the mixing of the materials was such that roughly one-third of the *S*s had blocks consisting of faces–houses or airplanes–men-in-motion, one-third had blocks of houses–airplanes or faces–men-in-motion, and the remaining third had blocks of airplanes–faces or houses–men-in-motion; (*c*) the blocks were alternated so that each picture was shown rightside-up as often as it was upside-down; and (*d*) the sexes were balanced with regard to all of these conditions.

TABLE 1

Mean Errors, Exp. I

Material	Presentation			
	Test and inspection series upright		Test and inspection series inverted	
	M	*SD*	*M*	*SD*
Faces	.89	1.09	4.35	1.41
Houses	2.23	1.60	3.42	1.36
Airplanes	3.65	1.69	3.85	2.03
Men in motion	2.35	1.27	3.27	1.58

Results

The mean errors, with their standard deviations, appear in Table 1. An analysis of variance of the error scores showed that there were significant differences as a function of presentation, $F\ (1, 25) = 90.90$, $p < .0005$, materials, $F\ (3, 75) = 6.63$, $p < .001$, and their interaction, $F\ (3, 75) = 9.18$, $p < .0005$.

Although all materials were more difficult in the inverted presentation, the extent to which each type of material contributed to this effect varied. Using *t* tests for matched pairs, two-tailed, the effect of inversion was

greatest for faces, t (25) = 8.48, $p < .001$, significant but not as great for the houses, t (25) = 3.01, $p < .01$, and the men in motion, t (25) = 2.15, $p < .05$, and not significant for the airplanes, t (25) < 1.

The materials also differed in their overall difficulty. Although this finding is not of primary interest here, the major reason for it was that the airplanes tended to be the most difficult material in either presentation.

Of greater interest is the fact that the Presentations × Materials interaction was significant. Further analysis showed that this was due mainly to the faces, which were easier than all the other materials when viewed upright, t (25) = 7.31, $p < .001$, but more difficult than the rest when viewed upside-down, t (25) = 2.53, $p < .02$. Examination of the individual scores produced added evidence of the existence of a difference between faces and the other materials. In general, those who did better in the inverted orientation also tended to be the ones who did better in the upright orientation. However, for faces, the reverse was true. Taking the average inverted score for houses, airplanes, and men in motion, and arbitrarily assigning all *S*s to a better group ($n = 14$, average error = 2.88) and a worse group ($n = 12$, average error = 4.25), the better group is also better on the upright task (average error = 2.36), while the worse group is still the worse one (average error = 3.19). Using a t test for independent samples, two-tailed, the difference between the two groups in their upright scores is significant at the $p < .05$ level, t (24) = 2.46.

On the other hand, arbitrarily assigning all *S*s by their score on the inverted-face task to a better group ($n = 14$, average error = 3.29) and a worse group ($n = 12$, average error = 5.58), we find that the better group is now the *worse* one on the upright-face task (average error = 1.29), while the worse group is the *better* one (average error = .42). This difference on the upright-face task is significant at the $p < .05$ level, t (24) = 2.14.

Sex differences.—Men and women did not differ in their total upright or inverted scores. There were differences between materials, however, in that the men's average upright and inverted score for airplanes was better than that of the women, t (24) = 2.26, $p < .05$, while the women's average upright and inverted score for houses was better, but not significantly, than that of the men, t (24) = 1.91, $p < .10$. In both cases the t test was for independent samples and two-tailed. There were no sex differences for either the faces or the men in motion.

Order of presentation.—There were no differences when the groups were characterized by viewing order, upright first or inverted first, or by the mixture of the materials in the different blocks.

TABLE 2

MEAN ERRORS, EXP. II

Material	Presentation			
	Up-Down		Down-Up	
	M	*SD*	*M*	*SD*
Faces	3.81	1.71	5.14	1.39
Houses	2.86	1.83	3.43	1.47
Airplanes	3.19	1.94	4.14	1.58
Men in motion	4.05	1.79	4.24	1.72

EXPERIMENT II

Experiment II required *S*s to make a mental inversion of the materials, presenting the inspection and test series in opposing orientations.

Method

Subjects.—There were 21 paid volunteers, 13 men and 8 women, ranging from 18 to 26 yr. of age (mean age = 21.1 yr.). The general nature of the sample was the same as that of Exp. I.

Materials and procedure.—The materials were the same as those used in Exp. I, and the overall procedure was exactly the same with one exception: For each *S* the two presentations were up-down (inspection series presented upright and test series inverted) or down-up (inspection series presented inverted and test series upright). As in Exp. I, each *S* performed in all experimental conditions, viewing the four materials in both presentations.

Results

Table 2 contains the mean errors with their standard deviations. An analysis of variance of the errors shows that there were

significant differences as a function of presentation, F (1, 20) = 11.67, $p < .01$, and materials, F (3, 60) = 4.37, $p < .01$, but not of their interaction, F (3, 60) = 1.09.

Although all materials were worse in the down-up presentation than the up-down presentation, faces were the most affected. Using t tests for matched pairs, two-tailed, the difference in presentation was significant for the faces, t (20) = 3.12, $p < .01$, but not for the houses, t (20) = 1.44, airplanes, t (20) = 1.59, or men in motion, t (20) < 1. The materials again differed in overall difficulty, this time mainly because the houses were easiest in both presentations.

Sex differences.—As in Exp. I, men and women did not differ in their total scores. Men tended to do better on airplanes in both presentations, but there were no differences for the other materials.

Order of presentation.—There were no differences due to order of presentation.

Comparison of results between Experiments I and II.—In general, for each material the up-down performance (Exp. II) tended to be worse than the upright performance (Exp. I) by about the same amount that the down-up (Exp. II) was worse than the inverted (Exp. I). This consistent decline reflects the added difficulty imposed by the necessity for inverting the pictures mentally.

With the faces, however, the up-down performance was disproportionately worse than that of the upright. This is apparent if for each material, one compares the up-down and down-up difference from Exp. II with the upright-inverted difference from Exp. I. Using t tests for independent samples, two-tailed, the difference between these differences is significant for faces, t (45) = 3.55, $p < .001$, but not for houses, t (45) = 1.09, airplanes, t (45) = −.99, or men in motion, t (45) = 1.26. Thus, while all the materials tended to become more difficult in Exp. II, the upright faces were disproportionately affected.

The major finding from the first two experiments is that faces are different from the other materials in two ways. First, although all the materials were more difficult when viewed upside-down, faces were especially difficult (Exp. I). Second, although all the materials were more difficult when S was required to make a mental inversion, the upright face was again disproportionately affected (Exp. II).

At least two interpretations of these results may be made. The first is that there is something special about faces that makes them particularly difficult even when compared with other mono-oriented objects. The second is that the difference between faces and the other materials is due solely to differences in degree of difficulty among the materials when presented upright. According to this interpretation, the easier a material when upright, the more it will be affected by inversion, and thus the disproportionate difficulty in remembering upside-down faces merely reflects the fact that the faces were the easiest material when viewed rightside-up.

To try to differentiate between these two interpretations of the results, a third experiment was designed in which memory for faces was compared with memory for another class of objects which, while meeting all the previous criteria in being mono-oriented, complex, familiar, and not easily verbalized, would also be as easy to remember as faces in the upright presentation. In addition, since the faces used in the first two experiments were studio pictures, the third experiment also investigated the possibility that the difficulty in remembering faces could be attributed solely to the special effects of light and shadow inherent in such pictures. Therefore an artist's line drawings of adult male faces, made to specification so that they were similar to the studio pictures but with all light and shadow cues eliminated, were used.

Experiment III

Method

Subjects.—There were 23 paid volunteers, all male undergraduates attending the regular school session.

Materials.—There were 36 different pictures, all black and white, of two types of materials: artist's sketches of faces and drawings of faceless figures clothed in different period costumes. The sketches were cropped very severely, so that no hair, ears, or chin lines were present. The costumed figures

were also cropped so that only the faceless head and torso of each figure were shown.

Procedure.—The procedure was the same as that of Exp. I, except that the inspection and test series were both shorter. The inspection series contained only 18 pictures, while the test series contained 18 pairs of pictures. Each block of inspection and test series was composed of equal numbers of faces and costumes, and each *S* viewed two blocks, one rightside-up and the other upside-down.

Results

For faces, the upright errors were $M = 1.35$, $SD = 1.13$, and the inverted errors were $M = 2.69$, $SD = 1.40$. For the costumes, the upright errors were $M = .48$, $SD = .71$, and the inverted errors were $M = .78$, $SD = .78$. Using t tests for matched pairs, two-tailed, the difference between upright and inverted errors was significant for the faces, $t\ (22) = 4.00$, $p < .001$, and strong but not quite significant for the costumes, $t\ (22) = 1.91$, $p < .10$. More important, performance for the costumes was better than that for the faces in the upright presentation, $t\ (22) = 3.14$, $p < .01$, as well as in the inverted presentation, $t\ (22) = 5.31$, $p < .001$. Thus the faces, although not the easier material in the upright presentation, were still more affected by inversion when compared with the costumes.

Discussion

The results of the third experiment indicate that upside-down faces are difficult to remember even when the differences between materials are such that the faces are not the easiest to remember in the upright presentation. In addition, the difficulty is not limited to studio photographs, but can also be shown to exist with line drawings.

The data from all three experiments support the interpretation that the inverted face is especially difficult to remember because of two factors: a general factor of familiarity with mono-oriented objects and a special factor involving only the faces. The general factor is seen as affecting all of the materials used, making them more difficult to recognize when upside-down; the special factor relates to the disproportionate difficulty created by the inverted face.

It is interesting to speculate what such a special factor might involve, even though this question is unanswerable from the present experiments. One clue may be provided by verbal reports from *S*s when they are asked how they tried to remember the various materials. They seemed to use two alternative strategies, either searching for some distinguishing feature or attempting to get a general impression of the whole picture. The first tended to be used for most of the materials; the second was used mostly for faces, with *S* trying to remember some personal impression made by the face. None of the *S*s, however, reported being able to use the second strategy when looking at the inverted face. Whatever the relevant variables, further investigation into the difficulty with inverted faces may by implication tell us something about how people recognize normal (i.e., upright) faces and how we distinguish one face from another.

References

Arnheim, R. *Art and visual perception: A psychology of the creative eye.* Berkeley: University of California Press, 1954.

Attneave, F. Criteria for a tenable theory of form perception. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form.* Cambridge: M.I.T. Press, 1967.

Brooks, R. M., & Goldstein, A. G. Recognition by children of inverted photographs of faces. *Child Development,* 1963, **34**, 1033–1040.

Dallett, K., Wilcox, S. G., & D'Andrea, L. Picture memory experiments. *Journal of Experimental Psychology,* 1968, **76**, 312–320.

Ghent, L. Recognition by children of realistic figures presented in various orientations. *Canadian Journal of Psychology,* 1960, **14**, 249–256.

Goldstein, A. G. Learning of inverted and normally oriented faces in children and adults. *Psychonomic Science,* 1965, **3**, 447–448.

Henle, M. An experimental investigation of past experience as a determinant of visual form perception. *Journal of Experimental Psychology,* 1942, **30**, 1–22.

Hochberg, J., & Galper, R. E. Recognition of faces: I. An exploratory study. *Psychonomic Science,* 1967, **9**, 619–620.

Köhler, W. *Dynamics in psychology.* New York: Liveright, 1940.

(Received September 11, 1968)

Parts and Wholes in Face Recognition

Article *in* The Quarterly Journal of Experimental Psychology A · June 1993

DOI: 10.1080/14640749308401045 · Source: PubMed

2 authors, including:

J. W. Tanaka
University of Victoria
61 PUBLICATIONS **6,028** CITATIONS

SEE PROFILE

THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY, 1993, *46A* (2) 225–245

Parts and Wholes in Face Recognition

James W. Tanaka

Oberlin College, Oberlin, Ohio, U.S.A.

Martha J. Farah

University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A.

Are faces recognized using more *holistic* representations than other types of stimuli? Taking holistic representation to mean representation without an internal part structure, we interpret the available evidence on this issue and then design new empirical tests. Based on previous research, we reasoned that if a portion of an object corresponds to an explicitly represented part in a hierarchical visual representation, then when that portion is presented in isolation it will be identified relatively more easily than if it did not have the status of an explicitly represented part. The hypothesis that face recognition is holistic therefore predicts that a part of a face will be *disproportionately* more easily recognized in the whole face than as an isolated part, relative to recognition of the parts and wholes of other kinds of stimuli. This prediction was borne out in three experiments: subjects were more accurate at identifying the parts of faces, presented in the whole object, than they were at identifying the same part presented in isolation, even though both parts and wholes were tested in a forced-choice format and the whole faces differed only by one part. In contrast, three other types of stimuli—scrambled faces, inverted faces, and houses—did not show this advantage for part identification in whole object recognition.

Parts and Wholes in Face Recognition

An important issue in face recognition research is whether faces are recognized on the basis of their individual features or more holistically, on the basis of their overall shape. As long ago as the nineteenth century, Galton (1879) proposed that holistic information may be more vital to face recog-

Requests for reprints should be sent to James W. Tanaka, Department of Psychology, Severance Lab, Oberlin College, Oberlin, OH 44074, U.S.A.

We would like to thank John Duncan, Andy Young, Paddy McMullen, and an anonymous reviewer for their comments on this manuscript. We are also grateful to Jonathan Schooler for the use of his Mac-a-Mug software and Safman Aly for designing the house stimuli used in Experiment 3. This research was supported by ONR contract N0014-89-J3016, NIH grants NS23458 and NS06209, NIH RCDA K04NS01405, and NIMH training grant 1 T32 MH 19102-02.

nition than the identification of individual features, and modern researchers continue to pursue this hypothesis (see Bruce, 1988, for a detailed review and evaluation). However, the empirical evidence to substantiate such a claim remains equivocal.

One factor that has contributed to the difficulty of resolving this issue is the lack of clear, generally accepted definitions of the concepts *holistic* and *featural*. Without clear definitions of what these terms mean, it is difficult to operationalize them in experimental tests. In this article, we propose an explicit definition of the holistic/featural distinction. We then use that definition to interpret the available evidence and to design new empirical tests.

We take as a starting point the idea that visual object representations are hierarchically organized, such that the whole object is parsed into portions that are explicitly represented as parts (cf. Palmer, 1977). For example, a house might be decomposed by the visual system into a set of doors, windows, a roof, etc. The resulting representation of the house would consist of representations of these parts, somehow linked together. Some objects may be decomposed into many parts, others into relatively few or none at all. In this context, the claim that faces are recognized holistically would mean that the representation of a face used in face recognition is not composed of representations of the face's parts, but more as a whole face. Although visual information from the eyes, nose, etc. would of course be included in the face representation, that information would not be contained in representational *packets* corresponding to the parse of the face into these features. In other words, these parts or features would not be explicitly represented as structural units in their own right in the final face representation. Instead, faces would be recognized "all of a piece"—or, to use a somewhat embattled term, as *templates*. The alternative hypothesis, that faces are recognized featurally, implies that faces are represented in terms of representations of their component parts. The holistic/featural distinction need not be a strict dichotomy, as both types of representations may exist and be used to different degrees for different classes of objects. Because of this, we would like to recast the question of whether faces are recognized holistically as the question: does face recognition rely on holistic visual representations to a greater degree than do other forms of pattern recognition? Before presenting our experiments, we briefly review what is known about this issue from other studies.

Bradshaw and Wallace (1971) addressed the issue of whether faces are perceived featurally using a matching task in which pairs of simultaneously presented Identikit faces were to be judged "same" or "different". They found that the number of features by which a pair of faces differed predicted the latency of correct "different" responses, with shorter reaction times associated with more featural differences. Based on this finding, they argued that faces were inspected according to a serial self-terminating

search and were therefore perceived in terms of their features. In another study with Identikit faces in a simultaneous matching task, Mathews (1978) found evidence that faces are perceived both featurally and holistically. He found that subjects' reaction times increased linearly for detecting differences in eyebrow, nose, and mouth features, thus indicating a top-to-bottom serial comparison process. However, he also found that reaction times for detecting changes in hair, eyes, and chin were essentially the same across features, which he interpreted as evidence for a holostic or at least a parallel comparison process. Mathews reconciled these results by proposing a dual processing strategy in which features are checked both in serial and parallel.

Other researchers have also arrived at the conclusion that faces are perceived both featurally and holistically using slightly different paradigms and methods of data analysis. For example, Smith and Nielsen (1970) used a matching paradigm with schematic line drawings of faces, but introduced a delay between the two stimuli to be matched. At delays of 1 or 4 sec, they found results similar to those of Bradshaw and Wallace: the more features differed between two faces, the more quickly the faces were judged to be different. In addition, by varying the number of features present in the faces, they were able to examine the effects of number of features on the latencies of "same" judgments. They found that "same" judgments were not affected by the total number of features, conflicting with their findings from the "different" trials and suggesting that subjects were not serially comparing the individual features. However, at the longer delays of 10 sec, both "same" and "different" trials yielded patterns of reaction times consistent with a feature by feature comparison process. Sergent (1984) used a matching task with Photofit faces and reasoned that if features are processed independently, the time to make a "different" response when faces varied by two features should never be faster than the time to make a "different" response when faces varied by the most salient feature. Her data indicated that changes in chin contour led to faster reaction times than changes to either eyes or to a feature that she termed "internal spacing"—the distance between the nose and mouth. She found that the time to decide that two faces differed with respect to their chin contours and internal spacings was faster than the time to decide that two faces differed with respect to their chin contours only. Sergent concluded that whereas some features seemed to be processed independently of each other (e.g. eyes and chin contour), other features (e.g. chin contour and internal spacing) interact and are processed more holistically. However, as Bruce (1988) has pointed out, Sergent's conclusion was weakened by the fact that only one feature type—internal spacing—produced the holistic effect, and it was not a feature in the sense of being a part of the face but was, rather, a relation among parts.

The foregoing studies suggest that faces may be perceived both in terms

of their individual component features and in terms of more holistic ensembles of those features. However, these studies fall short of answering the question posed earlier—does face recognition rely on holistic visual representations to a greater degree than other forms of pattern recognition?—for several reasons.

First, with the exception of Sergent's study, all of the experiments appear to be based on the assumption that the number of features in a face, or the number of features by which two faces differ, would not affect performance if subjects were using a holistic representation. However, this is not necessarily true. The more features that are in a face, the more information there is in the holistic representation, and the longer it could take to use that holistic representation. Similarly, the more features differ between two faces, the more different their two holistic representations will be, and the more easily a discrepancy will be discovered. These studies might better be regarded as testing whether face matching is carried out independent of capacity limitations, regardless of whether the matching is holistic, parallel featural, or serial featural.

Second, none of the studies is designed to distinguish between the possibility that faces are represented by features that can be processed in parallel, and the possibility that faces are represented holistically, that is, without explicit representations of the features.[1] The question of whether facial features, when and if they are explicitly represented, can be compared in parallel or only serially is an interesting one, but not the one to which we are addressing ourselves in this article.

Third, it is not clear how similar the visual processes elicited by these tasks are to those used in normal face recognition. All of the studies described above involved face matching rather than face recognition. Subjects may well use different strategies and, as a result, different types of visual representations when they can consult a percept or short-term memory representation of the face to be matched, rather than having to consult the long-term memories used for face recognition. The generalizability of these studies to real face recognition can also be called into question on the grounds of the stimulus pictures used, some of which were highly artificial and schematic.

Finally, without comparing the results obtained with faces in these paradigms to results obtained with objects other than faces, we cannot assess the extent to which the holistic or featural representation of faces is special

[1]Some researchers have couched the question in terms of configurational versus featural processing, but these studies do not address this distinction either. Although parallel processing of multiple features is presumably necessary for configural processing, they are not the same thing. Features could be processed in parallel without representing their spatial configuration. We will return to the issue of configuration in face recognition and how it relates to the idea of holistic representation in the General Discussion.

to faces. A series of experiments by Bruce, Doyle, Dench, and Burton (1991) avoids many of these problems. They presented subjects with sets of computer-generated faces with identical features, but slightly different spatial configurations in an incidental memory task. They found that subjects abstracted the prototypical configuration for each set, and that this tendency to identify the prototype as most familiar was greater for faces than for houses. This finding argues strongly for a special role of nonfeatural information in face recognition. However, it does not speak directly to the issue of holistic representation.

Our approach to the issue of holistic versus featural representations in face recognition is based on the following logic: if some portion of a stimulus is explicitly represented as a part in the stimulus representation, then it should be relatively more easily recognized as coming from that stimulus, when viewed in isolation, than if the stimulus representation does not contain it as an explicitly represented part. Similar reasoning has been used by Bower and Glass (1976) and Palmer (1977) to distinguish between psychologically real and less plausible parsings of patterns into parts. Bower and Glass showed subjects a set of abstract line drawings and then asked them to reproduce these drawings given fragments of the drawings as memory-retrieval cues. Fragments that corresponded to "good" parts according to Gestalt principles, which were therefore hypothesized to be explicitly represented as parts in a hierarchical representation of the visual pattern, were more effective in cueing memory than were other equally large and complex fragments.

Palmer (1977) gathered converging evidence that certain portions of abstract geometric patterns were explicitly represented as parts, for example by asking subjects to divide the patterns into their natural parts or rate the goodness of parts. He then showed that portions of a pattern that appeared to be explicitly represented as parts according to these criteria were also more easily verified as coming from their whole patterns than portions that were not. A similar finding was obtained by Reed (1974), although the aim of his research was not to elucidate the part structure of patterns but, rather, the information available in mental images. Reed found that subjects were able to verify the presence of pattern fragments in their mental images of the whole pattern only when the fragments corresponded to "good" parts.

The research of Bower and Glass (1976), Palmer (1977), and Reed (1974) suggests that when a portion of a stimulus pattern is explicitly represented as a part in the subject's representation of the pattern, it will be better recognized as having come from that pattern than if it is not explicitly represented as a part. The experiments to be reported here make use of this finding as a way of testing the *parthood* and *objecthood* of visual stimuli. If a portion of a stimulus is represented as a part in the visual

representation of the stimulus that underlies recognition, it should be identified more accurately than if it does not have the status of a part in the stimulus representation.

In each of the three experiments to be reported, subjects learn to recognize a set of normal faces and a set of some contrasting class of stimuli: scrambled faces (Experiment 1), inverted faces (Experiment 2), and houses (Experiment 3). Subjects are trained so that they are at least as accurate at recognizing the normal faces as the contrasting stimuli. We can then compare the identification of isolated features from normal faces with the identification of isolated features from the contrasting classes of stimuli. For face stimuli, the tested parts were the facial features of the eyes, nose, and mouth. These facial features are not only the nameable parts of a face but also correspond to the natural parsings of a face based on the discontinuities of its contours (Biederman, 1987; Hoffman & Richards, 1984). If face recognition is more holistic than the recognition of other kinds of stimuli, then identification of isolated features from the normal faces should be disproportionately less accurate than identification of isolated features from the contrasting stimulus classes, relative to the identification of the part in the whole face and whole contrast object.

EXPERIMENT 1

In this experiment, subjects were asked to memorize intact and scrambled faces. Scrambled faces were chosen as a contrasting stimulus class because their parts are the same as the parts of a normal face, and yet we would not expect special, face-specific recognition abilities to be used in recognizing scrambled faces. After learning the normal and scrambled faces, subjects were given a forced-choice recognition task in which they identified facial features presented in isolation and in whole-face context. The whole-face test items were constructed such that the target and foil faces differed only with respect to the feature being tested. Examples of these two types of test for intact and scrambled faces are shown in Figure 1. In the isolated part test condition, subjects would be asked to identify "Larry's nose". In the full-face test condition, subjects would be asked to identify "Larry". Note that the only difference between the "Larry" target and foil in the whole face test is the nose feature. That is, the information available for making the discrimination was exactly the same: the face outline, hair, eyes, and mouth were held constant.

If the recognition of normal faces involves representing their component parts to the same degree as the recognition of scrambled faces, then we should expect that identification of the features of normal faces will be just as good relative to the identification of the whole face as identification of the features of scrambled faces are relative to the identification of whole scrambled faces. However, if normal faces are recognized more holistically

than scrambled faces, then there should be a disadvantage for identifying isolated features compared to whole faces for normal faces, relative to part and whole test performance for scrambled faces.

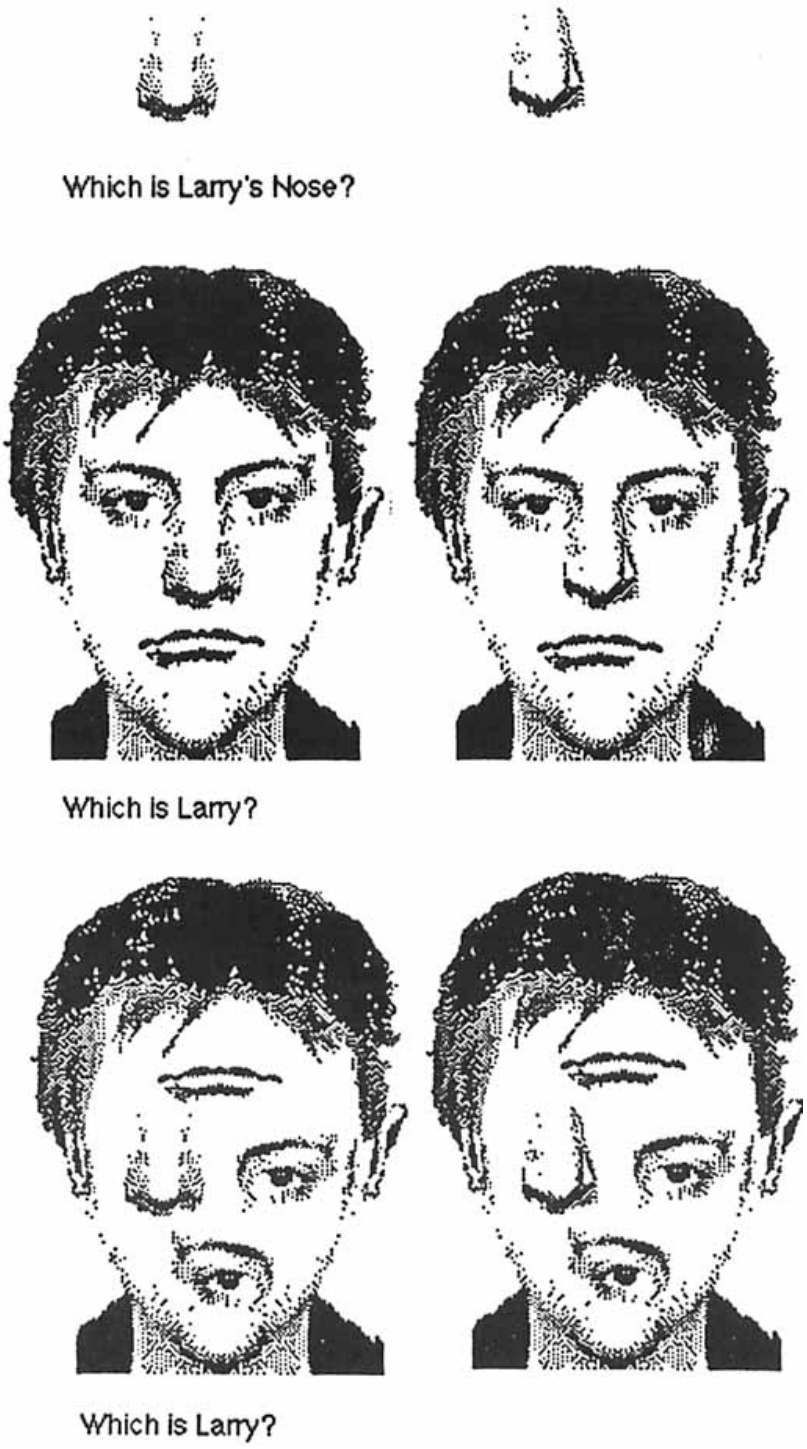


FIG. 1. Example of isolated part, intact face, and scrambled face test items.

Method

Subjects

Twenty first-year psychology students from Carnegie-Mellon University served as subjects in the experiment. Subjects were tested individually and received course credit for their participation.

Materials

Stimuli consisted of two groups of six male faces that were generated on a Macintosh computer using a Mac-a-Mug program. Faces were constructed by selecting one of the three exemplars for each of the three feature types (e.g. eyes, nose, mouth).[2] The exemplars for one group of faces are shown in Figure 2. For both groups, exemplars were placed within the same face outline. Face stimuli were constructed such that no one exemplar was unique to a particular face, with each exemplar present in two of the six faces in the group. Scrambled and intact versions of each face were generated. For the scrambled faces, the spatial positions of the features were consistent across faces (e.g. the nose was always located below and to the left of the mouth). Half of the subjects saw one group of faces as the scrambled set and the other group as the intact set. For the other half of subjects, the versions of the face groups were reversed. Thus, each face appeared an equal number of times in its scrambled and intact version. Faces were photocopied onto 4″ × 5″ white card stock.

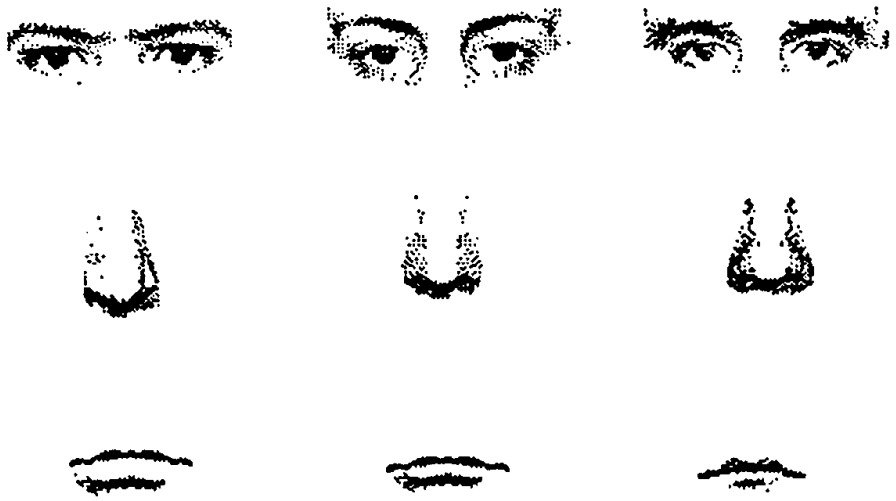


FIG. 2. Eye, nose, and mouth exemplars used for one group of faces in Experiment 1.

[2]We will reserve the term "feature" to refer to a discrete part of a face (e.g. eyes, nose, and mouth) and the term "exemplar" to refer to a particular instance of a feature (e.g. long nose).

Procedure

Subjects were seated at a table directly facing the experimenter at a viewing distance of approximately 2 m. Subjects were informed that they would be shown pictures of scrambled and intact faces paired with male names and their task was to learn the correct face–name associations.

Learning Phase. Learning and test trials were blocked according to face version (scrambled and intact). For each learning trial, the face stimulus was randomly presented for 5 sec accompanied by its verbally spoken name. Each learning block contained six learning trials, one trial per face. There were a total of five learning blocks per face version.

Test Phase. Immediately following learning, a two-choice recognition test was administered. One feature from each of the learned faces was included in the recognition test. An equal number of eyes, nose, and mouth features were tested. In the isolated part test condition, subjects identified isolated features of the learned faces (e.g. which is Bob's nose?). Item foils were taken from one of the other learned faces. In the full-face test condition, subjects were shown the same target features and their foils presented in the full-face configuration (scrambled or intact) and asked to identify the face that matched the given name (e.g. which is Bob?). The target and foil faces differed only with respect to the individual feature that was tested in the isolated test condition; all other feature information was held constant. The full-face foil did not correspond to any previously learned face. Thus, subjects identified each feature twice, once presented in isolation and once presented in the full face (intact or scrambled face). After the learning and test phases for one of the face versions (scrambled or intact) was completed, subjects learned and were tested on the other version with the initial face version counterbalanced across subjects.

Results and Discussion

As shown in Figure 3, subjects were able to identify isolated parts from intact faces correctly on 62% of the trials. When the same parts were tested in the whole face, performance improved to 73%. For scrambled faces, there was a different pattern of results. Subjects were actually better at identifying the parts tested in isolation (71% correct) than tested in the whole face (64% correct). An analysis of variance (ANOVA) with face version (intact and scrambled), test type (isolated part and whole face), and facial feature (eyes, nose, mouth) as within-subjects factors confirmed this interaction between face version and test type, $F(1, 19) = 7.55$, $p < 0.02$. Direct comparisons between part- and whole-face performance showed that the part–whole difference was reliable for normal, intact faces,

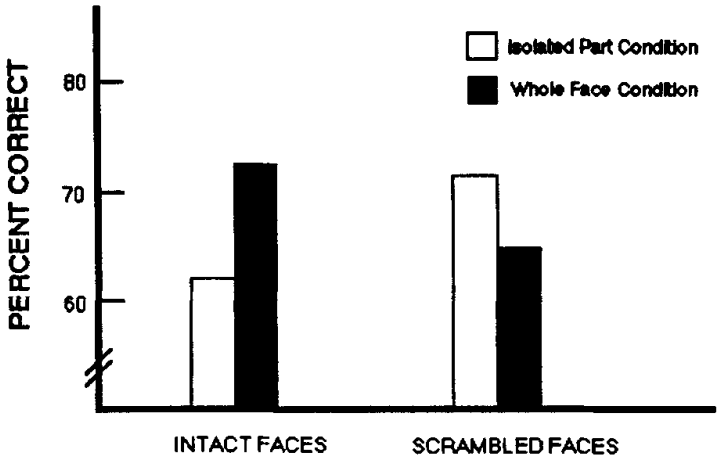


FIG. 3. Percentage of correctly identified isolated part- and whole-face test items for intact and scrambled faces.

$t(19) = 2.16$, $p < 0.05$, but not for scrambled faces, $t(19) = 1.25$. The advantage of whole-face recognition for normal, intact faces over scrambled faces suggests that the normal face is mentally represented more in terms of a whole object (holistically) as compared to the representation of a scrambled face, which is more in terms of its parts (featurally).

The main effect of facial feature was also reliable, $F(2, 38) = 7.96$, $p < 0.01$. Subjects were more accurate making eye judgments (80% correct) than they were making nose judgments (62% correct) or mouth judgments (63% correct). This finding is consistent with previous results involving simultaneous matching tasks (Sergent, 1984; Walker-Smith, 1978) in which eye features were perceptually more discriminable than nose or mouth features. No other main effects or interactions were reliable, $p > 0.10$.

The results of Experiment 1 indicate that subjects are better at identifying facial features from normal faces when they are presented in the whole face than when they are presented alone, relative to recognition of facial features from scrambled faces when presented as isolated parts and wholes. This is true despite the fact that the whole-face test items had no more discriminating information in them than did the isolated parts: for each choice between isolated parts, the corresponding whole-face test items differed only by those same parts. This outcome is consistent with the hypothesis that normal faces are recognized more holistically than are scrambled faces. Note that this result can be interpreted in either of two ways. It can be argued that part representations are less available for normal faces relative to scrambled faces or that holistic representations are more available for normal faces relative to scrambled faces. Direct comparisons of part- and whole-face recognition performance for intact and scrambled faces suggests that the latter interpretation might be more

accurate. Although difference in part recognition for intact and scrambled faces was not reliable, $t(19) = 1.52$, whole intact face recognition performance was reliably better than whole scrambled face recognition, $t(19) = 2.07$, $p < 0.05$. Thus, the recognition of intact faces differs from the recognition of scrambled faces primarily in engaging holistic representations.

EXPERIMENT 2

One could argue that scrambled faces are too unnatural to provide an appropriate comparison for the processing of normal faces. Perhaps scrambled objects in general would be more likely to be represented featurally than normal objects. If so, one could not conclude from the previous experiment that face recognition is particularly holistic. For this reason, we turned to different contrasting stimulus set, inverted faces. Inversion disproportionately impairs the recognition of faces more than it does the recognition of other types of objects, such as airplanes, buildings, or costumes (Yin, 1969). These effects appear to be fairly robust, and results have been obtained for a variety of face stimuli including famous and novel faces (Scapinello & Yarmey, 1970; Yarmey, 1971), simple line drawn faces (Yin, 1969), photographs of faces (Carey & Diamond, 1977; Diamond & Carey, 1986) in different experimental paradigms, including forced-choice recognition (Yin, 1969) and "old" versus "new" judgments (Valentine & Bruce, 1986). The face inversion effect has been taken to index the operation of specialized face recognition mechanisms not normally used for recognizing other kinds of objects (e.g. Carey & Diamond, 1977; Yin, 1969). Thus, inverted faces provide a contrasting stimulus set that includes the same parts, in the same relative configuration, as normal faces but does not engage the hypothesized face-specific recognition mechanisms. In this experiment, subjects learned the face–name associations for six upright faces and six inverted faces. In test, subjects were asked to identify both the individual features of the learned upright or inverted faces presented in isolation and whole upright and inverted faces. If upright features are recognized using more holistic representations than inverted faces, then subjects should be more accurate at recognizing upright features contained in a whole face than in isolation, relative to whole face and isolated part recognition of inverted features.

Method

Subjects

Twenty first-year psychology students from Carnegie-Mellon University served as subjects in the experiment. Subjects were tested individually and received course credit for their participation.

Materials

The same two groups of face stimuli used in the previous experiment were used in this experiment. Two versions of each set were prepared: one in its normal upright orientation, and one inverted by 180°. Instead of presenting the stimuli on cards, they were presented on a Macintosh computer screen. The test items were presented in the same orientation as the study items.

Procedure

During the learning phase of the experiment, subjects learned the name–face associations for six upright (inverted) faces presented on a Macintosh computer. Faces and their assigned names were blocked according to face orientation. One learning block consisted of six learning trials, one trial per face, and there were five learning blocks in total. Learning was self-paced. Immediately following learning, a two-choice recognition test was administered. In contrast to Experiment 1, isolated part- and whole-face test items were randomly presented with the restriction that features from the same face were separated by at least two test trials, and the same feature type (e.g. nose feature) was not tested on consecutive trials. Also different from Experiment 1, the eyes, nose, and mouth features from each face were tested in the isolated part- and whole-face test conditions. Presentation of the test items was initiated by the subject, and test items were displayed until a response was made. Responses were recorded by computer. After the learning and test phases were completed for faces in one orientation (upright or inverted), the faces in the other orientation were learned and tested. Half of the subjects learned one group of six faces in the upright orientation and the other six faces in the inverted orientation. For the other half of the subjects, the face groups and their orientation was reversed. Learning and test phases were blocked according to face orientation, and presentation order of the face orientation was counterbalanced across subjects.

Results and Discussion

As shown in Figure 4, recognition of inverted, whole faces and inverted parts was roughly equivalent, 65% accuracy for whole face and 64% accuracy for parts, respectively. However, for upright faces, whole face stimuli were better recognized than part face stimuli. That is, subjects correctly identified 74% of the whole-face stimuli as compared to 65% of the part-face stimuli. The recognition advantage found for whole upright faces, but not for whole inverted faces, relative to the recognition for their parts, is consistent with the interpretation that holistic processing is used for upright

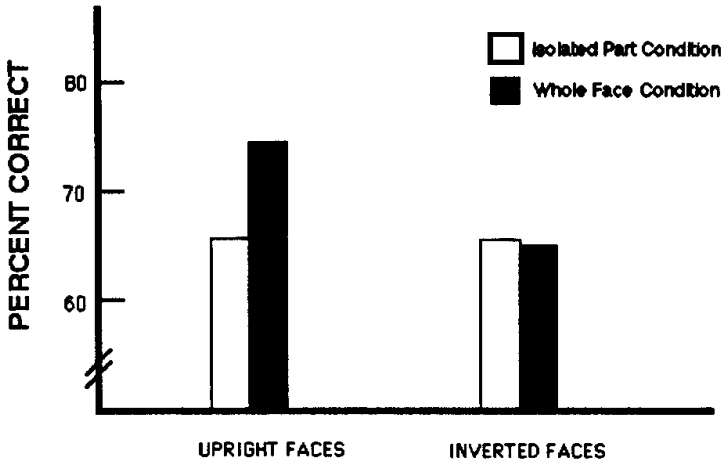


FIG. 4. Percentage of correctly identified isolated part- and whole-face test items for upright and inverted faces.

faces. An ANOVA with face orientation (upright and inverted), test type (isolated part and whole face), face feature (eyes, nose, mouth) as within-subjects factors and order as a between-subjects factor confirmed the reliable Face Orientation × Test Type interaction, $F(1, 18) = 8.92, p < 0.01$. Consistent with the results of Experiment 1, the direct comparison between isolated part and whole face recognition for upright faces was again reliable, $t(19) = 3.41$, $p < 0.01$. Further, whereas there was little difference in performance between recognition of isolated parts of upright and inverted faces, whole upright faces were reliably better recognized than whole inverted faces, $t(19) = 2.94, p < 0.01$, again suggesting that normal upright faces are recognized more holistically relative to inverted faces.

The main effect of test type was also reliable, $F(1, 18) = 8.47, p < 0.01$, indicating that overall, whole-face stimuli—either whole upright faces or whole inverted faces—were better recognized than were part-face stimuli. Consistent with the results of Experiment 1, the main effect of face feature was reliable, $F(2, 38) = 8.47, p < 0.001$, such that eye features were better recognized (76% correct) than nose features (64% correct) or mouth features (63% correct). However the relative saliency of the face features was affected by the orientation of the face as indicated by the reliable Orientation × Face Feature interaction, $F(2, 38) = 3.88, p < 0.05$. No other main effects or interactions were reliable ($p > 0.10$).

In Experiment 2, we found that subjects were poorer at recognizing the parts of upright faces when presented in isolation than they were at recognizing the whole face, even though they showed no disadvantage for parts over wholes when the same faces were inverted. As in Experiment 1, the part disadvantage for upright faces was observed despite the fact that the

same discriminating information was available in both the part and whole test items for all types of stimuli: whichever pair of feature exemplars was presented in the forced-choice test of part identification, the corresponding pair of whole stimuli differed only by those features. In a related study, Young, Hellawell, and Hay (1987) found that inversion improved recognition of the top or bottom halves of composite faces, which they attributed to the disruption of configural processes. Taken together, these results suggest that the face representations affected by inversion are relatively holistic representations. Given that inversion is more disruptive to face recognition than to the recognition of other kinds of stimuli, this supports the hypothesis that face recognition involves more holistic representations than the recognition of other stimuli.

EXPERIMENT 3

In Experiments 1 and 2 it was found that intact, upright faces were encoded more holistically than scrambled faces or inverted faces. It is possible that these results do not reflect anything special about face recognition per se, but only demonstrate holistic processing for the recognition of coherent, upright objects. The purpose of the present experiment is to contrast face recognition with the recognition of normal upright stimuli other than faces. Houses have been used as the contrast stimuli to faces in other studies (Bruce et al., 1991; Valentine & Bruce, 1986; Yin, 1969) and seemed particularly suited to goals of our research for several reasons. Like faces, houses have internal features (i.e. doors and windows) that share an overall configuration. Also like faces, the parts of a house can be varied independently of each other without disrupting the house schema. Finally, house stimuli can be constructed such that the number of house features corresponds to the number of face features. As shown in Figure 5, the house stimuli used in Experiment 3 had three features—a door, a large window, and two small windows—analogous to the mouth, nose, and eyes of a face. If holistic processing is not restricted to faces, then a disadvantage should also be evident for house parts relative to whole-house stimuli. On the other hand, if the use of holistic representations is a particular characteristic of face processing, houses should not show the relative part disadvantage.

Method

Subjects

Twenty first-year psychology students from Carnegie-Mellon University served as subjects in the experiment. Subjects were tested individually and received course credit for their participation.

Materials

House stimuli were generated on a Macintosh computer using an architectural design software package. As shown in Figure 5, similar to the faces, houses were constructed by selecting one of the three feature values for each of the three feature types (e.g. door, big window, small window). The six stimulus houses were created according to the exemplars specified by the face stimuli.

Procedure

The procedure was similar to the one used in Experiment 2. Subjects were informed that they would see a house (face) picture accompanied by a name, and their task was to learn the name–picture association. In the case of the houses, subjects were told that the name corresponded to the

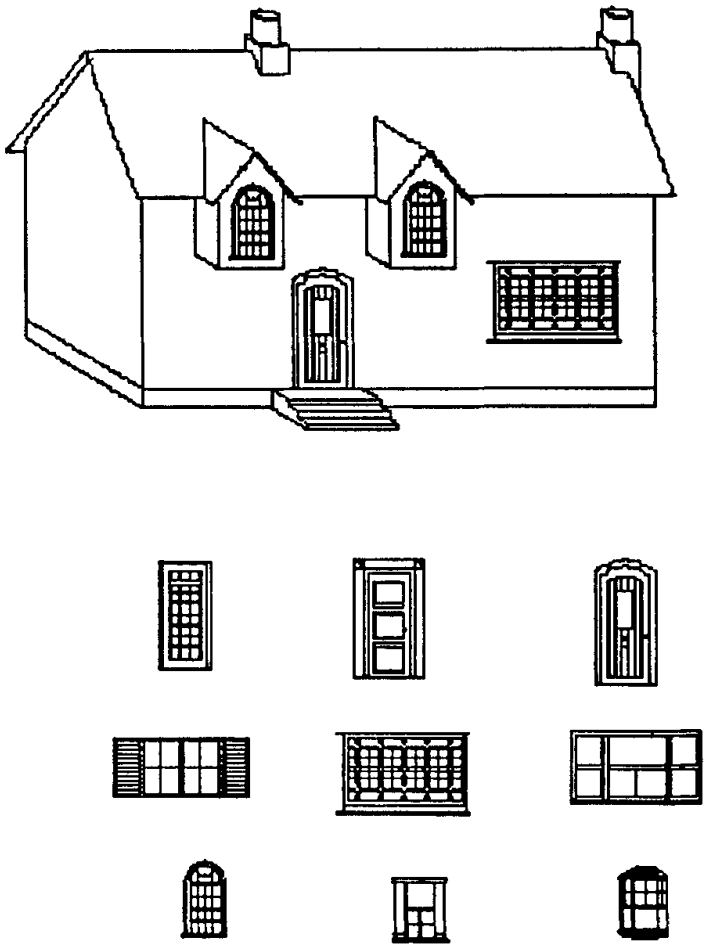


FIG. 5. Door, large window, small window exemplars and a sample stimulus house used in Experiment 3.

person who lived in the house. A learning block consisted of six learning trials, one trial per house (face) picture, and pictures were randomly presented on a Macintosh computer. There were five learning blocks per object type. After the learning phase was completed, recognition memory for the part shown in isolation and embedded in the whole object was randomly tested in a forced-choice paradigm. The same item order restrictions described in Experiment 2 were used. Whole-object foils (house and face) were constructed such that they were distinct from any previously learned object. Recognition memory was tested for the three house features (i.e. door, small window, big window) and three face features (eyes, nose, mouth) presented in isolation and in the whole-object conditions. Learning and test were blocked according to object type (house and face). The order of the object type presented for learning and test was counterbalanced across subjects.

Results and Discussion

As shown in Figure 6, whereas only 65% of the face features were recognized in isolation, recognition improved to 77% when the same features were shown in the whole-face context. This finding replicates the holistic effect found for faces demonstrated in the previous two experiments. In contrast, recognition of the house features was roughly equivalent in the isolated and whole-house test condition, 81% and 79% correct, respectively. Thus, unlike faces, no advantage was found for identifying house features as part of their whole object. An ANOVA with object type (houses and faces) and test type (isolated part and whole face) as within-subjects factors and

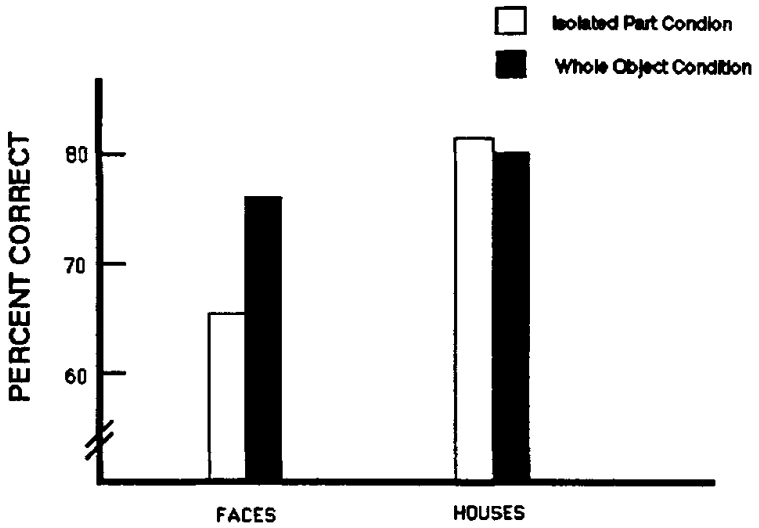


FIG. 6. Percentage of correctly identified isolated part- and whole-object test items for faces and houses.

order as a between-subjects factor revealed a reliable Object Type × Test Type interaction, $F(1, 18) = 17.47$, $p < 0.001$, as predicted. A direct comparison also showed that facial features were more readily recognized in the whole-face condition than in the isolated part condition, $t(19) = 4.46$, $p < 0.01$. The main factor of object type was also reliable, $F(1, 18) = 9.20$, $p < 0.01$, indicating that houses were recognized more accurately than were faces. A reliable effect was also found for test type, $F(1, 18) = 9.11$, $p < 0.01$; however, this effect should be interpreted as the result of its higher-order interaction with object type. Finally, the effect of order was also reliable, $F(1, 18) = 4.41$, $p < 0.05$, but order did not interact with any other factor. No other interactions were reliable, $p > 0.10$.

In comparing recognition for different types of objects, it is difficult to equate the relative discriminability of features—in this case, face features and house features. However, the focus of the present study was not on comparing part recognition across object types, only in comparing parts and wholes recognition within an object type. In this regard, we found an advantage for the recognition of the wholes of faces relative to the isolated face part, but found no difference between part and whole recognition for houses. Furthermore, the possibility that a difference in part discrimination across object types is, in some indirect way, responsible for a difference in reliance in part versus whole recognition cannot explain the results of Experiments 1 and 2 in which the part features were the same. Thus, the main finding of Experiment 3 is consistent with the claim that face recognition is different from the recognition of other objects, such as houses, in its relatively greater reliance on holistic representations.

GENERAL DISCUSSION

In these experiments we tested the hypothesis that face recognition is relatively more dependent on holistic representations than the recognition of other types of stimuli. By *holistic representation* we mean one without an internal part structure. Following other researchers, we reasoned that if a portion of an object corresponds to an explicitly represented part in a hierarchical visual representation, then when that portion is presented in isolation it will be identified relatively more easily than if it did not have the status of an explicitly represented part. The hypothesis that face recognition is holistic therefore predicts that the isolated parts of a face will be disproportionately more difficult to recognize than the whole face, relative to recognition of the parts and wholes of other kinds of stimuli. This prediction was borne out in three experiments: subjects were less accurate at identifying the parts of faces, presented in isolation, than they were at identifying whole faces, even though both parts and wholes were tested in a forced-choice format and the whole faces differed only by one part. In

contrast, three other types of stimuli—scrambled faces, inverted faces, and houses—did not show this disadvantage for part identification.

At first glance, these results are reminiscent of the *face superiority effect*, according to which the parts of a face are better perceived if presented in the context of a whole face than in the context of a scrambled face (e.g. Homa, Haver, & Schwartz, 1976; Mermelstein, Banks, & Prinzmetal, 1979). The two phenomena are indeed similar in that both reflect the influence of representations of wholes on subjects' performance. However, they are distinct phenomena, differing from each other in several ways. (1) The face superiority effect comes into play only under conditions of threshold vision, suggesting that its locus is in the visual encoding of facial features, not their access to stored memory representations. In contrast, our task did not tax visual encoding, but taxed memory access. (2) As Pomerantz (1981) has noted, in face and object superiority effects the perception of a part in context is as good as, but not better than, recognition of just the isolated part. Performance with the whole face is superior only to performance with a scrambled face. In contrast, we found that recognition of whole faces was better than recognition of isolated parts. (3) The face superiority effect does not appear to be specific to faces but is a more general phenomenon involving the visual encoding of parts in context, alongside the word superiority effect (Reicher, 1969; Wheeler, 1970) and object superiority effects for geometric forms (Enns & Gilani, 1988; Weisstein & Harris, 1974), and chairs (Davidoff & Donnelly, 1990). In contrast, the present results with faces were not found with the other types of tested stimuli.

How do these findings relate to the idea that face recognition is particularly dependent on "configuration"? If by a configurational representation we mean one in which the spatial relations among the parts of a face are as important as the shapes of the individual parts themselves (Haig, 1984; Hosie, Ellis, & Haig, 1988), then we would suggest that the concepts of configurational representation and holistic representation are highly similar, and possibly identical. The shapes of the individual parts are essentially within-part spatial relations. In the limiting case of configurational representation, in which between-part spatial relations are as precisely specified as the within-part relations, parts have lost much, if not all, of their special status. Presumably, for this reason, the terms *holistic* and *configurational* have often been used interchangeably in the face recognition literature.

Recent findings in neurophysiology and neuropsychology seem consistent with our conclusions regarding the relatively holistic representation of faces. It has been demonstrated that a subpopulation of neurons located in the superior temporal sulcus of the monkey responds selectively to the sight of face parts and whole faces (e.g. Desimone, Albright, Gross, &

Bruce, 1984; Perrett, Mistlin, & Chitty, 1987) and display some ability to discriminate among different faces (Baylis, Rolls, & Leonard, 1985). Although the responses of these neurons to a face are not greatly diminished by deleting a feature, they are abolished if all features are present but scrambled (Desimone et al., 1984), consistent with holistic rather than featural representation. Although many interpretations of this fact of anatomy are possible, it is at least consistent with the notion that face representations are relatively holistic.

The fact that the temporal cortex of monkeys also contains cells responsive to individual facial features, especially eyes, has been taken by some to indicate that faces are represented hierarchically, with explicitly represented component parts (Perrett et al., 1987). However, Desimone (1991) has raised the possibility that the "feature" cells may not be representing facial features per se. For example, a cell that responds to an eye in isolation might respond to any dark spot on a white background. Furthermore, it appears that the functional role of many of the "eye" cells may be to represent direction of eye gaze, an important form of social interaction among monkeys (Perrett et al., 1985). In our view, a critical test of the hierarchy hypothesis for interpreting the role of "feature" cells in face processing would be to verify that their latencies of response are, on average, shorter than the latencies of "face" cells. This test has not yet been carried out (Perrett, personal communication).

Human neuropsychology is also consistent with the hypothesis of relatively holistic face recognition. Brain-damaged patients may be impaired at face recognition, object recognition, or printed word recognition. In analysing the patterns of co-occurrence among these impairments, Farah (1991) found that two possible combinations of these impairments did not occur: object recognition impairments without either face or word impairments, and both face and word impairments without some degree of object impairment. This suggested the existence of two, rather than three, underlying representational capacities responsible for the recognition of faces, objects, and words, which are used in complementary ways: one that is essential for face recognition, needed to a lesser extent for the recognition of common objects and not needed at all for printed word recognition, and one that is essential for printed word recognition, needed to a lesser extent for the recognition of common objects, and not needed at all for face recognition. The representational capacity lacking in patients with impairments in printed word recognition appears to be the ability to represent multiple explicitly represented structural units (e.g. letters in a word; see Farah & Wallace, 1991, for a review of the evidence). The ability to represent shape holistically would seem a good candidate for a complementary representational capacity, and the neuropsychological evidence suggests that this capacity is particularly taxed by face recognition.

REFERENCES

Baylis, G.C., Rolls, E.T., & Leonard, C.M. (1985). Selectivity between faces in the responses of a population of neurons in the cortex of the superior temporal sulcus of the monkey. *Brain Research, 342*, 91–102.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Bower, G.H., & Glass, A.L. (1976). Structural units and the redintegrative power of picture fragments. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 456–466.

Bradshaw, J.L., & Wallace, G. (1971). Models for the processing and identification of faces. *Perception and Psychophysics, 9*, 443–448.

Bruce, V. (1988). *Recognizing faces*. Hove: Lawrence Erlbaum Associates Ltd.

Bruce, V., Doyle, T., Dench, N., & Burton, M. (1991). Remembering facial configurations. *Cognition, 38*, 109–144.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science, 195*, 312–314.

Davidoff, J., & Donnelly, N. (1990). Object superiority: A comparison of complete and part probes. *Acta Psychologica, 73*, 225–243.

Desimone, R. (1991). Face selective cells in temporal cortex of monkeys. *Journal of Cognitive Neuroscience, 3*, 1–8.

Desimone, R., Albright, T.D., Gross, C.G., & Bruce, C.J. (1984). Stimulus-selective responses of inferior temporal neurons in the macaque. *Journal of Neuroscience, 4*, 2051–2068.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115*, 107–117.

Enns, J.T., & Gilani, A.B. (1988). Three-dimensionality and discriminability in the object-superiority effect. *Perception and Psychophysics, 44*, 243–256.

Farah, M.J. (1991). Patterns of co-occurrence among the associative agnosias: Implications for visual object representation. *Cognitive Neuropsychology, 8*, 1–19.

Farah, M.J., & Wallace, M.A. (1991). Pure alexia as a visual impairment: A reconsideration. *Cognitive Neuropsychology, 8*, 313–334.

Galton, F. (1879). Composite portraits, made by combining those of many different persons into a single, resultant figure. *Journal of the Anthropological Institute, 8*, 132–144.

Haig, N.D. (1984). The effect of feature displacement on face recognition. *Perception, 13*, 104–109.

Hoffman, D.D., & Richards, W.A. (1984). Parts of recognition. In S. Pinker (Ed.), *Visual Cognition*. Cambridge, MA: MIT Press.

Homa, D., Haver, B., & Schwartz, T. (1976). Perceptibility of schematic face stimuli: Evidence for a perceptual Gestalt. *Memory and Cognition, 4*, 176–185.

Hosie, J.A., Ellis, H.D., & Haig, N.D. (1988). The effect of feature displacement on the perception of well-known faces. *Perception, 17*, 461–474.

Matthews, M.L. (1978). Discrimination of Identikit construction of faces: Evidence for a dual processing strategy. *Perception and Psychophysics, 23*, 153–161.

Mermelstein, R., Banks, W., & Prinzmetal, W. (1979). Figural goodness effects in perception and memory. *Perception and Psychophysics, 26*, 472–480.

Palmer, S.E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology, 9*, 441–474.

Perrett, D.I., Mistlin, A.J., & Chitty, A.J. (1987). Visual neurones responsive to faces. *Trends in Neuroscience, 10*, 358–364.

Perrett, D.I., Smith, P.A.J., Potter, D.D., Mistlin, A.J., Head, A.S., Milner, A.D., & Jeeves, M.A. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London, Series B, 223*, 293–317.

Pomerantz, J.R. (1981). Perceptual organization in information processing. In M. Kubovy & J.R. Pomerantz (Eds.), *Perceptual organization* (pp. 141–180). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Reed, S.K. (1974). Structural descriptions and the limitations of visual images. *Memory & Cognition, 2*, 329–336.

Reicher, G.M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 275–280.

Scapinello, K.F., & Yarmey, A.D. (1970). The role of familiarity and orientation in immediate and delayed recognition of pictorial stimuli. *Psychonomic Science, 21*, 329–330.

Sergent, J. (1984). An investigation into component and configural processes underlying face perception. *The British Journal of Psychology, 75*, 221–242.

Smith, E.E., & Nielsen, G.D. (1970). Representations and retrieval processes in short-term memory: Recognition and recall of faces. *Journal of Experimental Psychology, 85*, 397–405.

Valentine, T., & Bruce, V. (1986). The effect of race, inversion and encoding activity upon face recognition. *Acta Psychologica, 61*, 259–273.

Walker-Smith, G.J. (1978). The effects of delay and exposure duration in a face recognition task. *Perception, 6*, 63–70.

Weisstein, N., & Harris, C.S. (1974). Visual detection of line segments: An object-superiority effect. *Science, 186*, 752–755.

Wheeler, D.D. (1970). Process in word identification. *Cognitive Psychology, 1*, 59–85.

Yarmey, A.D. (1971). Recognition memory for familiar "public" faces: Effects of orientation and delay. *Psychonomic Science, 24*, 286–288.

Yim, R.K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*, 141–145.

Young, A.W., Hellawell, D., & Hay, D.C. (1987). Configuration information in face of perception. *Perception, 16*, 747–759.

Configural information in facial expression perception

Article *in* Journal of Experimental Psychology Human Perception & Performance · May 2000

DOI: 10.1037/0096-1523.26.2.527 · Source: PubMed

4 authors, including:

Michael P Dean
University College London
16 PUBLICATIONS **649** CITATIONS

SEE PROFILE

Journal of Experimental Psychology:
Human Perception and Performance
2000, Vol. 26, No. 2, 527–551

0096-1523/00/$5.00 DOI: 10.1037//0096-1523.26.2.527

Configural Information in Facial Expression Perception

Andrew J. Calder
Medical Research Council Cognition and Brain Sciences Unit

Andrew W. Young
University of York

Jill Keane
Medical Research Council Cognition and Brain Sciences Unit

Michael Dean
University of Sheffield

Composite facial expressions were prepared by aligning the top half of one expression (e.g., anger) with the bottom half of another (e.g., happiness). Experiment 1 shows that participants are slower to identify the expression in either half of these composite images relative to a "noncomposite" control condition in which the 2 halves are misaligned. This parallels the composite effect for facial identity (A. W. Young, D. Hellawell, & D. C. Hay, 1987), and like its identity counterpart, the effect is disrupted by inverting the stimuli (Experiment 2). Experiment 3 shows that no composite effect is found when the top and bottom sections contain different models' faces posing the same expression; this serves to exclude many nonconfigural interpretations of the composite effect (e.g., that composites are more "attention-grabbing" than noncomposites). Finally, Experiment 4 demonstrates that the composite effects for identity and expression operate independently of one another.

Bruce and Young's (1986) functional model of face recognition postulates separate parallel routes for the processing of facial identity (who the person is) and facial expression (what they are feeling). Over the years, this dissociation has been investigated by a number of studies using a range of different methodologies. These include cognitive studies of neurologically normal participants (Campbell, Brooks, de Haan, & Roberts, 1996; Young, McWeeny, Hay, & Ellis, 1986), double dissociations in brain-injured participants (Parry, Young, Saul, & Moss, 1991; Young, Newcombe, de Haan, Small, & Hay, 1993), single-cell recording in nonhuman primates (Hasselmo, Rolls, & Baylis, 1989), and, in more recent years, functional imaging studies of brain activation (George et al., 1993; Sergent, Ohta, MacDonald, & Zuck, 1994). Together, these studies provide substantial support for the idea that facial identity and facial expression recognition are dissociable cognitive functions, and this is perhaps one of the reasons why these two facial attributes have so often been the topics of separate examination. But their isolated investigation is possibly less to do with their proposed functional independence and more to do with the fact that traditionally, facial identity and facial expression processing have been studied within separate domains of psychology.

In general, facial expression recognition has been studied within a social psychology framework, where research has focused on the communicative value of signals of facial affect rather than their perceptual representation. Studies of facial identity processing, however, have been heavily influenced by research in cognitive psychology, and consequently, a firm emphasis has been placed on understanding the perceptual mechanisms involved. In the last 20 years, then, there has been an enrichment in our understanding of the perceptual representation of facial identity, whereas the perceptual mechanisms underlying facial expression recognition have not been so extensively investigated. Hence, although the work of Ekman and his colleagues has greatly enhanced our understanding of the anatomy used to produce facial expressions, knowledge of the perceptual processes needed to decode them remains scant.

Recent research has aimed to redress this imbalance (Calder, Young, Perrett, Etcoff, & Rowland, 1996; Calder, Young, Rowland, & Perrett, 1997; Ellison & Massaro, 1997; Etcoff & Magee, 1992; Young et al., 1997) by taking two approaches. First, these studies have built on the strong knowledge base provided by the social psychology literature, and second, they have applied perceptual paradigms developed within other areas of psychology to the study of facial affect processing. This latter approach has the added advantage of using tried and tested methods, and for the reasons outlined above, the facial identity literature provides a particularly good source of perceptual paradigms. Examples of these include the following: effects of stimulus orientation (Diamond & Carey, 1986; Farah, Tanaka, & Drain, 1995; Valentine, 1988), feature displacement (Haig, 1984), distinctiveness effects (Rhodes, Brennan, & Carey,

Andrew J. Calder and Jill Keane, Medical Research Council Cognition and Brain Sciences Unit, Cambridge, England; Andrew W. Young, Department of Psychology, University of York, Heslington, England; Michael Dean, Department of Psychology, University of Sheffield, Sheffield, England.

We are grateful to P. Ekman for giving us permission to use pictures from the Ekman and Friesen (1976) Pictures of Facial Affect series. We would also like to thank Brian Cox and Gary Jobe for their assistance in preparing figures.

Correspondence concerning this article should be addressed to Andrew J. Calder, Medical Research Council Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, England. Electronic mail may be sent to andy.calder@mrc-cbu.cam.ac.uk.

1987; Valentine, 1991), and image negation (Bruce & Langton, 1994; Hill & Bruce, 1996), all of which have provided valuable clues to how facial identity is coded. But perhaps the most consistent result to emerge from the facial identity literature is the important role of configural information in face recognition (Bruce, Doyle, Dench, & Burton, 1991; Carey & Diamond, 1977; Rhodes, 1988; Tanaka & Farah, 1993; Young, Hellawell, & Hay, 1987). It is highly pertinent, then, for us to investigate what role, if any, configural information may play in facial expression recognition.

Carey and Diamond (1977) introduced the term *configural information* to mean the interrelationship between different facial features (e.g., the relative shape and positioning of the mouth in relation to the shape and positioning of the nose, eyes, etc.); this type of facial information is seen as distinct from the structure and shape of individual features (e.g., eye, nose, mouth shape, etc.). Diamond and Carey (1986) identified two forms of configural information that they referred to as *first-order* and *second-order relational properties.* The former type refers to the raw inter feature relationships that are common to all normal faces—two horizontally positioned eyes, above a central nose, above a central mouth, etc.; effectively the spatial information that makes a face a face. Second-order relational properties are substantially more subtle and are what are more generally referred to as simply *configural features.* These features are the interrelationships between different feature positions and shapes that help distinguish one facial identity from all others (e.g., the distance between the eyes, position and shape of the nose in relation to the position and shape of the mouth, etc.).

The current consensus in facial identity research is that configural features are particularly important for face recognition; however, individual features may also contribute to some extent. We refer to this view as the configural model. Here, we investigate its applicability to the perception of facial signals of emotion. It worth mentioning that Tanaka and Farah (1993) have distinguished the configural model from their holistic model of face processing. For this latter model, it is proposed that faces are coded as Gestalt representations in which the constituent parts (eyes, nose, mouth, etc.) are not "explicitly represented." In support of their model, Tanaka and Farah showed that a single facial feature (eyes, nose, or mouth) is more readily identified as belonging to a particular person's face when it is shown in the context of the whole face, than when shown in isolation. The same was not shown to be true, however, of scrambled faces, inverted faces, or a set of structurally homogeneous houses (made up of doors and windows in place of facial features).

Recently, Ellison and Massaro (1997) have shown that Tanaka and Farah's (1993) holistic model is not applicable to facial affect recognition (see below). Instead, they suggest that their data are consistent with the antithesis of this model, one in which facial expressions are represented and identified in terms of their individual parts, or features (e.g., eye, nose, and mouth shape, etc.)—what we refer to as the part-based model.

Ellison and Massaro (1997) used facial expressions displayed on a synthetic (computer-generated) face in which just two facial features, the eyebrows and the corners of the mouth, were manipulated. The stimuli were produced by combining five levels of eyebrow displacement (ranging between eyebrows raised and eyebrows flattened) and five levels of mouth displacement (ranging between corners of the mouth turned up, and corners of the mouth turned down). *Prototype* expressions of happiness and anger were defined as eyebrows maximally raised with mouth corners maximally curled up, and eyebrows maximally flattened with mouth corners maximally curled down, respectively. All other combinations of the five mouth and five eyebrow displacements were generated to give a total of 25 full-face images. In addition, the five levels of eyebrow and five levels of mouth features were presented individually in the context of the upper and lower sections of the face, respectively. The participants' task was to decide whether each image signaled a happy or an angry expression.

By modeling their data using Massaro and colleague's fuzzy logical model of perception (FLMP; Massaro, 1998; Massaro & Cohen, 1990), Ellison and Massaro (1997) showed that participants' responses to the whole-face images could be reliably predicted from their responses to the half-face images. Consequently, they argued that their results were inconsistent with the holistic model (as defined by Tanaka and Farah, 1993). However, they pointed out that although their results provided no direct support for the configural model, they did not rule out the idea of configural encoding of facial affect altogether. Instead, they suggested that if configural features are used in the representation and recognition of facial expressions, their results demonstrated that they are unlikely to involve the spatial relationships between the features manipulated in their stimuli (eyebrows and mouth corners). But it is also worth considering that Ellison and Massaro may have failed to find evidence of configural processing because of the particular design and stimuli they used.

For example, Ellison and Massaro (1997) used facial expressions that were generated on a single synthetic face in which only the eyebrows and mouth corners were manipulated. Under these circumstances, the participants may have been able to treat these two altered features as separate objects, basing their decisions on their individual shapes rather than a more global impression of the face. It is also worth noting that manipulating one facial feature in a human face can often have secondary consequences for other features. For instance, changing the positions of the eyebrows can cause the brow to become wrinkled or furrowed, and manipulating the shape of the mouth can affect the shape of the cheeks. The fact that these more global changes were not present in the synthetic expressions used by Ellison and Massaro may also have served to minimize the configural encoding of these images.

In addition, the idea that configural information is important for facial expression recognition is not completely unfounded. In an investigation of the Thatcher illusion, Parks, Coss, and Coss (1985) found that the judged pleasant-

ness of upright and inverted smiling mouths was affected by two factors: (a) the location of the eyes in relation to the mouth (above or below), and (b) the distance between the eyes and the mouth; pleasantness ratings of the eyes showed a strikingly parallel pattern. Hence, even though the participants were being asked to rate just one facial feature (eyes or mouth), the configuration of the face influenced their judgment of the feature. In a separate study, Wallbott and Ricci-Bitti (1993) presented participants with single muscular movements (action units) in the context of an otherwise neutral face, and combinations of action units. The participants task was to rate the emotional intensity of the resultant expressions on seven scales (Happiness, Sadness, Anger, Fear, Disgust, Surprise, and Contempt). Wallbott and Ricci-Bitti found that the meaning of most single action units changes when presented in combination with other action units, and only a few action units transmit a specific emotional meaning that is retained across different contexts. Again, these results point to a role of configural processing in facial affect recognition, a role that Ekman and Friesen (1975) also identified, although not empirically, in their book *Unmasking the Face:* "With many facial expressions a change in just one area gives the impression that the rest of the facial features have changed as well" (p. 39).

Given the above observations, we felt that it was possible that evidence of configural processing of emotional facial expressions might be found using a different design to one used by Ellison and Massaro (1997).

The Composite Paradigm

Earlier we mentioned that contemporary facial expression research is in the fortunate position of being able to borrow tried and tested methodologies from the facial identity literature. Consequently, we felt that the most direct method of distinguishing between configural and part-based models of facial expression recognition was to adopt a paradigm that has been described by Bruce (1988) as "[a] compelling illustration of the power of configural processing of faces" (p. 41), the facial composite phenomenon originally shown by Young et al. (1987).

The composite effect shows that when the top half of one face is aligned with the bottom half of another's, the two halves fuse to create a perceptually "new" (composite) face (Figure 1). Consequently, people are significantly slower to name the top or bottom segments of these composite faces relative to a control condition in which the two halves are *misaligned* (noncomposite condition; Figure 1) so that they do not form a face shape. Young et al. (1987) suggested that this effect can be explained in terms of the important role that configural features play in facial identity recognition. In the composite condition, the top and bottom halves of two different faces align to form a novel configuration, and this interferes with the recognition of the identity shown in either of the two halves; that is, the novel configuration does not match the configural information for either the top or bottom identity. Misaligning the two halves, however, means that the image is no longer encoded as a configural whole, and the separate parts of the face can be accessed without interference from an inappropriate configuration. In a second experiment Young et al. (1987) bolstered this interpretation by showing that the composite effect is abolished when the stimuli are inverted (i.e., rotated by 180°; see also Carey & Diamond, 1994). This second finding is consistent with Carey and Diamond's (1977) earlier observation that configural information is more difficult to encode from inverted faces.

The advantage of the composite paradigm is that the same physical features (i.e., the top and bottom sections of the face) are present in both conditions (composite and noncomposite). The only difference between the two conditions is whether the two halves are aligned, to form a face, or misaligned, so that they do not. Consequently, if responses are slower for the composite condition, this demonstrates that the composite images are being processed differently to the noncomposites. In facial identity research, a number of investigators concur with Young et al.'s (1987) idea that

Figure 1. The composite effect shown by Young, Hellawell, and Hay (1987). The top half of one face is aligned with the bottom half of another's to create a "new" facial identity (composite). Young et al. (1987) showed that the top and bottom segments of faces are easier to identify in the misaligned (noncomposite) condition than in the aligned (composite) condition.

slower reaction times (RTs) for the composite condition can be attributed to a disruption of configural encoding (Bruce, 1988; Carey & Diamond, 1994; Endo, Masame, & Maruyama, 1989; Endo, Takahashi, & Maruyama, 1984; Hole, 1994). The composite effect, then, seems a highly appropriate paradigm to distinguish between configural and part-based models of facial expression processing.

Interestingly, historical research shows that Young et al. (1987) were not the first to use composite faces. They had originally been used some 60 years earlier for facial expression research (Dunlap, 1927). Here, they were not used to examine configural processing, however, but rather the relative contribution of the upper and lower face regions in expression recognition. For example, in one experiment, Dunlap presented his participants with frames containing four faces; two of the faces were posing different prototype expressions (selected from the list *natural, amusement, mirth, startle, expectation, pain, disgust, grief, strain,* and *relaxation*), and two were composite facial expressions prepared by combining the top half of one prototype with the bottom half of the other. For each composite expression, the participants were asked to decide which of the two prototype expressions it resembled most. The results showed that on 80% of the trials, participants selected the prototype that corresponded to the bottom half of the composite. Consequently, Dunlap concluded that the bottom region of the face is more important for facial expression recognition.

Since Dunlap (1927), other studies have addressed the issue of upper versus lower face dominance in emotion recognition, and the majority of these were reviewed by Ekman, Friesen, and Ellsworth (1972). Ekman et al. discussed the fact that Dunlap's findings proved difficult to replicate (Coleman, 1949; Frois-Wittmann, 1930), and subsequent investigations of this issue generally have found that the emotion is more readily recognizable from the upper face region for some facial expressions and the lower face region for others (Bassili, 1979; Hanawalt, 1944; Plutchik, 1962). Hence, these studies suggest that there are what we refer to as facial expressions with a *recognizable-top* or *recognizable-bottom* half.

For our own purpose of investigating a composite effect for facial expression, the results of these latter studies are highly relevant. This is because the participants' task in the composite paradigm is to identify the expressions in one half (top or bottom) of the composite or noncomposite images. Hence, by using composites prepared from the top segments of recognizable-top expressions and the bottom sections of recognizable-bottom expressions, we could ensure that the task was readily accomplishable.

The facial expressions used in this study were taken from Ekman and Friesen's (1976) pictures of facial affect series. This stimulus set is especially important because it is well validated, on the basis of exact anatomical criteria, and has been extensively used in other studies. The set contains pictures of facial expressions associated with six basic emotions (happiness, sadness, anger, fear, disgust, and surprise) posed by a number of different models. Ekman and his colleagues have shown that each emotion is associated with distinct facial musculatures that are recognized by a number of cultures throughout the world (Ekman, 1972; Ekman et al., 1987). As far as we are aware, there have been no attempts to determine which of the Ekman and Friesen faces can be identified from their top or bottom sections. Consequently, we conducted a preliminary experiment (described in Experiment 1) that identified that anger, fear, and sadness were more readily recognized from the top half of the face, whereas happiness and disgust were more recognizable from the bottom half; surprise was found to be equally recognizable from both top and bottom sections.

On the basis of this information, composite expressions were prepared for Experiment 1 by aligning the top half of a recognizable-top expression (e.g., anger) with the bottom half of a recognizable-bottom expression (e.g., happiness) posed by the same model. As a comparison condition, we used noncomposite images; these were identical to the composites except that the top and bottom halves were misaligned horizontally. Following Young et al. (1987), we reasoned that support for a configural model of facial expression recognition could be found if participants were slower to identify the top (or bottom) half of an expression when it was shown as part of a composite (face-like) image, relative to when it was presented as part of a noncomposite (non–face-like) image. If, on the other hand, configural information is relatively unimportant for facial expression identification (part-based model), then no significant difference should be found between the composite and noncomposite conditions. This would occur because if facial expression recognition is based largely on the analysis of individual features, then aligning or misaligning the top and bottom face halves should have little affect on the participants' ability to identify the emotion.

In Experiment 2, we studied the effect of stimulus inversion on the composite phenomenon for facial expression. As we have already noted, Young et al. (1987) found that the composite effect for identity was disrupted by inverting the stimuli. It was clearly of interest, then, whether a composite effect for expression would be similarly affected.

Having demonstrated a composite effect for facial expression in Experiments 1 and 2, in Experiment 3 we addressed the criticism that the longer RTs for the composite condition could be attributed to the composite images appearing somehow more "attention grabbing" than the noncomposites. This might occur for a number of reasons; for example, the join between the top and bottom face halves can produce abrupt changes in texture and unnatural contours in the middle of nose and cheeks, causing the face to look slightly unusual in appearance. A method of addressing this issue presented itself during the preparation of the composites.

While making the stimuli, we noted Young et al.'s (1987) original effect that aligning the top and bottom halves of two peoples' faces generates a perceptually new face (see also Hole, 1994). However, we also noted a second interesting phenomenon. When the two face halves are taken from two identities posing the same facial expression (e.g., happiness), the resultant composite expression is also readily identifiable as happiness. This suggested an interesting

prediction: that composite faces prepared from two identities posing the same expression (same-expression composites) should not show the composite effect for facial expression. Confirmation of this prediction would demonstrate that the composite effect observed in Experiments 1 and 2 cannot be attributed to some form of inherent quality of composite faces that causes them to produce slower response times (e.g., as a result of abrupt discontinuities in texture, etc.). On the other hand, a significant composite effect for the same-expression composites would question the idea that the composite paradigm taps configural processing. It was important, then, to address this issue.

Finally, Experiment 4 examined whether configural information for facial identity and facial expression recognition can be disrupted independently of one other. This was done by comparing participants' RTs to report the expression or identity shown in bottom half of composite faces containing the same or different expressions and same or different identities in the two facial halves.

Experiment 1

The first section of this experiment aimed to identify which of the expressions in the Ekman and Friesen (1976) series are identifiable from their top or bottom halves. This information was then used to create the composite and noncomposite images for Experiment 1.

Recognition Rates for Top and Bottom Sections of the Ekman and Friesen (1976) Faces

Method

Participants. Eight members of the MRC Cognition and Brain Sciences Unit subject panel (6 women, 2 men) participated in the experiment for payment. The participants were between the ages 21 and 40 years and had normal or corrected-to-normal vision.

Materials. The stimuli were prepared from gray-scale pictures from the Ekman and Friesen (1976) pictures of facial affect. Pictures of 10 people's faces (6 women, 4 men) were used, each posing one example of six facial expressions (happiness, sadness, anger, fear, disgust, and surprise). These 10 models were selected because a reliably recognized example of the six expressions was available for each. Each of these 60 pictures of facial expressions was divided into top and bottom segments. This was done by cutting each face along a horizontal line through the bridge of the nose. Examples of the stimuli are shown in Figure 2.

Design and procedure. One within-subjects factor, stimulus format (whole face, top segment, and bottom segment), was investigated. Participants saw the 60 faces (10 identities posing six

Figure 2. From Experiment 1, examples of the whole-face, top-segment, and bottom-segment stimuli. One example of six facial expressions (happiness, sadness, anger, fear, disgust, and surprise) posed by six models from the Ekman and Friesen (1976) series is shown in each of the three stimulus formats (whole face, top segment, and bottom segment). Images from *Pictures of Facial Affect*, by P. Ekman and W. V. Friesen, 1976. Copyright 1976 by P. Ekman and W. V. Friesen. Adapted with permission.

facial expressions) in each of these three stimulus formats. The 180 different stimuli were presented individually in random order on a 256 gray-scale computer screen. The top-segment images were presented in the location corresponding to the top half of the whole face and bottom-segment images in the corresponding bottom half location (see Figure 2). Each image subtended a horizontal visual angle of approximately 4.6°. The participant was asked to identify the emotion expressed in each image. Responses were made using a box with six labeled buttons (one for each emotion category, happiness, sadness, anger, fear, disgust, and surprise); the position of the emotion labels was counterbalanced across participants. The button box was interfaced with a Macintosh Power PC computer to record the participant's choice of emotion label and decision time.

On each trial, the image remained in view until the participant responded, and consecutive trials were separated by an interval of approximately 2.5 s. Participants were asked to respond quickly and accurately. After all 180 images had been presented, there was a short break, and then the same procedure was repeated in a second block.

To familiarize participants with the experimental format, the experiment began with 12 practice trials. These trials contained pictures of additional models from the Ekman and Friesen (1976) series posing the same six emotional expressions listed above in whole-face, top-segment, and bottom-segment formats. These practice faces were not seen in the main experimental trials.

Results

Participants' mean error proportions and mean correct RTs to identify the emotion displayed in the whole-face, top-segment, and bottom-segment images are listed in Table 1 by emotion category. Standard errors are shown in brackets.

Table 1
Data from Experiment 1

	Face format					
	Whole		Top		Bottom	
Emotion	*M*	*SE*	*M*	*SE*	*M*	*SE*
	Error proportions					
Anger	.22	.08	.28	.06	.49	.09
Fear	.25	.07	.33	.08	.56	.09
Sadness	.09	.03	.19	.05	.34	.08
Happiness	.01	.01	.20	.09	.01	.01
Disgust	.14	.05	.62	.10	.14	.04
Surprise	.21	.07	.21	.06	.33	.07
	Reaction times (in milliseconds)					
Anger	1,910	130	1,963	118	2,380	271
Fear	2,041	132	2,043	125	2,210	237
Sadness	1,742	202	1,803	142	2,400	332
Happiness	1,178	103	1,394	74	1,119	113
Disgust	1,738	258	2,320	206	1,413	124
Surprise	1,748	177	1,847	245	1,949	156

Note. Participants' error proportions and mean correct reaction times to identify the emotion displayed in examples of six facial expressions (happiness, sadness, anger, fear, disgust, and surprise). The faces were presented in three formats: whole face, top segment, and bottom segment.

Error rates. Our principal form of analysis involved error rates in identifying the emotions, because we were interested in determining which of the expressions could be identified accurately from their top halves (recognizable-top expressions), and which could be identified accurately from their bottom halves (recognizable-bottom expressions). Error proportions were arcsin transformed and submitted to two analysis of variance (ANOVAs), one by participants (F_1) the other by items (F_2). Two factors were investigated: face format (whole face, top segment, and bottom segment; repeated measure) and emotion (happiness, sadness, anger, fear, disgust, and surprise; repeated measure). Both analyses showed a significant effect of face format, $F_1(2, 14) = 30.44$, $p < .0001$, and $F_2(2, 18) = 11.94$, $p < .0005$. Post hoc t tests ($p < .05$) of the two analyses showed the same pattern; overall, the emotions were more accurately identified from the whole-face images than from the top or bottom segments, which did not reliably differ. This main effect was qualified by a significant interaction between emotion and face format, $F_1(10, 70) = 11.46$, $p < .0001$, and $F_2(10, 90) = 9.89$, $p < .0001$. Simple effects analyses by participants (F_1) and by items (F_2) showed significant effects of face format for all emotions except surprise. The F values of these simple effects analyses are listed by emotion category in the following section, and where appropriate, a summary of post hoc t tests ($p < .05$) of the simple effect is shown in brackets (note: in each case the post hoc effects were identical for the analyses by participants and by items): anger, $F_1(2, 14) = 10.07$, $p < .005$, and $F_2(2, 18) = 6.43$, $p < .01$ ([whole = top] < bottom); fear, $F_1(2, 14) = 27.10$, $p < .001$, and $F_2(2, 18) = 6.25$, $p < .01$ ([whole = top] < bottom); sadness, $F_1(2, 14) = 7.13$, $p < .01$, and $F_2(2, 18) = 19.28$, $p < .001$ ([whole = top] < bottom); happiness, $F_1(2, 14) = 5.77$, $p < .02$, and $F_2(2, 18) = 10.64$, $p < .001$ ([whole = bottom] < top); disgust, $F_1(2, 14) = 29.77$, $p < .001$, and $F_2(2, 18) = 33.12$, $p < .001$ ([whole = bottom] < top); and surprise, $F_1(2, 14) = 3.41$, $p > .05$, and $F_2(2, 18) = 1.97$, $p > .2$. Finally, both analyses also showed significant effects of emotion, $F_1(5, 35) = 7.94$, $p < .0001$, and $F_2(5, 45) = 7.46$, $p < .0001$. Post hoc t tests ($p < .05$) showed that, overall, happiness was more accurately recognized than the other emotions.

In summary, the results of the error rates analysis show that anger, fear, and sadness were more recognizable from the top half of the face (recognizable-top expressions), whereas happiness and disgust were more recognizable from the bottom half of the face (recognizable-bottom expressions). Surprise was equally recognizable from its top and bottom sections.

These results essentially replicate those of Bassili (1979), who examined the same six facial expressions, although his images were not taken from the Ekman and Friesen (1976) series, and they were animated. The only difference between Bassili's findings and our own is that Bassili's sadness expressions were equally recognizable from their whole, top, and bottom segments, whereas we found that the bottom segments of the Ekman and Friesen sadness expressions

were less accurately recognized than their whole or top segments, which did not reliably differ.

RTs. Two subsidiary analyses (one by participants, F_1, the other by items, F_2) were carried out on the RT data to check that the more accurate responses were not accompanied by slower RTs. Again, the factors investigated were face format (whole face, top segment, and bottom segment; repeated measure) and emotion (happiness, sadness, anger, fear, disgust, and surprise; repeated measure). Neither analysis showed a significant effect of face format, but both showed a significant interaction between emotion and face format, $F_1(10, 70) = 3.78$, $p < .0005$, and $F_2(10, 90) = 3.12$, $p < .005$. Simple effects analyses by participants (F_1) and by items (F_2) showed a significant effect of face format for happiness, $F_1(2, 14) = 7.15$, $p < .01$, and $F_2(2, 18) = 6.46$, $p < .01$; and disgust, $F_1(2, 14) = 8.06$, $p < .005$, and $F_2(2, 18) = 5.76$, $p < .05$, only. Post hoc t tests ($p < .05$) showed that participants were significantly slower to identify the happiness and disgust emotions from the top segment of the face; RTs to identify these emotions from the bottom-segment and whole-face images did not reliably differ. Thus, there was no evidence of participants trading accuracy for speed.

In summary, this preliminary study identified that anger, fear, and sadness are readily identified from the top section of the face (recognizable-top expressions), whereas happiness and disgust are readily identifiable from the bottom half of the face (recognizable-bottom expressions). Because surprise could be recognized from either part of the face, we used it as a recognizable-bottom expression to even up the number of expressions in each condition of our design.

In the next section of the experiment, we created composite facial expressions composed of the top halves of the recognizable-top expressions and bottom halves of the recognizable-bottom expressions (e.g., top = anger, bottom = happiness). These images allowed us to test whether a similar phenomenon to the composite effect for facial identity (Young et al., 1987) could be found with facial expressions. Following Young et al.'s reasoning, if configural information is important for facial expression recognition, then participants should be slower to identify the top or bottom half of a facial expression when it is presented as part of a composite image than when it is shown as part of a noncomposite (misaligned) image.

Identifying the Top and Bottom Sections of Composite and Noncomposite Expression Images

Method

Participants. Twelve people (9 women, 3 men) aged between 21 and 40 years and from the same population as the previous section participated in the experiment. None had taken part in the previous section.

Materials. The stimuli were prepared from pictures of four female models from the Ekman and Friesen (1976) series (C, NR, PF, and SW), each posing one example of the expressions happiness, sadness, anger, fear, disgust, and surprise; these pictures were selected from the stimuli used in the previous experiment. Composite and noncomposite stimuli comparable with the facial identity composites used by Young et al. (1987) were then prepared from these facial expressions. Their preparation is described below.

Composites. Composite facial expressions were prepared by aligning the top segment of a recognizable-top expression (e.g., anger) with the bottom segment of a recognizable-bottom expression (e.g., happiness) posed by the same model. For each of the four models, all nine possible combinations of these recognizable-top and recognizable-bottom segments were prepared; these combinations were as follows: anger–happiness, anger–disgust, anger–surprise, fear–happiness, fear–disgust, fear–surprise, sadness–happiness, sadness–disgust, and sadness–surprise (the first emotion of each pair indicates the top half of the composite). This gave a total of 36 composite faces.

Noncomposites. The noncomposite facial expressions were essentially identical to the composites except that the top and bottom segments were misaligned horizontally. This was done by aligning the middle of the nose in the top segment with the edge of the face in the bottom segment. For half of the images, the top segment was shifted to the left of the bottom segment, and for the other half, this positioning was reversed (see Figure 3). Note that when the noncomposites were presented in the center of the computer screen, neither the bottom or top half of the image was centralized in the screen. To allow for this fact, half of the composite stimuli were presented in the same position as the left section of the noncomposites and half in the same location as their right section (see Figure 3); positioning was counterbalanced across stimuli. This method of presentation follows the basic procedure used by Young et al. (1987).

Examples of composite and noncomposite facial expressions prepared from pictures of one of the models used in Experiment 2 are shown in Figure 3.

Design and procedure. Two within-subjects factors were investigated: stimulus type (composite and noncomposite) and task instructions ("identify the top-half expression" and "identify the bottom-half expression"). The experiment began with a block in which each of the 24 whole (prototype) facial expressions (four models, each posing six facial expressions) were presented individually in random order. The participant's task was to identify the emotion displayed in each face by pressing one of six buttons marked with the emotion labels *happiness, sadness, anger, fear, disgust,* and *surprise;* the position of these labels was counterbalanced across participants. Each face was preceded by a fixation cross for 500 ms followed by a blank interval of the same duration. The face remained in view until the participant responded, with their response initiating the next trial after an interval of approximately 2.5 s. All images were displayed on a 22-in. gray-scale computer screen using a Macintosh Power PC. The purpose of this block of trials was to familiarize the labeling task, to ensure that accuracies in the main experimental trials were sufficiently high to allow meaningful measurement of RTs.

Participants then completed two blocks of experimental trials. In one block, they were asked to identify the expression displayed in the top segment of the composite and noncomposite images (top-segment block) and in a second block the expression shown in the bottom segment of these same images (bottom-segment block); half of the participants did the bottom-segment block trials first. The general design of these two blocks was the same, so we only give a detailed description of the bottom-segment block.

The bottom-segment block began with a single presentation of the bottom segment of each of the three facial expressions (happiness, disgust, and surprise) posed by the four models (C, NR, PF, and SW); the presentation times were as for the whole-face

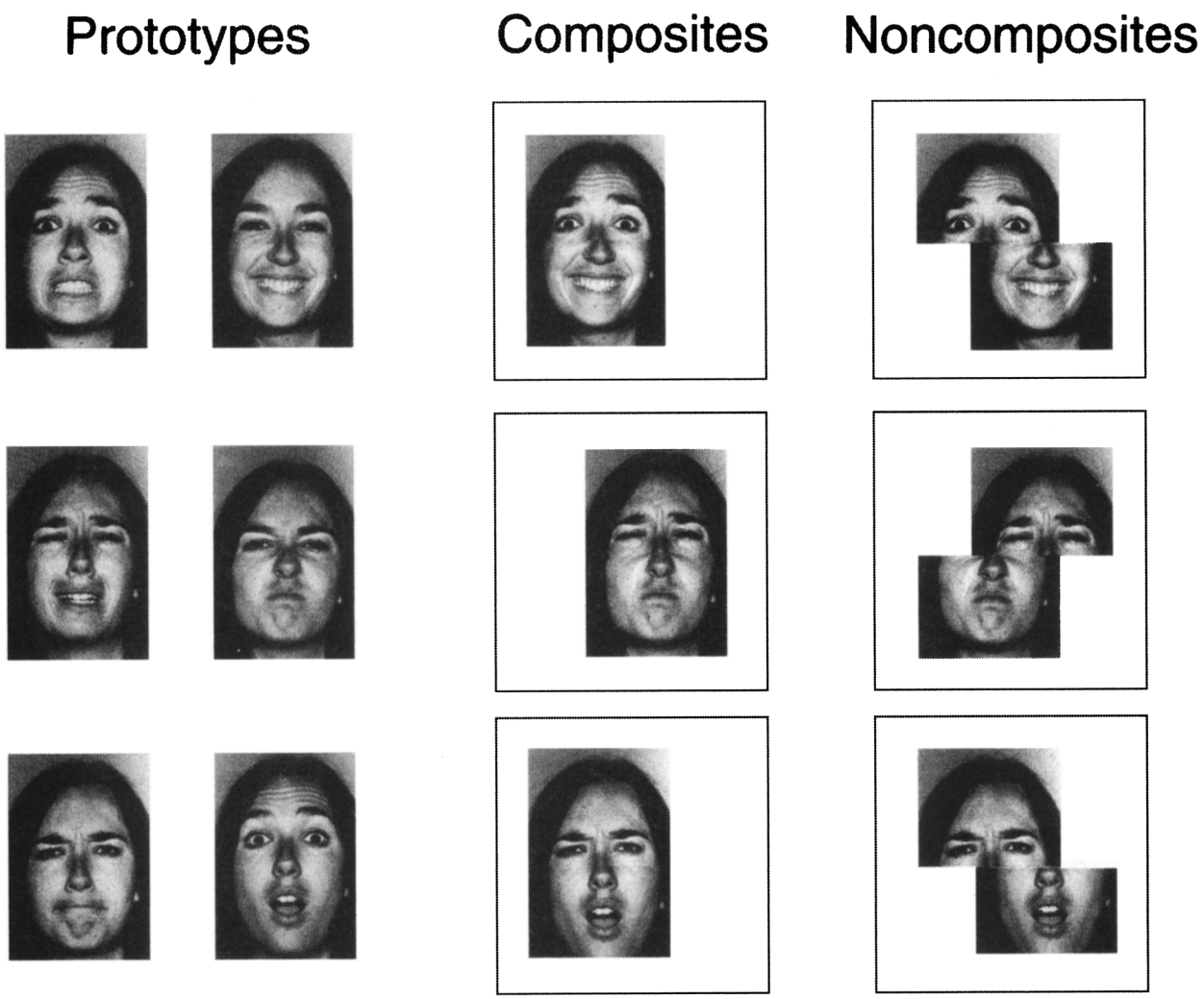


Figure 3. Examples of stimuli used in Experiment 1. The top and bottom segments of recognizable-top and recognizable-bottom prototype expressions (left), respectively, were combined to create composite (middle) and noncomposite (right) stimuli. The two face sections of each composite and noncomposite image were from pictures of the same model (Model C in the example shown). Images from *Pictures of Facial Affect,* by P. Ekman and W. V. Friesen, 1976. Copyright 1976 by P. Ekman and W. V. Friesen. Adapted with permission.

presentations described above. Participants were asked to make an identification decision by pressing one of three buttons labeled *happiness, disgust,* and *surprise.* Following this, the experiment proper began. This included one presentation of each of the 36 composite and 36 noncomposite stimuli described above. The images were presented in random order, and the participant was asked to identify the expression displayed in the bottom segment of the images as quickly and accurately as possible by pressing one of the three labeled keys. Again, the presentation times were identical to the whole-face presentations described earlier. To familiarize the participants with the composite and noncomposite images, the experiment was preceded by 10 practice trials selected at random from the 72 experimental trials. The composite images subtended a horizontal visual angle of approximately 4.6°, and for the noncomposites, a horizontal visual angle was approximately 5.7°; the vertical visual angle for both was approximately 6.3°.

For the top-segment block, the design was virtually identical. However, this time the block began with one presentation of the top segment of the facial expressions anger, fear, and sadness posed by the same four models. The participants were then presented with the same composite and noncomposite images seen in the bottom-segment block, but this time they were asked to identify the expression displayed in the top segment of the face. In both sections of the top-segment block, participants made their response by pressing one of three keys labeled *anger, fear,* and *sadness.*

Results

Participants' mean correct RTs (with standard error bars) to identify the top and bottom halves of the composite and noncomposite facial expressions are shown in the left graph

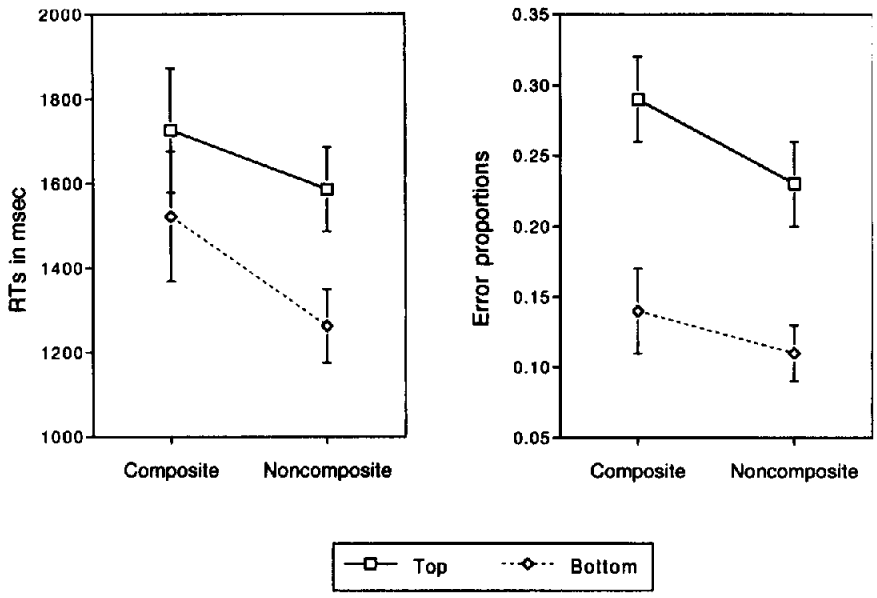


Figure 4. Data from Experiment 1. The left graph shows participants' mean correct reaction times (RTs; with standard error bars) to identify the expression displayed in top and bottom halves of the composite and noncomposite stimuli. The right graph shows participants' mean error proportions (with standard error bars) from the same experiment.

of Figure 4. The right graph shows participants' mean error proportions (with standard error bars) for the same experiment.

RTs. Our principal form of analysis involved RTs for correct responses. These were submitted to a two-factor ANOVA investigating stimulus type (composite and noncomposite; repeated measure) and task instructions ("identify top-half expression" and "identify bottom-half expression"; repeated measure). There was a significant effect of stimulus type, $F(1, 11) = 6.35, p < .05$, indicating that participants found it harder to identify the top and bottom segments of the images in the composite condition. There was also a significant effect of task instructions, $F(1, 11) = 17.26, p < .005$, demonstrating that, overall, participants were faster to recognize the expression shown in the bottom half of the images. There was no significant interaction between these two factors.

Error rates. A subsidiary analysis examined error rates to ensure that the slower RTs to the composite images were not accompanied by increased accuracy. Error proportions were arcsin transformed and submitted to a two-factor ANOVA investigating stimulus type (composite and noncomposite; repeated measure) and task instructions ("identify top-half expression" and "identify bottom-half expression"; repeated measure). This showed a significant main effect of task instructions, $F(1, 11) = 23.02, p < .001$, reflecting that, overall, participants were significantly more accurate at identifying the expressions shown in the bottom half of the images. There was also a marginally significant effect of stimulus type, $F(1, 11) = 4.65, .1 > p > .05$, but no significant interaction between these two factors. Thus, there was no evidence of participants trading accuracy for speed.

Discussion

The results of Experiment 1 support the configural model of facial expression recognition over the part-based model. Participants were significantly slower, and marginally less accurate, at identifying the expression shown in half of the composite images than the noncomposite images. Moreover, the effect was equally strong when they were asked to identify the top half of the images as when they were asked to identify their bottom half. This composite effect for facial expression is all the more striking when we consider that the participants were only asked to discriminate among three different facial expressions in each of the top-segment and bottom-segment blocks. Hence, although strategies were readily available to the participants (e.g., if the mouth is open wide, the bottom-half expression must be surprise, or if the eyes are wide open, the top-half expression must be fear), they did not, or were not able to, make full use of them. In this sense, these findings essentially parallel those found for facial identity (Carey & Diamond, 1994; Young et al., 1987), and a similar explanation can be invoked. Following Young et al.'s reasoning, we suggest that facial expressions are processed in terms of their configural make-up; that is, the shape and position of the mouth in anger may be coded relative to the shape and position of other features in the expression (e.g., furrowed brow, close-set eyebrows, etc.). Hence, when the top and bottom segments of different facial expressions are aligned, they fuse to form a perceptually new facial expression configuration that interferes with the processing of the constituent parts of the top and bottom sections. This effect can be seen in the examples shown in Figure 3. The top row shows the top half of a fear expression

combined with the bottom half of a happiness expression. The result is a wild expression that could not really be accurately described as happiness or fear.

It is important to emphasize, however, that a composite effect for facial affect does not mean that the individual features of facial expressions are not also encoded for identification. It simply implies that the configural relationship of the features plays a significant role in the encoding of facial expression.

Recall that Ellison and Massaro (1997) found that their data could be reliably modeled by the FLMP if one assumed that the information in the upper and lower sections of the face were evaluated independently and then integrated to produce an overall degree of support for a particular emotion category (e.g., happiness). Our own data do not concur with this finding. In Experiment 1, the same face halves were present in the composite and noncomposite conditions. Hence, if the face halves were being processed independently of one another, we would predict that the RTs for the two conditions should not significantly differ. However, this was not found: the participants responses were significantly slower for the composite condition. In other words, aligning the face halves to produce a facial image has a significant effect on the speed with which the participants can perform the task. For the present, then, we note there is a disagreement between Ellison and Massaro's results and our own, and in the General Discussion section we address possible explanations.

It is also worth emphasising that our results cannot simply be attributed to a Stroop (1935) interference effect between the different conceptual (or semantic) information conveyed by the top and bottom face halves. This is because the same halves are present in both composite and noncomposite conditions. Hence, although a Stroop effect between emotion concepts may operate in both experimental conditions, it can not be the source of the increased RTs found for the composite condition.

Experiment 2

As we discussed in the introduction, Young et al. (1987) found that the composite effect for facial identity was lost when the stimuli were inverted (see also Carey & Diamond, 1994, and Hole, 1994). This is consistent with Carey and Diamond's (1977) suggestion that configural information for identity is more difficult to process in inverted than upright faces. Therefore, in Experiment 2, we investigated the effect of stimulus inversion on the composite effect for facial expression. We reasoned that if configural processing constitutes the basis of the effect we have observed, then the composite effect for expression should be significantly disrupted when the stimuli are inverted.

In Experiment 1, the participants were asked to identify both top and bottom sections of the composite and noncomposite images, and no difference in the pattern of findings was noted across "identify-top" and "identify-bottom" conditions—both showed an equivalent composite effect. For Experiment 2, therefore, we arbitrarily selected the bottom section of the images for the participants to identify in both upright and inverted conditions.

Method

Participants. Twelve people (6 women, 6 men) aged between 19 and 45 years and from the same population as Experiments 1 and 2 participated in the experiment. All had normal or corrected-to-normal vision, and none had taken part in Experiments 1 and 2.

Materials. The stimuli were identical to those used in Experiment 1.

Design and procedure. In the previous experiment, participants identified the expression shown in the top half of the composite and noncomposite images in one block and the expression shown in the bottom half in a second block. In Experiment 2, participants were only asked to identify the expression shown in the bottom half of these same stimuli, but under two conditions: (a) when the stimuli were presented upright and (b) when the same stimuli were inverted. Hence, in the inverted condition, the bottom half of the face was effectively the top half of the image.

The beginning of the experiment was identical to Experiment 1; participants were presented with the 24 original whole-face images and asked to categorize each with one of six emotion labels (happiness, sadness, anger, fear, disgust, and surprise). Following this, half of the participants were assigned first to the upright condition and half to the inverted condition. The upright condition block was identical to the bottom-segment block described in Experiment 1. Hence, participants were first presented with the bottom segments of the expressions happiness, disgust, and surprise posed by four models and asked to categorize each image with one of three emotion labels (*happiness, disgust,* and *surprise*). In the experiment proper the same composite and noncomposite stimuli used in Experiment 1 were presented individually in random order. The participants' task was to categorize the bottom segment of each image with one of the same three emotion labels as quickly and accurately as possible.

The inverted condition block was essentially identical to the upright condition block, except that all of the stimuli were inverted. In all other respects, the design and procedure of Experiment 2 were the same as for Experiment 1.

Results

Participants' mean correct RTs (with standard error bars) to identify the bottom half of the composite and noncomposite facial expressions in upright and inverted formats are shown in the left graph of Figure 5. The right graph shows participants' mean error proportions (with standard error bars) for the same experiment.

Reaction times. Our principal form of analysis involved RTs for correct responses. These were submitted to a two-factor ANOVA investigating stimulus type (composite and noncomposite; repeated measure) and stimulus orientation (upright and inverted; repeated measure). There was a significant effect of stimulus type, $F(1, 11) = 9.74, p < .01$, indicating that participants found it harder to identify the expression shown in the bottom half of the composite images. This was qualified by a significant interaction between stimulus type and stimulus orientation, $F(1, 11) =$

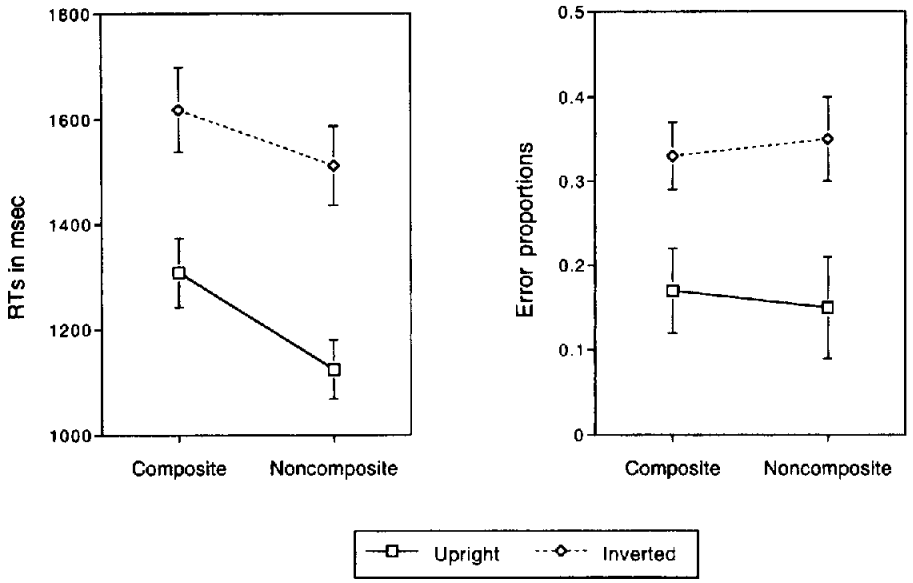


Figure 5. Data from Experiment 2. The left graph shows participants' mean correct reaction times (RTs; with standard error bars) to identify the expression displayed in the bottom half of the composite and noncomposite images presented in upright and inverted formats. The right graph shows participants' mean error proportions (with standard error bars) from the same experiment.

6.62, $p < .05$. Simple effects analyses of the interaction effect showed a significant effect of stimulus type (composite and noncomposite) for the upright condition, $F(1, 11) = 15.14$, $p < .005$, and a borderline, nonsignificant effect for the inverted condition, $F(1, 11) = 4.02$, $.1 > p > .05$. Finally, there was also a significant effect of stimulus orientation, $F(1, 11) = 14.06$, $p < .005$, demonstrating that, overall, participants found the task easier when the stimuli were upright.

Error rates. A subsidiary analysis examined participants' error rates to check that the slower responses in the composite condition were not also accompanied by more accurate performance. Error proportions were arcsin transformed and submitted to a two-factor ANOVA investigating stimulus type (composite and noncomposite; repeated measure) and stimulus orientation (upright and inverted; repeated measure). This showed a significant main effect of stimulus orientation, $F(1, 11) = 11.80$, $p < .01$, reflecting that, overall, participants were significantly more accurate at identifying the expression shown in the upright images. There were no other significant effects (Fs < 1.10). Thus, there was no statistical evidence of participants trading accuracy for speed.

Discussion

Experiment 2 demonstrates three findings. First, the results replicated the findings of Experiment 1. In the upright condition, participants were significantly slower (but no more accurate) to identify the expression shown in the bottom half of the composite images relative to their performance with the noncomposite images. Second, inverting the images significantly disrupted the composite effect for facial expressions; the composite effect was statistically reliable for the upright condition only. This second finding is similar to Young et al.'s (1987) observation that the composite effect for facial identity is lost when the stimuli are inverted (see also Carey & Diamond, 1994; Hole, 1994). Finally, Experiment 2 also showed a significant main effect of stimulus inversion. This indicates that, overall, participants were significantly slower to identify the expression shown in the bottom half of the composite and noncomposite stimuli when they were inverted. This is consistent with McKelvie's (1995) finding that facial expressions are more difficult to recognize in inverted faces (see also Valentine & Bruce, 1988).

It is important to note that the negative effect of inversion on facial identity recognition is usually attributed to the idea that configural features are more difficult to process in inverted faces or, less specifically, that holistic processing of faces is made more difficult by stimulus inversion. The fact that the composite effect for facial expression is also disrupted by inversion converges on the idea that configural features may also be used to encode facial expressions. Hence, the results of Experiment 2 provide further support for the configural model rather than the part-based model of facial affect recognition.

Given that our interpretation of the results of Experiments 1 and 2 has substantial implications for the understanding of facial expression perception, it was important to consider whether there were any alternative interpretations of the composite effect we had found. We considered that one possibility was that the composite stimuli were simply more

attention-grabbing than the noncomposites, possibly because facial composites prepared from the top and bottom halves of two different pictures inevitably look like unusual (or distorted) faces, causing participants to look longer at them before deciding on a response. Experiment 3 addressed this alternative explanation.

Experiment 3

Although it seems likely that the composite effect found in Experiments 1 and 2 is attributable to a disruption of configural processing, an alternative explanation may exist. As we have said, it is possible that the participants may have been distracted by the slightly distorted and unusual appearance of the composite stimuli and, hence, slower to make their response. Clearly, it was important to address this alternative explanation, and one means of testing it became evident while we were preparing the stimuli for Experiments 1 and 2.

When creating stimuli, we noted that if the two face halves are taken from different identities posing the *same* facial expression (e.g., happiness), the resultant composite face also looks happy. This result suggested a prediction: If the composite effect we had observed was due to a disruption of configural information for facial expression, then composite faces prepared from two different identities posing the same expression (same-expression composites) should not show the effect. This is because the top and bottom segments of these images contain configural information relating to the same facial affect, meaning that there is no conflict between the configural information for expression in the two halves, even though the identities are different. Alternatively, if the effect we had observed was due to the composite stimuli being more attention-grabbing than the noncomposites (as a result of discontinuities in texture across the two face halves, etc.), then a significant composite effect should be observed for the same-expression composites. This was tested in Experiment 3.

As a comparison condition, we also included composite images prepared from different identities posing different facial expressions (different-expression composites). We predicted that these images should produce the same composite effect found in Experiments 1 and 2, because for the different-expression composites, there is a conflict of configural information for facial expression across the two halves of the image. Hence, our suggestion that the composite effect reflected configural processing of the images would hold true if Experiment 3 showed a significant interaction effect between stimulus type (composite and noncomposite) and top–bottom expression congruency (same expression and different expression).

Method

Participants. Twelve participants (7 women, 5 men) aged between 18 and 40 years and from the same population as Experiments 1 and 2 took part in the experiment. All had normal or corrected-to-normal vision, and none had participated in the previous experiments.

Materials. The stimuli were prepared from pictures of the same four models (C, NR, PF, and SW) used in Experiments 1 and 2 posing the facial expressions happiness, disgust, and surprise. The top and bottom halves of these faces were combined to produce all possible composite expressions in which the two halves were taken from different models' faces. For 36 of these images, the top and bottom halves showed the same expression (same-expression composite; e.g., top = happiness Model C, bottom = happiness Model NR), and for the remaining 72, the two halves showed different expressions (different-expression composite; e.g., top = happiness Model C, bottom = disgust Model NR).

Noncomposite versions of the same stimuli were produced using the method described in Experiment 1. Recall that there are two possible versions of noncomposite stimuli: (a) top half shifted to the right of the bottom half and (b) top half shifted to the left of the bottom half. Given that there were twice as many different-expression composites as same-expression composites, both versions of noncomposite were produced for each of the same-expression images, whereas for the different-expression images, the two versions were counterbalanced across stimuli. Examples of composite and noncomposite images prepared from two of the four models used in Experiment 3 are shown in Figure 6.

Design and procedure. Two within-subjects factors were investigated: stimulus type (composite and noncomposite) and top–bottom expression congruency (same expression and different expression).

All stages of the experiment used the presentation format described in Experiment 2 (i.e., 500-ms fixation, 500-ms blank interstimulus interval followed by the stimulus, which remained in view until the participant responded). The experiment began with a session in which the original 12 whole faces (four models, each posing three facial expressions) used to prepare the composites were presented individually in random order. Each face was shown three times, and the participant identified the emotion displayed by pressing one of three keys marked with the labels *happiness, disgust,* and *surprise;* label positions were counterbalanced across participants. Next, half of the participants were presented with the top segments of these same faces and half with the bottom segments. Again, each image was presented three times, and the participants' task was to identify the facial expression as one of happiness, disgust, or surprise. After this, the participants that had seen the top sections were presented with the bottom sections of the same facial expressions and vice versa. Their task was the same, namely to identify the emotion.

In the experiment proper the participants were presented with equal numbers (36) of same-expression composites, same-expression noncomposites, different-expression composites, and different-expression noncomposites in random order. The stimuli were counterbalanced across two stimulus sets to accommodate the different numbers of same-expression and different-expression images; half of the participants were assigned to one stimulus set and half to the other. Participants were instructed to identify the expression displayed on the bottom half of each image by pressing the appropriate response key (happiness, surprise, or disgust) as quickly and accurately as possible. To familiarize the participants with the composite and noncomposite images, the experiment proper was preceded by 10 practice trials selected at random from the experimental trials. The composite images subtended a horizontal visual angle of approximately 4.6°, and the noncomposite images subtended a horizontal visual angle of approximately 5.7°; the vertical visual angle for both was approximately 6.3°.

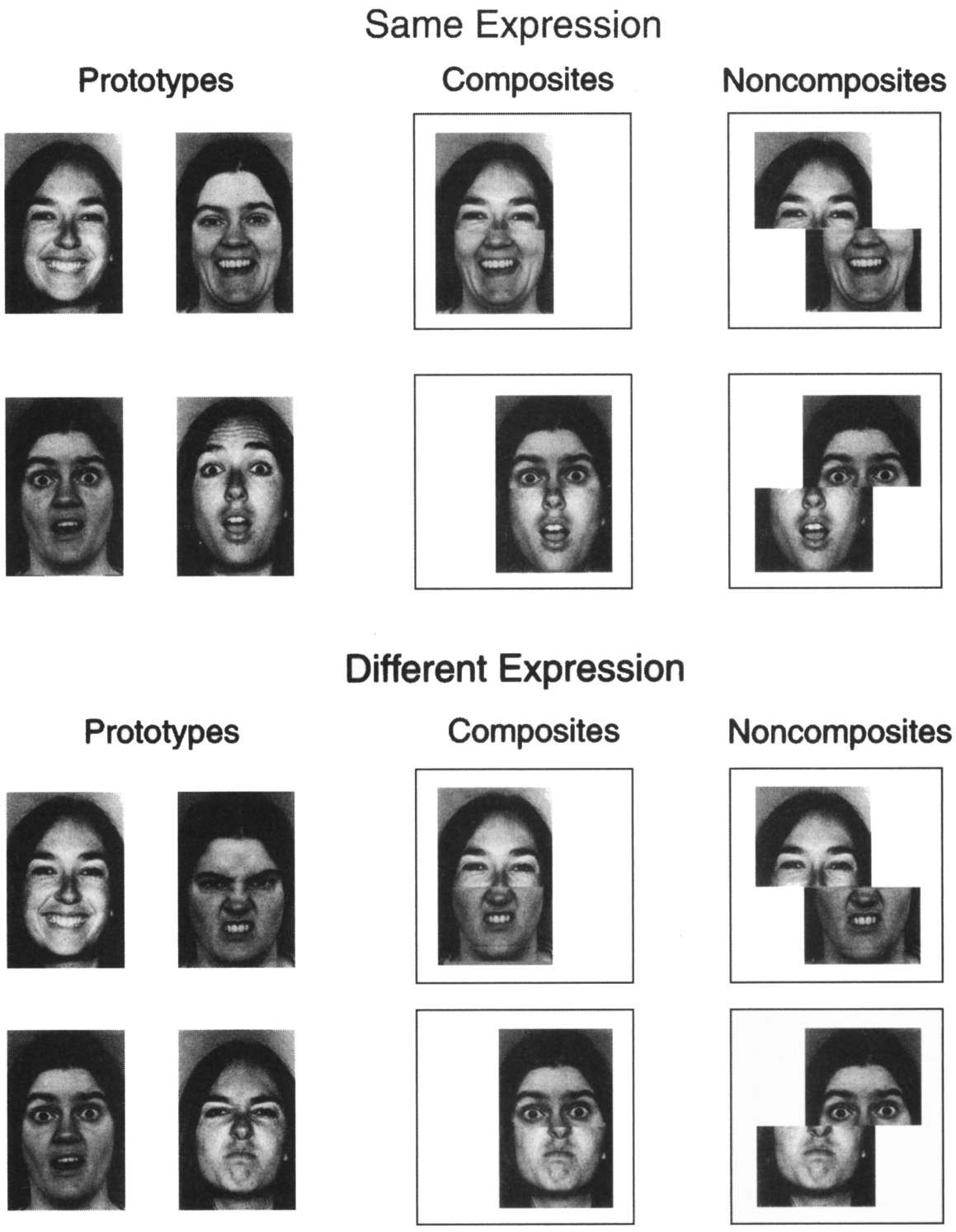


Figure 6. Examples of the stimuli used in Experiment 3. Composite (middle) and noncomposite (right) facial expression stimuli were prepared from the top and bottom halves of two models' faces (left) posing the same expression (same-expression images; top row) or different expressions (different-expression images; bottom row). Images from *Pictures of Facial Affect,* by P. Ekman and W. V. Friesen, 1976. Copyright 1976 by P. Ekman and W. V. Friesen. Adapted with permission.

Results

Participants' mean correct RTs (with standard error bars) to identify the bottom half of the composite and noncomposite facial expressions in same-expression and different-expression conditions are shown in the left graph of Figure 7. The right graph shows participants' mean error proportions (with standard error bars) for the same experiment.

RTs. Our principal form of analysis involved RTs for correct responses. These were submitted to a two-factor ANOVA investigating stimulus type (composite and noncomposite; repeated measure) and top–bottom expression congruency (same expression and different expression; repeated measure). There was a significant effect of stimulus type, $F(1, 11) = 10.10$, $p < .01$, indicating that, overall, participants were slower to identify the expression shown in the bottom half of the composite images. This was qualified by a significant interaction between stimulus type and top–bottom expression congruency, $F(1, 11) = 12.94$, $p < .005$. Simple effects analyses showed a significant effect of stimulus type for the different-expression images, $F(1, 11) = 24.00$, $p < .0001$, but not for the same-expression images ($F < 1.00$). There was also a significant effect of expression congruency, $F(1, 11) = 9.67$, $p < .01$, demonstrating that, overall, participants were significantly slower to identify the expression in the different-expression images; post hoc t tests ($p < .05$) showed that this held for the composite images (same expression < different expression) but not for the noncomposite images.

Error rates. A subsidiary analysis examined participants' error rates to check that the slower responses were not accompanied by more accurate performance. Error proportions were arcsin transformed and submitted to a two-factor ANOVA investigating stimulus type (composite and noncomposite; repeated measure) and top–bottom expression congruency (same expression and different expression; repeated measure). There was a marginal main effect of stimulus type, $F(1, 11) = 3.86$, $.1 > p > .05$, reflecting an overall trend toward more errors with the composite stimuli. This was qualified by a significant interaction between stimulus type and top–bottom expression congruency, $F(1, 11) = 5.61$, $p < .05$. Simple effects analyses showed a significant effect of stimulus type for the different-expression images, $F(1, 11) = 15.88$, $p < .005$, but not for the same-expression images ($F < 1.00$). There was also a significant main effect of top–bottom expression congruency, $F(1, 11) = 7.23$, $p < .05$, indicating that, overall, participants made significantly more errors with the different-expression images. Post hoc t tests ($p < .05$) indicated that this effect held for the composite and noncomposite images (same expression < different expression). The results of the error rates analysis, then, show no evidence of participants trading speed for accuracy.

Discussion

The results of Experiment 3 can be summarized as follows. First, a composite effect for facial expressions was found when the images compose the top and bottom segments of different people's faces posing different expressions; this replicates and extends the findings of Experiments 1 and 2. Second, no composite effect was observed when the stimuli were prepared from the top and bottom

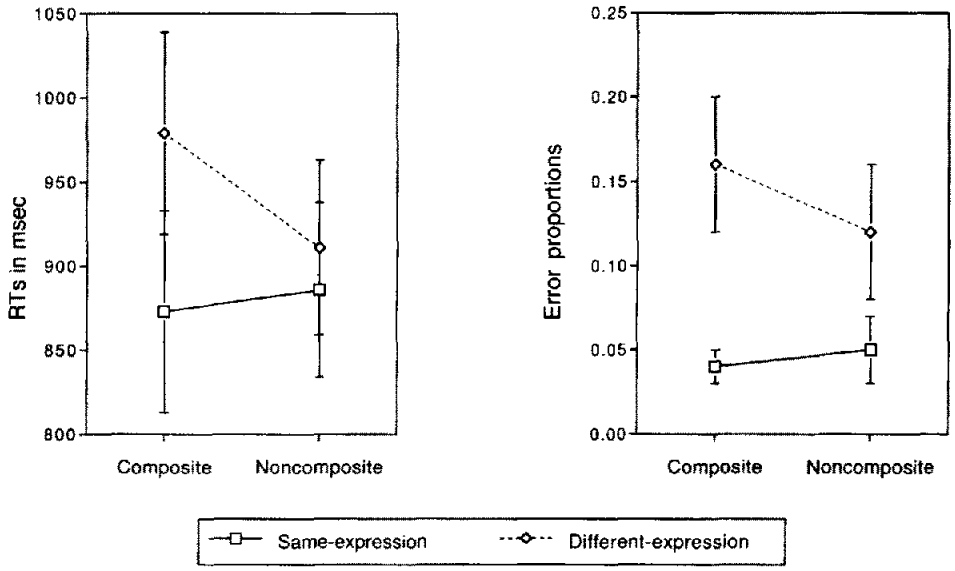


Figure 7. Data from Experiment 3. The left graph shows participants' mean correct reaction times (RTs; with standard error bars) to identify the expression displayed in bottom half of composite and noncomposite images containing the same facial expression (same-expression images) or different facial expressions (different-expression images). The right graph shows participants' mean error proportions (with standard error bars) from the same experiment.

segments of different people's faces posing the same expression.

The significant factor that differentiated these two types of stimuli was that for one stimulus, the expressions shown in the top and bottom segments were different (different-expression composite), whereas for the other stimulus they were the same (same-expression composites). The results of Experiment 3, then, confirm that the composite effect found in Experiments 1 and 2 cannot be attributed to the idea that composite faces are more attention-grabbing than noncomposites. Instead, these findings are consistent with the suggestion that for the different-expression composites, there is a conflict between the configural information in the two face halves, whereas for the same-expression composites there is no such conflict. Once again, then, our data are consistent with the configural model of facial affect processing.

Note that in Experiments 1 and 2, the top and bottom halves of each composite expression were taken from pictures of the same model; hence, it was relatively easy to align the nose, hairline, and so forth, in these images and to avoid abrupt discontinuities of stimulus texture. This was considerably more difficult in Experiment 3 because the two halves belonged to the faces of different models. Experiment 3, then, demonstrated that the composite effect for facial expressions is robust, because the effect was found even when the composite images did not look (on close inspection) like fully credible faces.

Finally, it is worth noting one further point. In Experiment 3, our composite stimuli were prepared from the top and bottom halves of recognizable-bottom facial expressions. However, we still observed a large composite effect in the different-expression condition. This would suggest that the face half that the participant was not instructed to attend to (i.e., the top half of the face in Experiment 3), did not need to display a highly recognizable expression for the effect to occur; the average recognition rates from Experiment 1 for the top halves of these expressions across the four models used were as follows: happiness, 68.5%; disgust, 43.75%; and surprise, 84.75% (chance = 16.67%). Contrast these rates with the recognition rates for the bottom halves of the same expressions (happiness, 100%; disgust, 87.5%; and surprise, 85.75%). Hence, the critical factor for a composite effect to be observed may be that the configural information in the unattended half is inconsistent with that in the attended half, rather than that the unattended half itself contains another readily identifiable facial expression.

Experiment 4

Figure 6 illustrates a point demonstrated in Experiment 3 that the top and bottom halves of two faces of different people posing the same expression (e.g., happiness) can be combined to generate a perceptually new facial identity without disrupting facial expression (i.e., the face still looks happy). In other words, the composite face is a poor match for either of the two original models' faces but a good match for the expression posed by both models. This observation implies that the configural information used to encode facial identity may be different from that used to encode facial expression. We reasoned that support for this hypothesis could be found by showing that the configural processing of facial identity and facial expression can be selectively disrupted.

To demonstrate this, we used three types of composite stimuli prepared from (a) pictures of the same person posing different facial expressions (same-identity–different-expression composites), (b) pictures of different people posing the same facial expression (different-identity–same-expression composites), and (c) pictures of different identities posing different facial expressions (different-identity–different-expression composites). We predicted that if participants could selectively attend to the configural information that specifies either facial identity or facial expression, then we should find different patterns of performance with different task instructions. Hence, when the instructions are to indicate the identity shown in the bottom half of a composite face, participants' responses should be fastest when the top and bottom segments contain the same person's face (same-identity–different-expression composites). However, when the instructions are to identify the expression shown in the bottom half, participants responses should be fastest when the top and bottom segments contain the same facial expression (different-identity–same-expression composites).

The third type of composites (different-identity–different-expression composites) was used for the following reason. If different configural information is used to specify identity and expression, then although there should be a significant cost to RTs when the attribute (identity or expression) that the participants are asked to attend to is incongruent across the two face halves, there should be no additional cost when the two halves are incongruent with respect to both facial attributes. So, for example, participants' RTs to identify the expression shown in the bottom half of the same-identity–different-expression and different-identity–different-expression composites should not differ. Similarly, there should be no reliable difference between their RTs to indicate the identity shown in the bottom half of the different-identity–same-expression and different-identity–different-expression composites.

To test these hypotheses, it was necessary to use either pictures of already familiar faces posing different expressions or faces from the Ekman and Friesen (1976) series, which were made familiar to the participants at the beginning of the experiment. The latter method of making unfamiliar faces familiar in the course of the experiment has been successfully used by Young et al. (1987) and Carey and Diamond (1994) in their investigations of the composite effect for facial identity, and, on balance, we selected it for two reasons. First, this method facilitates comparison with other experiments in this article and, second, full-face pictures of personally familiar people or of celebrities posing different facial expressions are difficult to obtain.

Note that we did not use noncomposite images in Experiment 4 for the following reason. Experiments 1, 2, and 3 were all consistent with the suggestion that facial

expression composites are encoded configurally. This effect was observed despite the participants being instructed to attend to only one section of the face. Hence, although participants would have improved their performance for the composite condition by only processing the information in the attended half, they were apparently not able to use this strategy. These results strongly suggest that under these circumstances, the configural encoding of facial expression is an automatic process that is beyond conscious control. Similarly, the results of previous studies examining the composite effect for facial identity (Carey & Diamond, 1994; Young et al., 1987) suggest that the same is true for the perception of configural information relating to facial identity.

Given that Experiment 4 used the same basic task used in the experiments discussed above (i.e., identify the person, or identify the expression shown in the bottom half of the composite), we could see little reason for including a series of noncomposite conditions as a check of an effect that is apparently beyond the participants' control. In addition, Experiment 3 had demonstrated that RTs to identify the expression shown in the bottom section of same-expression–different-identity and different-expression–different-identity composites were significantly different. This was consistent with the idea that the configural information for facial expression was disrupted in one condition (different-expression–different-identity) but not the other (same-expression–different-identity). Hence, the same-expression–different-identity composites essentially served as a control for the expression recognition task (i.e., a similar role to the noncomposite images used in Experiments 1, 2, and 3). Similarly, we used the different-expression–same-identity images as a control for the identity recognition task. This is because we predicted that participants should show significantly less interference with these images compared with the composites in which the top and bottom sections contained different people's faces. As it turned out, our predictions were confirmed.

Method

Participants. Fifteen participants (13 women, 2 men) aged between 18 and 50 years and from the same population as Experiments 1–3 took part in the experiment. All had normal or corrected-to-normal vision, and none had participated in the previous experiments.

Materials. All stimuli in Experiment 4 were prepared from pictures of three models (C, NR, and SW) from the Ekman and Friesen (1976) series, each posing three different facial expressions (happiness, disgust, and surprise). All possible combinations of the top and bottom halves of these nine different pictures were produced to give a total of 72 different composite faces. For 18 of these, the top and bottom halves displayed different expressions posed by the same model (different-expression–same-identity composites; e.g., top = happiness Model C, bottom = disgust Model C); for 36, the top and bottom halves displayed different expressions posed by the different models (different-expression–different-identity composites; e.g., top = happiness Model C, bottom = disgust Model NR) and for the remaining 18 the top and bottom halves displayed the same expression posed by different models (same-expression–different-identity composites; e.g., top = happiness Model C, bottom = happiness Model NR). Note that the original images (literally same expression–same identity) were not used in the experiment proper. Examples of the three image types prepared from two of the models used in Experiment 4 are shown in Figure 8.

Design and procedure. Two within-subjects factors were investigated: composite type (different expression–same identity, same expression–different identity, and different expression–different identity) and task instructions ("identify the person" and "identify the expression"). All stages of the experiment used the presentation format described for Experiment 2 (i.e., 500-ms fixation, 500-ms blank interstimulus interval and then the stimulus that remained in view until the participant responded).

The experiment consisted of two sections corresponding to the two levels of the task instructions factor (identify the person and identify the expression); half of the participants were assigned to the identify-the-person section first and half to the identify-the-expression section. Both sections used the same basic design.

Identify-the-person trials. The section began with a training session. In this training session, the three models' faces were presented with neutral facial expressions (expressionless faces), and each was accompanied by an arbitrarily assigned first name (Model C = Susan, Model NR = Margaret, Model SW = Tracy); these names were printed in uppercase letters and positioned below the face. Each face–name pair was presented five times for 5 s each in random order. The participant was instructed to look at the faces and try to remember the models' names because later they would be tested on them. Following this, pictures of the same models posing the three facial expressions, happiness, disgust, and surprise, were presented individually and without name labels. The participant was asked to identify each model's name by pressing one of three buttons on a box interfaced with the computer; the keys were marked with the names (Susan, Margaret, and Tracy), and their positions were counterbalanced across participants. Each of the nine different pictures (three models × three facial expressions) were presented three times in random order. If participants made an error (e.g., pressed the Susan button in response to Margaret's face), the computer made a "beep" noise, and they were invited to try again until the correct response was made. Participants who each made more than 3 errors (out of a total of 27 trials; maximum total errors = 54; i.e., 2 errors per trial) in this stage of the experiment were excluded from the analysis.

Next, half of the participants were presented with the top halves of the same faces and half of the participants with the bottom halves of the same faces. Again, each image was presented individually, three times in random order, and the participant's task was to indicate each model's name by making a button-press response. Following this, the participants who had seen the top sections were presented with the bottom sections of the same faces and vice versa. Their task was the same: to identify the models' names.

In the experiment proper the participants were presented with equal numbers (18) of the three types of composite faces (different expression–same identity, different expression–different identity, and same expression–different identity) in random order; each image was presented twice. The stimuli were counterbalanced across two different stimulus sets to accommodate the different number of images in the three levels of the composite type condition. Half of the participants were assigned to one stimulus set and half to the other. The composite images subtended a horizontal visual angle of approximately 4.6°, and a vertical visual angle of approximately 6.3°. Participants were instructed to identify the name of the model shown in the bottom half of each composite image by pressing one of the three keys listed above. To familiarize

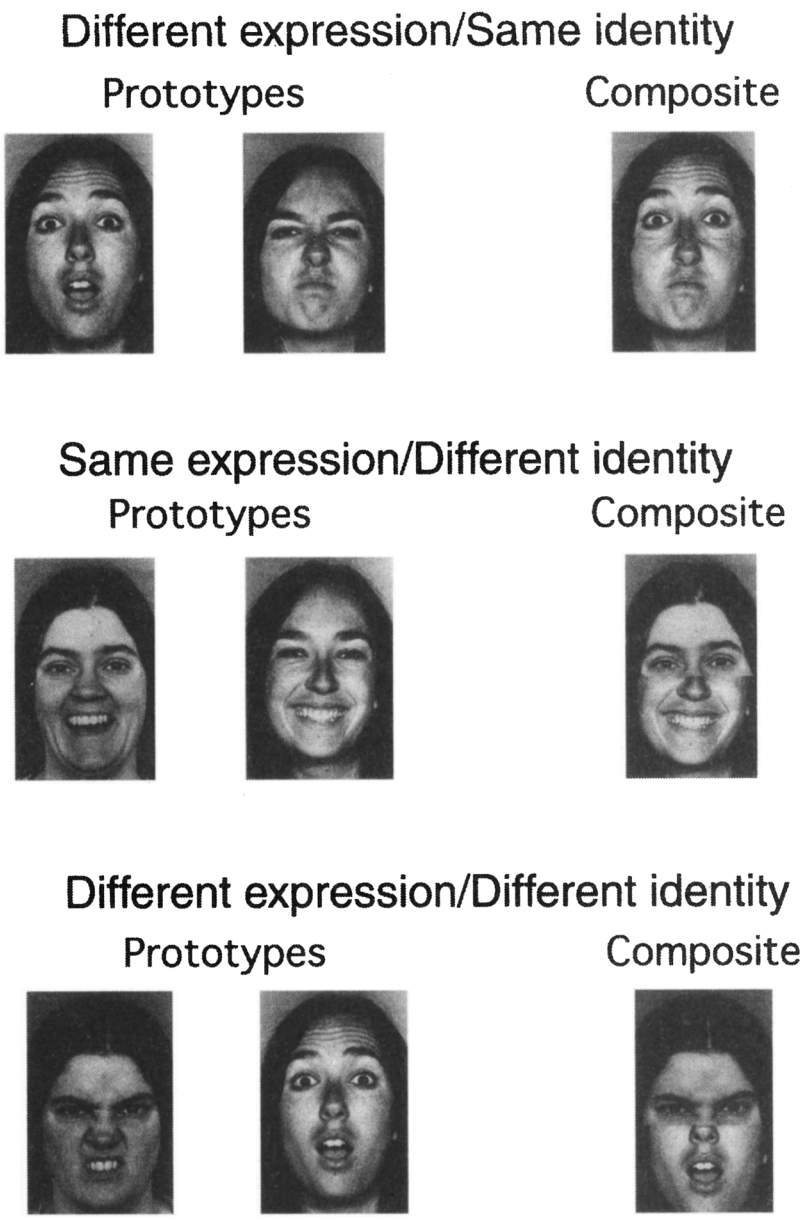


Figure 8. Examples of the stimuli used in Experiment 4. Composite facial expressions were prepared from the top and bottom halves of (a) two different prototype expressions posed by the same model (different expression–same identity; top row), (b) the same prototype expression posed by different models (same expression–different identity; middle row), and (c) two different prototype expressions posed by different models (different expression–different identity; bottom row). Images from *Pictures of Facial Affect,* by P. Ekman and W. V. Friesen, 1976. Copyright 1976 by P. Ekman and W. V. Friesen. Adapted with permission.

the participants with the composite images, the experiment proper was preceded by six practice trials selected at random from the 54 experimental trials.

Identify-the-expression trials. The identify-the-expression section began with exactly the same training session described above, in which each model was presented five times with a neutral expression and their name label. Next, the same three models were presented posing the expressions happiness, surprise, and disgust, and participants were asked to identify their facial expression by pressing one of three keys labeled *happiness, disgust,* and *surprise;* each face was presented three times in random order. If the participant made an error in his or her choice, the computer made a beep noise and they were asked to try again. Any participant who made more than 3 errors out of a total of 54 was again excluded from the analysis. In all other respects, the design of the identify-the-expression trials was the same as the identify-the-person trials described above. The only difference was in the task instructions.

Hence, participants were next presented with the top segments of the stimuli in one block and bottom segments of the same faces in another block; order of presentation of these two blocks was counterbalanced across participants. Their task in each case was to identify the facial expression by pressing one of three keys marked *happiness*, *disgust*, and *surprise*. In the experiment proper, they identified the expression shown in the bottom half of the composite stimuli.

Results

In the first block of the identify-the-person trials in which the participants were shown the whole faces and asked to indicate the models' name, 3 of the participants made three errors (the criterion number for rejection). In the corresponding section of the identify-the-expression block, the same 3 participants reached or exceeded this criterion error rate. These participants were therefore excluded from the following analysis, leaving data from 12 participants (10 women, 2 men; 18–50 years). The mean number of correct responses made by these 12 participants when identifying the expression and person in the whole-face block in the respective sections were as follows: Identify the person, $M = 26.42$, $SD = 0.90$; and identify the expression, $M = 26.50$, $SD = 0.80$. Clearly, then, these participants had little difficulty in identifying the models' names or their facial expressions.

Participants' mean correct RTs (with standard error bars) to identify the person and expression in the bottom half of the three types of composite images (different expression–same identity, different expression–different identity, and same expression–different identity) are shown in the left graph of Figure 9. The right graph shows participants' mean error proportions (with standard error bars) for the same experiment.

RTs. Our principal form of analysis involved RTs for correct responses. These were submitted to a two-factor ANOVA investigating composite type (different expression–same identity, different expression–different identity, and same expression–different identity; repeated measure) and task instructions (identify the person and identify the expression; repeated measure). There was a significant effect of composite type, $F(2, 22) = 7.39, p < .005$. Post hoc t tests ($p < .05$) indicated that, overall, participants were slowest to identify the bottom segments of the different-identity–different-expression composites; RTs to the different-identity–same-expression and same-identity–different-expression composites did not reliably differ. The main effect of composite type was qualified by a significant interaction between composite type and task instructions, $F(2, 22) = 14.39, p < .0001$. Simple effects analyses showed a significant effect of composite type for both levels of task instructions condition: Identify the person, $F(2, 22) = 12.01, p < .0001$, and identify the expression, $F(2, 22) = 9.82, p < .001$; however, the pattern of the effects in these two conditions was different. Post hoc t tests ($p < .05$) showed that for the identify-the-person condition, RTs to the different-expression–same-identity composites were significantly faster than those to the same-expression–different-identity and different-expression–different-identity compos-

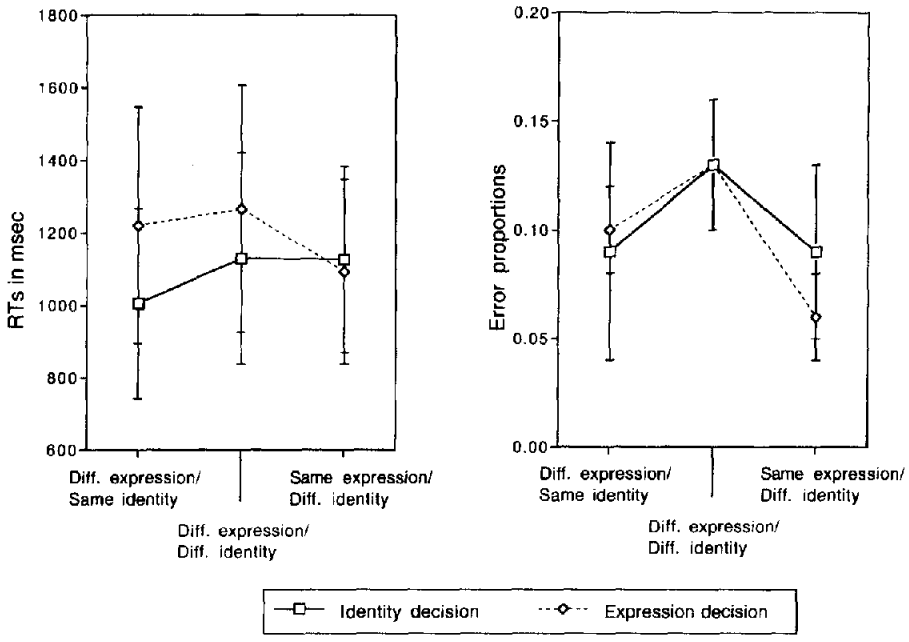


Figure 9. Data from Experiment 4. The left graph shows participants' mean correct reaction times (RTs; with standard error bars) to identify the expression (expression decision) or identity (identity decision) in the bottom segment of three types of composite image (different expression–same identity, different expression–different identity, and same expression–different identity). The right graph shows participants' mean error proportions (with standard error bars) from the same experiment.

ites, which did not reliably differ (different expression–same identity < [same expression–different identity = different expression–different identity]). For the identify-the-expression condition, RTs to the same-expression–different-identity composites were significantly faster than those to the different-expression–same-identity and different-expression–different-identity composites, which did not reliably differ (same expression–different identity < [different expression–same identity = different expression–different identity]). Finally, there was no overall significant effect of task instructions, indicating that participants' RTs to perform the identify-the-person and identify-the-expression tasks did not reliably differ; this shows that the two tasks were of comparable difficulty, as assessed by the RTs measure.

In summary, the results of the RTs analysis demonstrate that participants were significantly slower to perform the task when the attribute (expression or identity) that they were asked to attend to was incongruent across the two face halves. Moreover, there was no additional significant cost when the unattended attribute was also incongruent across the two halves. Note that this result is consistent with the breakdown of the main effect of composite type. This showed that, overall, participants were slowest to classify composites in which the top and bottom halves were different identities and different expressions. This result is to be expected because we predicted that the different-expression–different-identity condition should show slower RTs in both levels of the task instructions condition (identify the person and identify the expression).

Error rates. A subsidiary analysis examined participants' error rates to check that the slower responses were not also accompanied by more accurate performance. Error proportions were arcsin transformed and submitted to a two-factor ANOVA investigating composite type (different expression–same identity, same expression–different identity, and different expression–different identity; repeated measure) and task instructions (identify the person and identify the expression; repeated measure). The only significant effect was the main effect of composite type (different expression–same identity, same expression–different identity, and different expression–different identity), $F(2, 22) = 3.76, p < .05$. Post hoc t tests ($p < .05$) showed that, overall, participants made significantly more errors in the different-expression–different-identity condition compared with the same-expression–different-identity condition. There were no other statistically reliable effects. Thus, there was no evidence of participants trading accuracy for speed. Furthermore, the absence of a significant effect of task instructions indicates that the two tasks (identify the person and identify the expression) were of comparable difficulty, as assessed by error proportions; this is consistent with the findings of the RT analysis.

Discussion

The results of Experiment 4 showed that when viewing the same stimulus set, participants produced different patterns of RTs depending on whether they were asked to perform a facial identity task or a facial expression task.

Three types of composite image were used in which the top and bottom halves were (a) different expressions posed by the same model (different-expression–same-identity composites), (b) the same expression posed by different models (same-expression–different-identity composites), and (c) different expressions posed by different models (different-expression–different-identity composites). When the task instruction was to identify the facial expression shown in the bottom half of these composites, participants' RTs were significantly faster when the top and bottom halves contained the same expression (same-expression–different-identity composites), than when they contained different expressions (different expression–same identity and different expression–different identity). Moreover, for the two conditions in which the top and bottom segments of the composite showed different expressions (different expression–same identity and different expression–different identity), there was no significant additional cost when the two halves contained both different expressions and different identities (different expression–different identity).

When the task was to recognize the identity shown in the bottom half of the composites a different pattern of performance was observed. Here participants' RTs were significantly faster when the top and bottom segments showed the same identity (different-expression–same-identity composites) than when they contained different identities (same expression–different identity and different expression–different identity). And for the two conditions in which the two halves of the composite contained different identities (same-expression–different-identity and different-expression–different-identity composites), there was no additional significant cost when they displayed both different identities and different expressions (different-expression–different-identity composites).

These results are consistent with previous findings showing that participants can selectively attend to information in a face that is relevant to its expression and discard information relevant to its identity, or vice versa (Campbell et al., 1996; Etcoff, 1984; Young et al., 1986). However, the results of Experiment 4 go beyond these previous studies and offer an impressive demonstration of participants' ability to selectively attend to different types of configural information; one relating to the representation of facial identity the other to the representation of facial expression. Experiment 4 also demonstrates that these two forms of configural information can be selectively disrupted. One implication of this finding is that the configural information used for facial identity and facial expression perception is different.

Finally, it is worth noting that these results were obtained using a within-subjects design. This shows that the participants have an impressive ability to shift their attention from processing configural information that is relevant to facial identity in one block to processing configural information that is relevant to facial expression in another without experiencing any substantial interference from the immediately preceding task.

In summary, Experiment 4 suggests that different configural information is used to encode facial identity and facial expression. The nature of these two types of configural features is discussed in the following section.

General Discussion

The results of these experiments demonstrate a number of points. These can be summarized as follows.

1. Facial expressions of the basic emotions can be divided into recognizable-top and recognizable-bottom categories. Experiment 1 found that anger, fear, and sadness showed a significant recognizable-top bias, whereas happiness and disgust showed a significant recognizable-bottom bias. Surprise showed no significant bias and was equally distinguishable from whole-face, top and bottom segments. These results largely replicate the findings of a previous study by Bassili (1979).

2. Composite facial expressions were prepared by aligning the top half of one facial expression (e.g., anger) with the bottom half of another (e.g., happiness). In three separate experiments, we have demonstrated that participants are significantly slower to identify the expression in either half of these composite images relative to a noncomposite control condition in which the two halves are misaligned (Experiments 1, 2, and 3). These results parallel Young et al.'s (1987) earlier finding of a similar effect for facial identity.

3. Young et al. (1987) demonstrated that the composite effect for facial identity is abolished when the stimuli are inverted. The results of Experiment 3 demonstrate that the composite effect for facial expression is also significantly disrupted by stimulus inversion.

4. The composite effect for facial expression is found when the top and bottom segments are taken from pictures of different expressions posed by the same model (Experiments 1 and 2) or two different models (Experiment 3). However, this effect is not found when the two segments are taken from pictures of different models posing the same facial expression (Experiment 3). This result serves to exclude the suggestion that the composite effect is an artefact of stimulus quality—for example, the composite stimuli appearing slightly distorted and unusual as faces and, hence, more attention-grabbing than the noncomposite images.

5. In Experiment 4, participants were presented with three types of composite faces in which the top and bottom segments were (a) different expressions posed by the same model, (b) the same expression posed by different models, and (c) different expressions posed by different models. When participants were asked to name the identity shown in the bottom half of these images, their RTs were significantly slower for those composites prepared from the top and bottom halves of different models' faces (b and c above). However, when they were asked to identify the expression shown in the same half, significantly slower RTs were found for images in which expression was incongruent across the two halves (a and c above). Moreover, no added cost was found when the attribute that participants had not been instructed to report (e.g., expression, in the identity task) was also incongruent across the two face halves. These findings suggest that the composite effects for facial expression and facial identity may disrupt the perception of different configural features.

These results have important implications for the perceptual representations of facial expressions and we deal with each of them in turn.

As we discussed in the introduction, previous studies have shown that some facial expressions are more recognizable from the top half of the face (recognizable-top expressions), whereas others are more readily recognized from the bottom half (recognizable-bottom expressions, Bassili, 1979; Hanawalt, 1944; Plutchik, 1962). Our results confirm these observations and are highly consistent with Bassili's findings that were obtained using animated examples of the same facial expressions from a different image set.

It is worth emphasizing that observations of top–bottom expression dominance are not inconsistent with the idea that configural information is important for facial expression recognition. It is possible for the overall configuration of a facial expression to contribute toward its recognition, despite the sufficient information for accurate recognition of an the emotion being contained largely in the top (or bottom) section of the face. Likewise, the observation that a person's identity is more readily recognized from the eye region than the mouth region (see Shepherd, Davies, & Ellis, 1981, for a review) does not detract from the well-established finding that configural features are important for facial identity recognition. We should also point out that in the preliminary experiment conducted in Experiment 1 (see Table 1), none of the top or bottom sections of the expressions were recognized at chance (chance error proportion = 0.83). This means that for all six expressions, both top and bottom sections of the face contained information that was associated with the emotion.

In the introduction, we outlined a configural model and a part-based model of facial expression recognition. Consistent with a configural model, Experiments 1, 2, and 3 showed that a facial composite effect, similar to the one shown for facial identity by Young et al. (1987), can also be found for facial expression. Young et al. suggested that the composite effect for facial identity reflects a disruption of configural information processing, because when the top and bottom halves of two identities' faces are aligned, they produce a new facial configuration that interferes with one's ability to recognize the identity in the top or bottom part of the face. We think that a similar explanation can be applied to the composite effect for facial expression. That is, the top and bottom halves of the two expressions align to produce a new facial expression configuration. Consequently, this interferes with identifying the emotion shown in either half of the composite expressions. Misaligning the two face halves, however (noncomposite condition), means that the face is no longer encoded as a configural whole, and, hence, the feature information relating to the expression in the top and bottom halves can be accessed faster. The results of our experiments, then, suggest that the composite effects for

facial identity and facial expression are somewhat similar. This similarity is further emphasized by our observation that the composite effect for facial expression is disrupted by inverting the stimuli.

Recall that Young et al. (1987) showed that the composite effect for facial identity is only found when the stimuli are presented upright; inverting the stimuli (i.e., 180° rotation) abolished the effect (see also Carey & Diamond, 1994). This finding is consistent with Carey and Diamond's (1977) suggestion that inversion impairs the perception of configural information. Hence, our observation that the composite effect for facial expression is also disrupted by stimulus inversion further supports a configural model of facial expression recognition.

In Experiment 3, we addressed the criticism that the composite effect could be attributed to some inherent quality of composite faces that makes them more attention-grabbing than the noncomposites. However, Experiment 3 discounted this interpretation by showing that a composite effect is not found when the top and bottom sections contain the same facial expression posed by different models (same-expression composites). This finding also suggested a further hypothesis.

From examining the same-expression composites, our intuition was that the configural features used for facial expression recognition were different to those used for facial identity. We noted that although composites composed of the top and bottom sections of two people's faces no longer resemble either of the original faces, if both faces are posing the same expression (e.g., happiness), the composite face also looks happy. Similarly, inspection of the stimuli used in Experiment 1 showed that the opposite was true. That is, composites prepared from the top and bottom sections of different expressions posed by the same model were also highly identifiable as that particular model, although the composite facial expression resembled neither of the two starting expressions. This seemed to suggest that the composite effects for facial identity and facial expression were tapping two different types of configural processing. The results of Experiment 4 were consistent with this intuition.

Experiment 4 showed that either form of configural interference (identity or expression) can be produced from the same set of composite faces depending on whether participants are instructed to attend to the faces' identity or their expression. Thus, when asked to report the identity shown in the bottom half of a composite face, participants were significantly slower if the two halves contained different models' faces. Likewise, participants were slower to report facial expression if the two halves showed different expressions. More important, however, no significant cost was produced if the unattended attribute (e.g., expression in the identity task) was incongruent across the two face halves. Nor was there any additional cost when both attended and unattended attributes were incongruent relative to the condition in which the attended attribute alone was incongruent. These observations suggest that participants were encoding different types of configural information when processing facial identity and facial expression.

Configural Information for Facial Identity and Facial Expression

As we discussed in the introduction section, cues to facial expression and facial identity are generally thought to be processed by separate cognitive routes (Bruce & Young, 1986; Hay & Young, 1982; Young & Bruce, 1991; Young et al., 1993). It seems entirely plausible, then, that these parallel processing routes should use different types of visual information from the same facial image. What is slightly more contentious, however, is the idea that these two routes should process different types of configural information. However, in line with this idea, it is worth remembering that Diamond and Carey (1986) identified two forms of configural features, which they referred to as first-order and second-order relational properties.

The term first-order relational properties refers to the raw interfeature relationships that are common to all normal faces—two horizontally positioned eyes, above a central nose, above a central mouth, and so forth—which is effectively the spatial information that makes up a face. Second-order relational properties are substantially more subtle and are what are more generally referred to as simply configural features. As we discussed earlier, these features are the interrelationships between different feature positions and shapes that help distinguish one facial identity from all others (e.g., the distance between the eyes; position and shape of the nose in relation to the position and shape of the mouth, etc.). Furthermore, it is generally thought to be these second-order features that are disrupted by inversion and by the composite effect for facial identity. At first glance, then, it seems natural to assume that second-order features are also disrupted in the composite effect for facial affect shown here. But, as we have already discussed, this explanation is inconsistent with the observation that configural information for facial identity and facial expression can be selectively disrupted (Experiment 4). Instead, this finding points to the conclusion that the configural cues to these two facial attributes are different. Hence, one possibility is that the composite effect for facial expression may reflect a disruption of a more coarse form of configural information, one more akin to first-order relational properties.

As applied to facial identity by Diamond and Carey (1986), first-order relational properties are the interfeature relationships that are common to all faces (i.e., the average or prototype configuration associated with all faces one has encountered). For facial expressions, we suggest that there are the interfeature relationships that make a surprise expression surprised, or happiness expression happy, etc. In other words, we suggest that each emotional facial expression is associated with its own average configuration. The average configuration could be regarded as a distinct representation that is abstracted from encountered exemplars of each type of facial expression (happiness, sadness, anger, fear, disgust, surprise, etc.). Alternatively, it could be envisaged as the centroid of a cluster of stored exemplar representations, with separate clusters for each emotion

category. In other words, it is not necessary for the average to exist as a distinct prototype structure in its own right.

In relation to this discussion of average expression representations, it is worth remembering that Ekman and his colleagues have shown that each emotion category is associated with more than one facial structure. For example, in the case of surprise, the mouth could be anything between wide open and closed, or, in anger, the eyes can be narrowed or wide open. Nonetheless, the different variants of each expression contain enough common features for them to form a cluster around an average or prototype configuration.

We should point out that we do not wish to imply that facial expressions are coded in terms of configural information alone, and we have no fundamental objection to the idea that the individual features of facial expressions are also important for recognition (Ellison & Massaro, 1997). Hence, a facial expression, such as surprise, might be encoded in terms of its individual features (i.e., raised eyebrows, wide open eyes, and a wide open mouth) and in terms of its facial configuration (i.e., a symmetrical arrangement of raised eyebrows, above wide open eyes, above a wide open mouth). Consequently, when the eye and eyebrow regions from a surprise expression are aligned with the bottom half of a face displaying a different expression (e.g., one in which the upper lip is raised to signal disgust), the overall representation no longer resembles the average configuration for surprise (or disgust). Hence, although participants are able to use the individual features in the top (or bottom) half of the face to identify the expression, this process is slowed by the configural mismatch. In the noncomposite condition, however, there is no conflicting configural information because the face halves are misaligned. Hence, the participants can use the information in one face half to identify the emotion without experiencing interference from an unusual facial configuration.

Configural and Part-Based Models of Facial Expression Recognition: Weighing Up the Evidence

So where does the above discussion leave us in relation to Ellison and Massaro's (1997) largely part-based account of facial expression recognition? As we discussed in the introduction, these authors asked participants to identify whole-face expressions in which two different features (eyebrows and mouth corners) were manipulated. Participants were also presented with the individual features shown in the context of the upper and lower sections of the face. For each image, they were asked to make a binary decision response: Is the emotion expressed happiness or anger? By modeling the data using the fuzzy logical model of perception, Ellison and Massaro showed that participants' categorization of whole-facial expressions could be reliably modeled by assuming that the critical features of the face (eyebrow and mouth sections) are evaluated separately.

The composite paradigm is not dissimilar to Ellison and Massaro's (1997) task in that it also uses stimuli prepared by recombining the upper and lower sections of different facial features. Nonetheless, our data do not concur with their findings. It is relevant, then, that we consider why, but in doing so, it is important to recognize two points.

First, the FLMP model does not exclude the possibility that configural features are used to encode facial expressions (Ellison & Massaro, 1997; Massaro, 1998). The only constraint the FLMP makes on the information used to encode facial affect is that each feature must be evaluated as an independent perceptual unit. Hence, if one assumes that configural information can be encoded as one or more independent units, then our data are not at odds with the FLMP.

The second important point to take note of is that our own data do not exclude the possibility that part-based information is used for facial expression recognition. The data simply rule out the idea that configural information is not used for recognizing facial expressions. In actual fact, we think that our data actively support the suggestion that both configural and part-based information are used to decode facial affect; otherwise, the participants would have found it virtually impossible to identify the emotions in the upper or lower parts of the face, because with the exception of Experiment 3 (same-expression composites), the overall configuration did not match the emotion shown in either half of the face.

With these factors in mind, what are the differences between the two studies that may account for the different results? First, perhaps the main difference between the two studies is that Ellison and Massaro (1997) used identification rates and affect ratings as their dependent measures, whereas the composite paradigm uses RTs as the principal measure of interest. In relation to this point, it may be relevant that in our own series of experiments, a significant composite effect was found for the RT measure in all of the experiments, but only Experiment 3 showed a reliable composite effect for the error data, although there was no evidence of a speed–accuracy trade-off in the other experiments. Hence, it is possible that RTs provide a more sensitive measure of the configural contribution.

It is interesting that Ellison and Massaro (1997) reported that participants' RTs were longer for "ambiguous expressions" in their study; these included faces in which the top and bottom sections displayed different emotional signals (e.g., top = anger, bottom = happiness). Without a noncomposite condition, however, it is difficult to determine whether the longer RTs reflect interference between the different emotional concepts expressed in the two face halves or whether (as we have found) there was some additional interference from the inappropriate configuration.

The second point to take note of is that our experiments used a number of human models' faces from the Ekman and Friesen (1976) set with expressions associated with six basic emotions (happiness, sadness, anger, fear, disgust, and surprise). These images are based on an anatomical analysis of facial affect, and their repeated use in psychological studies verifies that the expressions are highly recognizable. A fact that is further substantiated by the preliminary experiment described in Experiment 1, this experiment also demonstrated that the upper and lower face sections used to

prepare the composites were reliably identified using a six-way, forced-choice task. Ellison and Massaro's (1997) study, on the other hand, used a single (computer-generated) synthetic face posing two facial expressions, anger and happiness. Hence, there were differences between the number of expressions and number of different examples of the expressions used in the two experiments.

To expand further on the last point, it may also be relevant that we used a three-way decision task, whereas Ellison and Massaro (1997) used a binary response task. One of the problems in using a binary decision task is that one cannot be sure that the participants spontaneously recognize the facial signals used as the intended emotions (e.g., happy and angry). For example, it is possible that participants might use a strategy of classifying the images as "happy" and "not happy." It is interesting that previous studies that have applied the composite effect to the recognition of facial identity have used a vocal response task with at least four response options (Carey & Diamond, 1994; Young et al., 1987). Hence, it is possible that by making the task more demanding (by increasing the number of response options), one can increase the paradigm's sensitivity to configural interference.

We also feel that it worth emphasizing once again that the composite paradigm is particularly suited to differentiating between part-based and configural accounts of facial expression recognition, because the same part-based information is present in both composite and noncomposite conditions. This means that for the two conditions, any interference between the emotional concepts expressed in the two separate halves should be constant. Hence, any difference in the RTs between the composite and noncomposite conditions would appear to reflect the difference in coding a facial (composite) as opposed to a nonfacial (noncomposite) image. In this sense, the data speak for themselves: RTs are slower when the two halves are aligned to form a facial expression configuration (composite condition) than when they are misaligned and there is no facial expression configuration (noncomposite condition). Having demonstrated this finding in a number of experiments, we conclude that these data are consistent with a configural model of facial affect recognition in which both configural and part-based information is used to identify the emotion. The use of part-based information is further substantiated by the findings of Ellison and Massaro (1997), and by our own observations that recognition of expressions of partial faces does not fall to chance level.

As we have already emphasized, this conclusion is not inconsistent with the FLMP, provided that one assumes the configural information can be evaluated as one or more independent perceptual units. To successfully demonstrate that this is the case, however, one would first have to identify the important configural features and then assess their contribution using the sort of extended factorial design that has become associated with FLMP research (Massaro, 1998). Given that 20 years of research into configural coding of facial identity have failed to identify the precise nature of the configural information for facial identity, this seems a tall order for facial expression research.

Finally, it worth noting that our results are also in line with a recent model facial expression production outlined by Smith and Scott (1997). These authors suggested that each emotional facial expression is made up of a number of individual features (or components) and that at least some of these features are in themselves meaningful. However, they also suggested that when the individual features are produced in combination (i.e., in the form of a facial expression of anger, disgust, or sadness, etc.), the overall facial configuration may convey additional information that is not captured by the individual features themselves. In other words, these authors suggested that for facial expression production, the whole is more than the sum of the parts. Similarly, we think that our own results demonstrate that when faced with a facial expression image, the perceptual system not only analyzes cues present in individual features but also the configuration of interrelationships between these features.

References

Bassili, J. N. (1979). Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face. *Journal of Personality and Social Psychology, 37,* 2049–2058.

Bruce, V. (1988). *Recognising faces.* London: Erlbaum.

Bruce, V., Doyle, Y., Dench, N., & Burton, M. (1991). Remembering facial configurations. *Cognition, 38,* 109–144.

Bruce, V., & Langton, S. (1994). The use of pigmentation and shading information in recognising the sex and identity of faces. *Perception, 1994,* 803–822.

Bruce, V., & Young, A. W. (1986). Understanding face recognition. *British Journal of Psychology, 77,* 305–327.

Calder, A. J., Young, A. W., Perrett, D. I., Etcoff, N. L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition, 3,* 81–117.

Calder, A. J., Young, A. W., Rowland, D., & Perrett, D. I. (1997). Computer-enhanced emotion in facial expressions. *Proceedings of the Royal Society London, 264,* 919–925.

Campbell, R., Brooks, B., de Haan, E., & Roberts, T. (1996). Dissociating face processing skills: Decisions about lip-read speech, expression and identity. *Quarterly Journal of Experimental Psychology, 49A,* 295–314.

Carey, S., & Diamond, R. (1977). From piecemeal to configurational representation of faces. *Science, 195,* 312–314.

Carey, S., & Diamond, R. (1994). Are faces perceived as configurations more by adults than by children? *Visual Cognition, 1,* 253–274.

Coleman, J. C. (1949). Facial expressions of emotion. *Genetic Psychology Monographs, 63*(1), Whole No. 296.

Diamond, R., & Carey, S. (1986). Why faces are and are not special: An effect of expertise. *Journal of Experimental Psychology: General, 115,* 107–117.

Dunlap, K. (1927). The role of eye-muscles and mouth-muscles in the expression of the emotions. *Genetic Psychology Monographs, 2,* 199–233.

Ekman, P. (1972). Universals and cultural differences in facial expressions of emotion. In J. K. Cole (Ed.), *Nebraska Symposium on Motivation* (pp. 207–283). Lincoln: University of Nebraska Press.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues.* Englewood Cliffs, NJ: Prentice Hall.

Ekman, P., & Friesen, W. V. (1976). *Pictures of facial affect.* Palo Alto, California: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). *Emotion and the human face: Guidelines for research and an integration of findings.* New York: Pergamon Press.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., Ayhan LeCompte, W., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., & Tomita, M. (1987). Universals and cultural differences in the judgement of facial expressions of emotion. *Journal of Personality and Social Psychology, 53,* 712–717.

Ellison, J. W., & Massaro, D. W. (1997). Featural evaluation, integration, and judgment of facial affect. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 213–226.

Endo, M., Masame, K., & Maruyama, K. (1989). Interference from configuration of a schematic face onto the recognition of its constituent parts. *Tohoku Psychologica Folia, 48,* 97–106.

Endo, M., Takahashi, K., & Maruyama, K. (1984). Effects of observer's attitude on the familiarity of faces: Using the difference in cue value between central and peripheral facial elements as an index of familiarity. *Tohoku Psychologica Folia, 43,* 23–34.

Etcoff, N. L. (1984). Selective attention to facial identity and facial emotion. *Neuropsychologia, 22,* 281–295.

Etcoff, N. L., & Magee, J. J. (1992). Categorical perception of facial expressions. *Cognition, 44,* 227–240.

Farah, M. J., Tanaka, J. W., & Drain, H. M. (1995). What causes the inversion effect? *Journal of Experimental Psychology: Human Perception and Performance, 21,* 628–634.

Frois-Wittmann, J. (1930). The judgement of facial expression. *Journal of Experimental Psychology, 13,* 113–151.

George, M. S., Ketter, T. A., Gill, D. S., Haxby, J. V., Ungerleider, L. G., Herscovitch, P., & Post, R. M. (1993). Brain regions involved in recognizing facial emotion or identity: An oxygen-15 PET study. *Journal of Neuropsychiatry, 5,* 384–394.

Haig, N. D. (1984). The effect of feature displacement on face recognition. *Perception, 13,* 505–512.

Hanawalt, N. G. (1944). The role of the upper and lower parts of the face as the basis for judging facial expressions: II. In posed expressions and "candid camera" pictures. *Journal of General Psychology, 31,* 23–36.

Hasselmo, M. E., Rolls, E. T., & Baylis, G. C. (1989). The role of expression and identity in face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research, 32,* 203–218.

Hay, D. C., & Young, A. W. (1982). The human face. In A. W. Ellis (Ed.), *Normality and pathology in cognitive functions* (pp. 173–202). London: Academic Press.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 986–1004.

Hole, G. J. (1994). Configurational factors in the perception of unfamiliar faces. *Perception, 23,* 65–74.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioural principle.* Cambridge, MA: MIT Press.

Massaro, D. W., & Cohen, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science, 1,* 55–63.

McKelvie, S. J. (1995). Emotional expression in upside-down faces: Evidence for configurational and componential processing. *British Journal of Social Psychology, 34,* 325–334.

Parks, T. E., Coss, R. G., & Coss, C. S. (1985). Thatcher and the Cheshire cat: Context and the processing of facial features. *Perception, 14,* 747–754.

Parry, F. M., Young, A. W., Saul, J. S. M., & Moss, A. (1991). Dissociable face processing impairments after brain injury. *Journal of Clinical and Experimental Neuropsychology, 13,* 545–558.

Plutchik, R. (1962). *The emotions: Facts theories and a new model.* New York: Random House.

Rhodes, G. (1988). Looking at faces: First-order and second-order features as determinants of facial appearance. *Perception, 17,* 43–63.

Rhodes, G., Brennan, S. E., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology, 19,* 473–497.

Sergent, J., Ohta, S., MacDonald, B., & Zuck, E. (1994). Segregated processing of emotional identity and emotion in the human brain: A PET study. *Visual Cognition, 1,* 349–369.

Shepherd, J. W., Davies, G. M., & Ellis, H. D. (1981). Studies of cue saliency. In G. M. Davies, H. D. Ellis, & J. Shepherd (Eds.), *Perceiving and remembering faces* (pp. 105–131). London: Academic Press.

Smith, C. A., & Scott, H. S. (1997). A componential approach to the meaning of facial expressions. In J. A. Russell & J. M. Fernandez-Dols (Eds.), *The psychology of facial expressions* (pp. 229–254). Cambridge, England: Cambridge University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18,* 643–662.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46A,* 225–245.

Valentine, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology, 79,* 471–491.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43A,* 161–204.

Valentine, T., & Bruce, V. (1988). Mental rotation of faces: Memory and Cognition. *Memory and Cognition, 16,* 556–566.

Wallbott, H. G., & Ricci-Bitti, P. (1993). Decoders' processing of emotional facial expression—A top-down or bottom-up mechanism. *European Journal of Social Psychology, 23,* 427–443.

Young, A. W., & Bruce, V. (1991). Perceptual categories and the computation of "grandmother." *European Journal of Cognitive Psychology, 3,* 5–49.

Young, A. W., Hellawell, D., & Hay, D. C. (1987). Configurational information in face perception. *Perception, 16,* 747–759.

Young, A. W., McWeeny, K. H., Hay, D. C., & Ellis, A. W. (1986). Matching familiar and unfamiliar faces on identity and expression. *Psychological Research, 48,* 63–68.

Young, A. W., Newcombe, F., de Haan, E. H. F., Small, M., & Hay, D. C. (1993). Face perception after brain injury: Selective impairments affecting identity and expression. *Brain, 116,* 941–959.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Megamixing facial expressions. *Cognition, 63,* 271–313.

Appendix

Experimental Faces: Identifier in Ekman and Friesen's (1976) Series and Percentage Recognition as This Emotion in Their Norms

Experiment 1: Identification of the Top and Bottom Sections of Ekman and Friesen (1976) Faces

Happiness (M = 99%). 7 C-2-18; 14 EM-4-07; 34 JJ-4-07; 48 MF-1-06; 57 MO-1-04; 66 NR-1-06; 74 PE-2-12; 85 PF-1-06; 93 SW-3-09; 101 WF-2-12.

Surprise (M = 91%). 11 C-1-10; 19 EM-2-11; 39 JJ-4-13; 54 MF-1-09; 63 MO-1-14; 70 NR-1-14; 81 PE-6-02; 90 PF-1-16; 97 SW-1-16; 107 WF-2-16.

Fear (M = 90%). 9 C-1-23; 16 EM-5-21; 37 JJ-5-13; 50 MF-1-26; 59 MO-1-23; 68 NR-1-19; 79 PE-3-21; 88 PF-2-30; 95 SW-2-30; 104 WF-3-16.

Sadness (M = 90%). 8 C-1-18; 15 EM-4-24; 36 JJ-5-05; 49 MF-1-30; 58 MO-1-30; 67 NR-2-15; 75 PE-2-31; 86 PF-2-12; 94 SW-2-16; 102 WF-3-28.

Disgust (M = 93%). 12 C-1-04; 20 EM-4-17; 40 JJ-3-20; 55 MF-2-13; 64 MO-2-18; 71 NR-3-29; 82 PE-4-05; 91 PF-1-24; 98 SW-1-30; 108 WF-3-11.

Anger (M = 90%). 10 C-2-12; 18 EM-5-14; 38 JJ-3-12; 53 MF-2-07; 61 MO-2-11; 69 NR-2-07; 80 PE-2-21; 89 PF-2-04; 96 SW-4-09; 105 WF-3-01.

Experiments 1, 2, and 3

Happiness (M = 98%). 7 C-2-18; NR-1-06; 85 PF-1-06; 93 SW-3-09.

Surprise (M = 92%). 11 C-1-10; 70 NR-1-14; 90 PF-1-16; 97 SW-1-16.

Fear (M = 88%). 9 C-1-23; 68 NR-1-19; 88 PF-2-30; 95 SW-2-30.

Sadness (M = 94%). 8 C-1-18; 67 NR-2-15; 86 PF-2-12; 94 SW-2-16.

Disgust (M = 95%). 12 C-1-04; 71 NR-3-29; 91 PF-1-24; 98 SW-1-30.

Anger (M = 88%). 10 C-2-12; 69 NR-2-07; 89 PF-2-04; 96 SW-4-09.

Experiment 4

Happiness (M = 97%). 7 C-2-18; NR-1-06; 93 SW-3-09.

Surprise (M = 92%). 11 C-1-10; 70 NR-1-14; 97 SW-1-16.

Disgust (M = 94%). 12 C-1-04; 71 NR-3-29; 98 SW-1-30.

Received May 6, 1998
Revision received January 12, 1999
Accepted April 20, 1999 ■

Face Aftereffects Predict Individual Differences in Face Recognition Ability

Article *in* Psychological Science · October 2012

DOI: 10.1177/0956797612446350 · Source: PubMed

4 authors, including:

Elinor Mckone
Australian National University
109 PUBLICATIONS **4,884** CITATIONS

SEE PROFILE

Mark Edwards
Australian National University
106 PUBLICATIONS **2,060** CITATIONS

SEE PROFILE

Tirta Susilo
Victoria University of Wellington
40 PUBLICATIONS **772** CITATIONS

SEE PROFILE

Face Aftereffects Predict Individual Differences in Face Recognition Ability

Hugh W. Dennett, Elinor McKone, Mark Edwards and Tirta Susilo

Psychological Science published online 16 October 2012
DOI: 10.1177/0956797612446350

Psychological Science OnlineFirst, published on October 16, 2012 as doi:10.1177/0956797612446350

Research Report

Psychological Science
XX(X) 1–9

DOI: 10.1177/0956797612446350
http://pss.sagepub.com

Face Aftereffects Predict Individual Differences in Face Recognition Ability

Hugh W. Dennett[1], Elinor McKone[1,2], Mark Edwards[1], and Tirta Susilo[1,3]

[1]Department of Psychology, Australian National University; [2]ARC Centre of Excellence in Cognition and its Disorders, Australian National University; and [3]Department of Psychological and Brain Sciences, Dartmouth College

Abstract

Face aftereffects are widely studied on the assumption that they provide a useful tool for investigating face-space coding of identity. However, a long-standing issue concerns the extent to which face aftereffects originate in face-level processes as opposed to earlier stages of visual processing. For example, some recent studies failed to find atypical face aftereffects in individuals with clinically poor face recognition. We show that in individuals within the normal range of face recognition abilities, there is an association between face memory ability and a figural face aftereffect that is argued to reflect the steepness of broadband-opponent neural response functions in underlying face-space. We further show that this correlation arises from face-level processing, by reporting results of tests of nonface memory and nonface aftereffects. We conclude that face aftereffects can tap high-level face-space, and that face-space coding differs in quality between individuals and contributes to face recognition ability.

Keywords

face perception, individual differences, cognitive ability, visual memory, performance

Received 12/12/11; Revision accepted 3/6/12

Most humans have a remarkable ability to recognize faces, although there are surprisingly large individual differences in this ability (Bowles et al., 2009; Wilmer et al., 2010). In the study reported here, we investigated whether these individual differences might be partially attributable to the quality of face-space coding (Fig. 1), as measured using figural face aftereffects (Fig. 2a). It has been argued that face-space facilitates individuation of faces (Valentine, 1991), and the widespread investigation of face aftereffects is based on the common assumption that they reflect face-space coding (Leopold, O'Toole, Vetter, & Blanz, 2001; Nishimura, Doyle, Humphreys, & Behrmann, 2010; Rhodes & Jeffery, 2006; Robbins, McKone, & Edwards, 2007; Webster & MacLin, 1999). If these assumptions are correct, there should be a relationship between face aftereffects and face recognition ability, because of their common origin in face-space coding.

However, the degree to which face aftereffects originate in face-level coding has been a long-standing issue in the literature (Rhodes & Leopold, 2011; Webster & MacLin, 1999). Several studies have shown that face aftereffects can partly originate in low- and mid-level stages of the visual stream (Afraz & Cavanagh, 2008; Dickinson, Almeida, Bell, & Badcock, 2010; Susilo, McKone, & Edwards, 2010a). Moreover, two studies failed to show that face aftereffects are related to face recognition ability: These studies found normal face aftereffects in individuals who could not recognize faces because of their developmental prosopagnosia (DP; $N = 6$ in Nishimura et al., 2010; $N = 1$ in Susilo et al., 2011).[1] If face aftereffects arise even partly from face-space processes, and face-space is important for face recognition, then how could such individuals exhibit normal face aftereffects? We see two possibilities, both of which informed the design of our present study.

First, although face-space is likely coded in posterior face areas (Freiwald, Tsao, & Livingstone, 2009; Loffler, Yourganov, Wilkinson, & Wilson, 2005), the problem in some individuals with DP appears to be not in these areas but instead in weak connections from these areas to anterior face areas (Thomas et al., 2009). This means that failure to find abnormal aftereffects in individuals with DP does not rule out an association between face-space coding and face recognition within the normal population, in whom the forward connections are

Corresponding Author:
Hugh W. Dennett, Department of Psychology (Building 39), The Australian National University, Canberra ACT 0200, Australia
E-mail: hugh.dennett@anu.edu.au or hughdennett@gmail.com

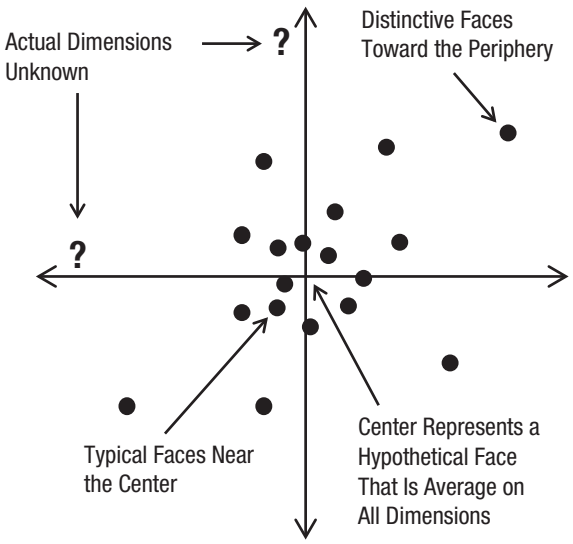


Fig. 1. Face-space coding. Each individual face is coded as a point in a multidimensional perceptual space that has dimensions corresponding to attributes that vary across faces and that has the average face at the center (Valentine, 1991). Face aftereffects are commonly interpreted as arising from a shift in the location of this average.

intact. Thus, in the present study, we tested only individuals in the normal range of face recognition ability.

Second, certain types of face aftereffects might be more effective at capturing face-space processes than others. A group analysis of 14 individuals with DP (Palermo, Rivolta, Wilson, & Jeffery, 2011) revealed a normal-sized aftereffect for a *figural* manipulation in which an expanded-face adaptor causes a different undistorted face to appear contracted, but an impaired aftereffect for an *identity* manipulation, in which adaptation to one person's face (e.g., "Dan") causes the average face to be perceived as resembling an individual opposite to the adaptor face on all face attributes (i.e., "anti-Dan"). Palermo et al. accounted for this difference by proposing that the identity aftereffect taps face-specific processes to a greater extent than does the more shape-generic expansion-contraction aftereffect. Thus, in the present study, we tested participants using a particular type of figural aftereffect (manipulation of eye height; Fig. 2a) that has previously been shown to have a substantial face-specific component (Susilo et al., 2010a).

Our basic question was whether, within the normal range, face recognition ability as measured using the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) correlates with the magnitude of the eye-height aftereffect. Researchers (Nishimura et al., 2010; Palermo et al., 2011; Pellicano, Jeffery, Burr, & Rhodes, 2007) have implicitly assumed that the direction of the correlation should be positive—that is, that poorer face-space coding (in clinical conditions) should be associated with a smaller aftereffect. However, there has been no explicit rationale given for assuming this direction. We chose to study the eye-height aftereffect because recent evidence regarding its neural coding provides an empirical rationale for a positive correlation (Susilo, McKone, & Edwards, 2010b).

The relevant neural coding properties (Fig. 3) are broadband-opponent (two-pool) coding and linear response functions. In opponent coding (a neural implementation of norm-based coding), one pool of cells responds maximally to one end of the attribute dimension (e.g., low eyes), whereas the other responds maximally to the opposite end (e.g., high eyes). (Note that the low-eye and high-eye pools should be thought of not as pools of eye-height detectors, but rather as slices through each neuron's multidimensional response profile; individual face cells respond

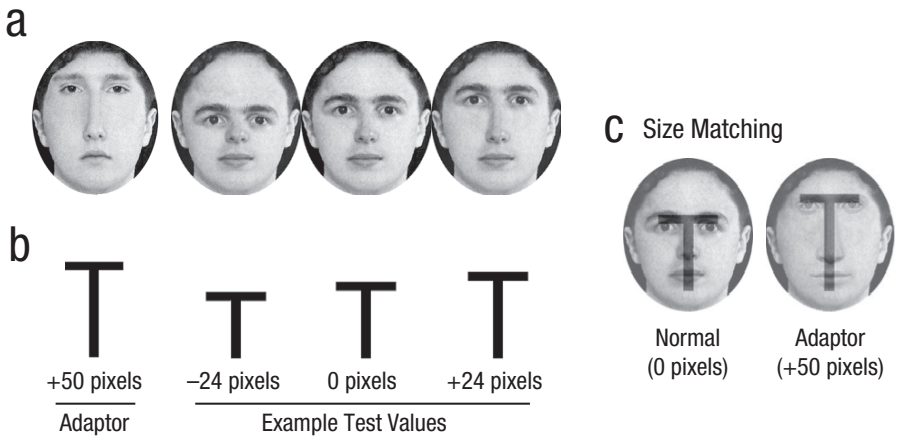


Fig. 2. The eye-height and T-shape aftereffects. In the eye-height aftereffect (a), adaptation to a distorted face in which the eyes are higher than in the original face (the +50-pixel deviation we used in our adaptors is shown here) makes the eyes in test faces (including the undistorted face with 0-pixel deviation) appear to be lower than they appeared before adaptation. In the T-shape aftereffect (b), adaptation to a T-shape with the horizontal bar shifted upward (the +50-pixel deviation we used in our adaptors is shown here) makes the bar in test T-shapes appear to be lower than it appeared before adaptation (Susilo, McKone, & Edwards, 2010a). In our study, the T-shapes used as stimuli for measuring the T-shape aftereffect were matched in size to the T-shaped region of the faces used to measure the eye-height aftereffect (c).

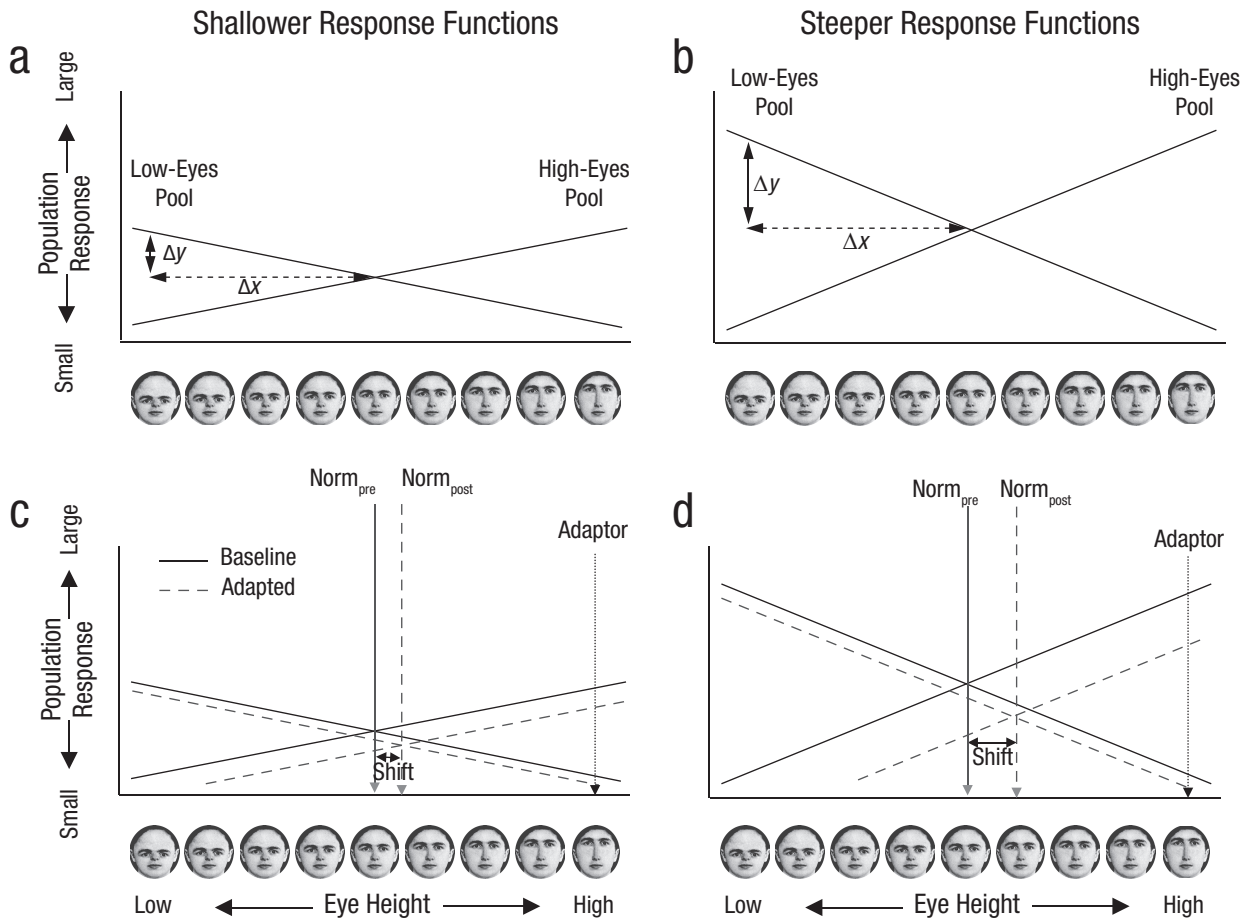


Fig. 3. Basis of the prediction that better face recognition should be positively correlated with larger eye-height aftereffects. Eye height is coded by the comparative activation of two pools of neurons (each with linear response): one that responds maximally to low eyes and one that responds maximally to high eyes. The eye height that elicits equal responses in the two pools would be perceived as normal or average. As illustrated in (a) and (b), steeper response functions of these pools mean that a unit change in eye height (Δx) would elicit a greater change in the population response (Δy). Thus, steeper response functions should yield better discrimination of eye height. As illustrated in (c) and (d), postadaptation responses (dashed lines) are lowered relative to preadaptation responses (solid lines) in proportion to the preadaptation response (Maddess, McCourt, Blakeslee, & Cunningham, 1988; Movshon & Lennie, 1979); this means that adapting to any eye height except for the norm would cause differential adaptation in the two pools, and hence cause the face perceived as normal before adaptation ($Norm_{pre}$) to shift toward the adaptor after adaptation ($Norm_{post}$). Steep functions would cause a larger shift than shallow functions because the initial difference in response between the two pools would be larger for steeper functions. Thus, steeper response functions predict both better discrimination of eye height and a larger eye-height aftereffect, and there should therefore be a positive correlation between face discriminability and the magnitude of the aftereffect.

to several face attributes; Freiwald et al., 2009.) Linear opponent coding is supported by neurophysiological evidence, which has revealed face-selective cells in monkeys with linear ramp-shaped response functions for many face attributes (Freiwald et al., 2009). Psychophysical evidence also supports this type of coding specifically for eye height (Robbins et al., 2007; Susilo et al., 2010b); moreover, the response functions remain linear across the full range of eye heights up to eyes approaching the hairline (Susilo et al., 2010b).

Together, these properties predict a positive correlation between the size of the aftereffect and ability to recognize faces, because both the aftereffect and discrimination ability derive from the slope of the individual's response functions. Steeper response functions should yield better face recognition because steeper slopes produce better discrimination of a unit change in eye height (Figs. 3a and 3b); steeper response functions also should yield larger eye-height aftereffects because the eye height perceived as normal will shift more after adaptation (Figs. 3c and 3d). Further, the linearity across the full range of eye heights (Susilo et al., 2010b) means that one can test for the predicted correlation using only one eye-height distortion in the adaptors. We used adaptors with very high eyes (Fig. 2a) because in broadband-opponent coding, adaptors furthest from the norm elicit the largest aftereffects (Fig. 3), and thus maximize the potential to observe individual differences in aftereffect magnitude.

We included two nonface control tasks in our study. The first was memory for cars (Cambridge Car Memory Test, or CCMT; Dennett et al., 2012). The second was a task measuring a T-shape aftereffect (Fig. 2b; Susilo et al., 2010a). The

T-shape task was designed to capture the shape-generic component of the eye-height manipulation by using a letter *T* matched to the T-shaped region of the face formed by the eyes, nose, and mouth. These control tasks allowed us to assess the extent to which any correlation between face aftereffects and face recognition arose specifically from face-level coding.

Method

Participants

Participants received course credit or were paid $30. To ensure that we were not testing individuals with prosopagnosia, we excluded 7 participants with CFMT scores in the lowest 5% of the population (using norms from 248 young adult Australians; McKone et al., 2011). We excluded an additional 5 participants whose data from the adaptation tasks had poor psychometric fits (see the section on curve fitting), as well as 2 participants who were extreme univariate outliers ($z > 3.32$) on the adaptation tasks. The final sample consisted of 78 participants (48 female, 30 male; ages 18–45 years, $M = 20.69$, $SD = 5.34$). All either were Caucasian ($n = 75$) or had very high Caucasian exposure (i.e., had one Caucasian parent and were raised in Australia; $n = 3$).

Tasks

For the CFMT, the method was as described in Duchaine and Nakayama (2006). Briefly, participants learned six Caucasian male faces—each in three views, to encourage face rather than image learning. Participants later discriminated these targets from similar-looking distractor faces (untimed three-alternative, forced-choice task, with simultaneous presentation of the faces; Fig. 4a). The CFMT has good psychometric properties and produces large individual differences (Bowles et al., 2009; Wilmer et al., 2010).

For the face eye-height adaptation task, the method was as in Susilo et al. (2010a, 2010b). In the preadaptation phase (348 trials), participants viewed faces that varied in eye height (29 levels ranging from –24 pixels to +24 pixels; negative numbers indicate eyes shifted down from the unaltered "zero" face, and positive numbers indicate eyes shifted up from the unaltered face; Fig. 2a). Participants indicated whether the eyes

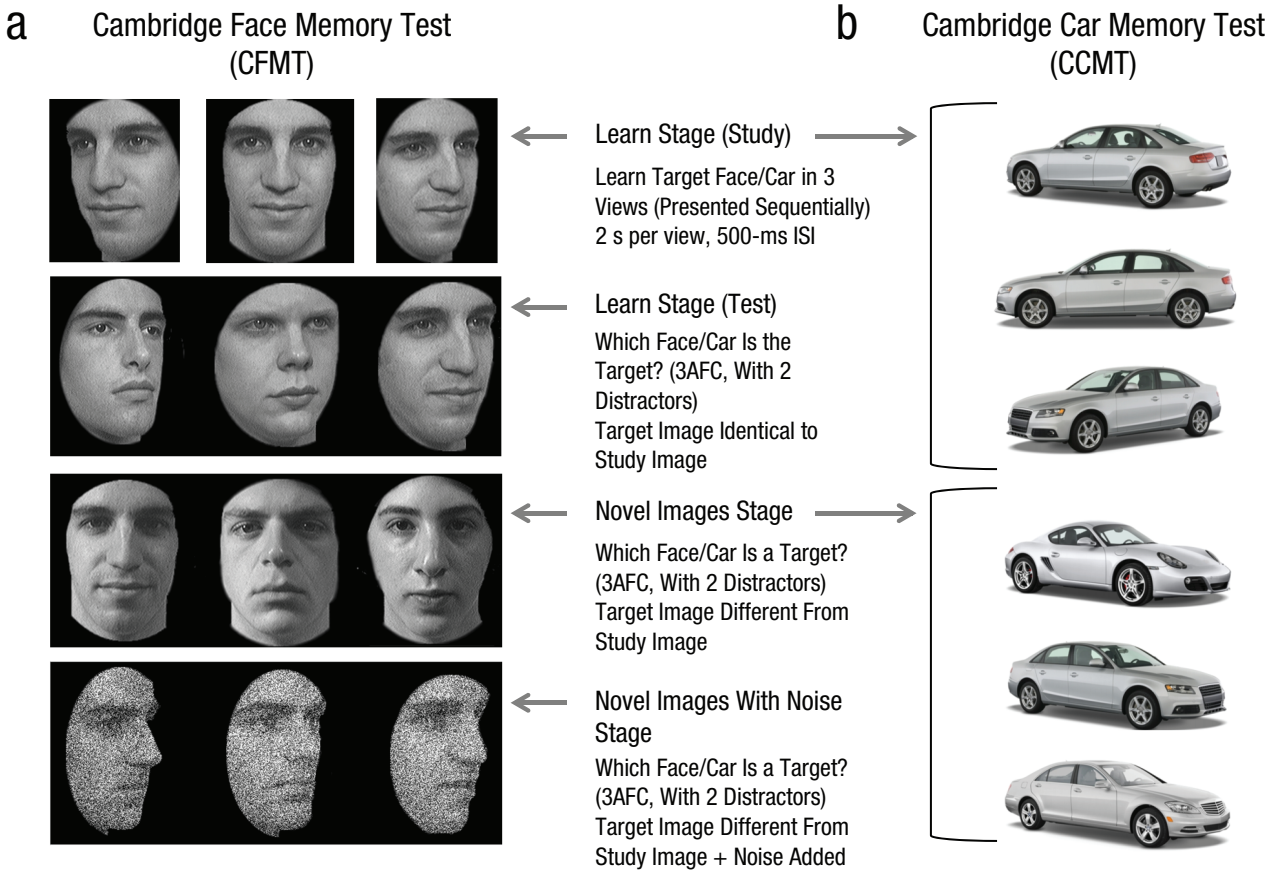


Fig. 4. Illustration of (a) the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) and (b) the Cambridge Car Memory Test (CCMT; Dennett et al., 2012). These tests are very similar, differing only in the stimulus category. For the CFMT, the figure illustrates three stages: Learn (including study and test trials), Novel Images, and Novel Images With Noise. For the CCMT, the figure illustrates only the Learn stage (study trial) and Novel Images stage. In the Learn stage, participants learn target faces or cars and are tested on recognition of images identical to the learned targets; in the Novel Images and Novel Images With Noise stages, participants are tested on recognition of targets in novel views, novel lighting, or both. 3AFC = three-alternative forced choice; ISI = interstimulus interval.

were "high" or "low" relative to their idea of a normal face. The postadaptation phase was the same except that each test face was preceded by a 4,000-ms adaptor face with an eye height of +50 pixels (Fig. 2a). Adaptor faces were smaller than test faces, to minimize contributions to the aftereffect from retinotopic low-level vision.

For the CCMT, the method was as described in Dennett et al. (2012). The CCMT has the same structure, format, and scoring as the CFMT, but the stimuli are cars instead of faces (Fig. 4b).

For the T-shape adaptation task, the method was as in Susilo et al. (2010a). This task matched the eye-height adaptation task in method, except that the adaptors and test stimuli were T-shapes, matched in size to the T-shaped eyes-nose-mouth region of the faces in the face-height task (Fig. 2c).

Curve fitting and calculation of aftereffects (eye-height and T-shape tasks)

Psychometric functions were fitted to the data from the adaptation tasks (details in Susilo et al., 2010b) to determine the point of subjective equality (PSE; see Fig. 5 for an example), that is, the eye height or T-shape that each observer perceived as being most normal, before and after adaptation. Observers with poor fits ($R^2 < .8$ averaged across the pre- and postadaptation phases, resulting in an unreliable shift score) were excluded. For the 78 participants in the final sample, the mean R^2 across all face fits was .92 (SD = .04), and the mean R^2 across all T-shape fits was .91 (SD = .05).

Aftereffect magnitude was calculated as the difference (in pixels) between each participant's postadaptation PSE and his or her preadaptation PSE (postadaptation minus preadaptation), expressed as a percentage of the distance of the adaptor from the participant's preadaptation norm (Fig. 5). This measure was used because there were individual differences in the preadaptation norm: Although the mean preadaptation PSEs were close to zero, there was noticeable spread around the means (see the standard deviations in Table 1).

Results

Table 1 shows that, as required for individual differences studies (Wilmer, 2008), all tasks had high reliability, means well away from ceiling and floor, and large standard deviations (i.e., wide spread of scores). All tasks also had scores that were normally distributed (Kolmogorov-Smirnov tests, all ps > .05). There were no multivariate outliers. For all correlations reported in this section, N was equal to 78.

The first key finding was that face aftereffects correlated with face recognition abilities, in the predicted direction: There was a significant positive correlation between the magnitude of the eye-height aftereffect and face memory (Fig. 6).

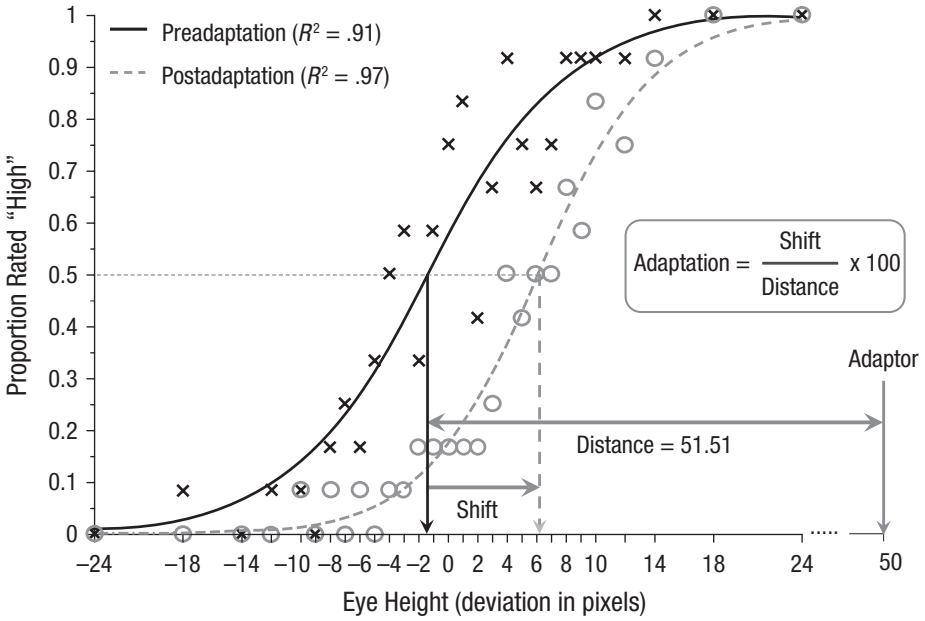


Fig. 5. Example of curve fitting and calculation of the eye-height aftereffect for a participant. The graph shows the proportion of faces that the participant rated as having "high" eyes as a function of eye height (deviation from the undistorted face; positive = up, negative = down). From the psychometric curves fitted to each participant's data, the locations of the preadaptation norm (solid black arrow) and the postadaptation norm (dashed arrow) were determined. Adaptation was calculated as the shift in the norm divided by the distance between the preadaptation norm and the adaptor face, multiplied by 100. As illustrated by this example, the distance of the adaptor from the preadaptation norm was not exactly 50 pixels for every participant, because of individual variation in the preadaptation norm.

Table 1. Descriptive Statistics for All Variables (N = 78)

Variable	Reliability	Mean	*SD*	Minimum	Maximum
Eye-height task: preadaptation PSE	.93	–0.18	3.73	–10.66	15.07
Eye-height task: postadaptation PSE	.96	3.98	5.19	–7.75	19.07
T-shape task: preadaptation PSE	.91	–6.66	4.15	–18.86	1.80
T-shape task: postadaptation PSE	.96	–1.64	5.46	–20.86	11.30
Eye-height aftereffect	.86	8.31	7.60	–6.30	33.40
T-shape aftereffect	.89	8.76	8.13	–11.10	30.60
CFMT	.85	79.02	10.95	58.33	100.00
CCMT	.83	74.25	11.36	47.22	98.61

Note: Points of subjective equality (PSEs) are expressed as the deviation (in number of pixels) from the zero (undistorted) face or T-shape. Eye-height and T-shape aftereffects are expressed as the shift in PSE as a percentage of the distance to the adaptor from the preadaptation norm. Results for the Cambridge Face Memory Test (CFMT; Duchaine & Nakayama, 2006) and Cambridge Car Memory Test (CCMT; Dennett et al., 2012) are reported as the percentage correct, out of 72 trials. For these two tests, the reported reliabilities are Cronbach's alphas; all other reliabilities are Spearman-Brown corrected split-half correlations. CFMT scores in this table should not be used as test norms because individuals in the lowest-scoring 5% of the population have been excluded.

The second key finding was that this correlation was specific to faces. If it arose from shape-generic processes—for example, if individuals with larger face aftereffects simply had better memory, and larger aftereffects, for all shapes—scores for all tasks should have correlated positively with each other. This was not the case.

First, despite the strong physical similarity of the T-shape manipulation to the eye-height manipulation, the two aftereffects were uncorrelated, r = –.02, p = .90, 95% confidence interval (CI): [–.24, .21]. Thus, it was not the case that some participants were generically "more adaptable" than others. Second, the face aftereffect was uncorrelated with car memory

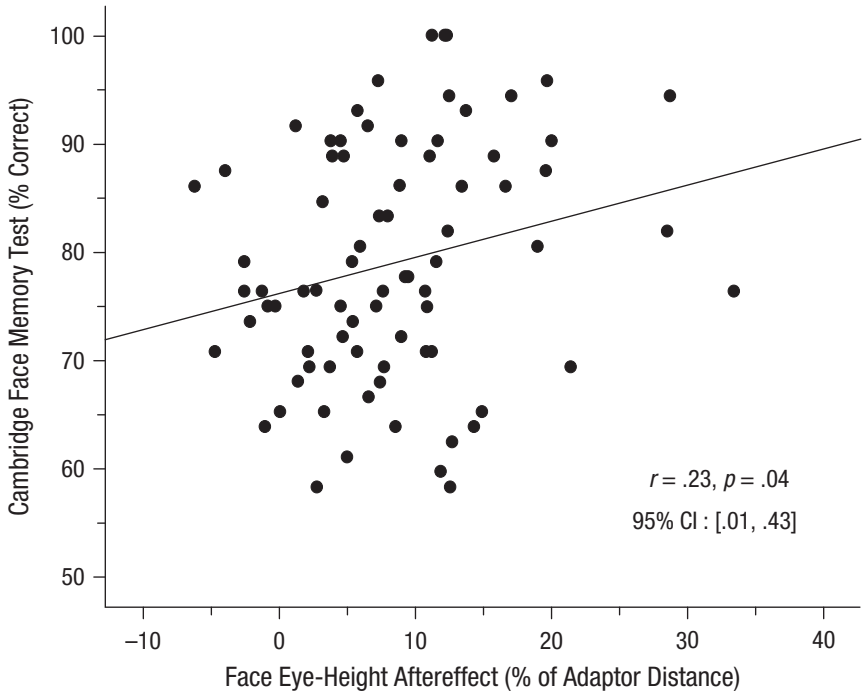


Fig. 6. Scatter plot (with best-fitting regression line) illustrating the Pearson correlation (r) between face recognition ability, as measured using the Cambridge Face Memory Test (Duchaine & Nakayama, 2006), and the magnitude of the face eye-height aftereffect. 95% CI = 95% confidence interval on the correlation value.

(CCMT), $r = .04$, $p = .76$, 95% CI: [–.19, .26]. This shows that the face aftereffect did not predict within-class object recognition memory generally, but predicted only face memory. Third, although there was a significant correlation between the T-shape aftereffect and face memory (CFMT), $r = –.25$, $p = .03$, 95% CI: [–.45, –.03], this correlation was negative, which means it cannot provide a shape-generic explanation for the positive association between magnitude of the face aftereffect and face memory. Finally, multiple regression revealed that the face aftereffect was a predictor of unique variance in face memory (CFMT). When the two nonface variables were added as predictors in the model, the negative relationship between magnitude of the T-shape aftereffect and CFMT was reduced to nonsignificance, $\beta = –0.16$, $t(74) = 1.58$, $p = .12$, and a significant relationship between car memory and face memory was revealed, $\beta = 0.39$, $t(74) = 3.81$, $p < .001$. Crucially, however, addition of these variables had no effect on the relationship between magnitude of the face aftereffect and CFMT, $\beta = 0.21$, $t(74) = 2.13$, $p = .04$ (cf. $r = .23$ for the simple bivariate correlation). These multiple regression results show that there was some overlap in CFMT variance explained by the two nonface tasks (T-shape aftereffect and CCMT), and that the variance in CFMT performance explained by the face aftereffect did not overlap with that explained by either of the nonface tasks.

Discussion

These results provide the first empirical evidence that individual differences in the quality of face-space coding exist, and that these contribute to individual differences in face recognition ability. The results support continued use of face aftereffects as a tool to investigate face-space. They further indicate that a figural (not just identity) face aftereffect can tap face-space (cf. Palermo et al., 2011). Finally, these results support a key prediction of a broadband-opponent (two-pool) face-space, namely, that steeper neural response functions should be associated with better face recognition ability.

Note that our results do not imply that all face aftereffects can necessarily be used as a tool to investigate face-space. Rather, several factors will affect the correlation between face aftereffects and face recognition. First, only face aftereffects that have a significant face-level component are suitable for investigating face-space. For eye height, the aftereffect has been argued to derive approximately 50% from face-level processes and 50% from shape-generic processes (Susilo et al., 2010a). Other types of figural aftereffects, however, might have a smaller face-level component, and would therefore be expected to have weaker relationships with face recognition ability. This may explain results showing normal expansion-contraction aftereffects in individuals with DP (Palermo et al., 2011).

Second, it may be that the correlation with face recognition depends on a direct link between aftereffect size and the slope of neural response functions in face-space. Linear functions allow use of a single adaptor value to measure response slope (Fig. 3), and there is evidence supporting linearity of eye-height coding (Susilo et al., 2010b). However, little is known about the shape of neural response functions underlying neural coding of other face attributes, and some findings suggest nonlinearity (Dakin & Omigie, 2009; Tanaka & Corneille, 2007; Wilson, Loffler, & Wilkinson, 2002; for discussion, see Susilo et al., 2010b). In the case of nonlinear functions, the aftereffect for a single adaptor value would not necessarily be related to discriminability—which would vary across the continuum—and thus the aftereffect would not necessarily correlate with face recognition.

Third, correlations between face aftereffects and face recognition might be masked if individual differences in preadaptation baseline are ignored. Theoretically, the link between aftereffect size and the slope of an individual's neural response functions requires that the deviation of the adaptor be measured from that individual's preadaptation norm, which was not zero pixels for all participants in our study. Ignoring individual differences in the distance of the adaptor from the norm would therefore potentially reduce statistical power by adding noise to any correlation.[2] Note that the traditional approach of calculating aftereffects as raw shift scores (postadaptation PSE minus preadaptation PSE) remains valid for group studies that average across participants (i.e., average adaptor distance = 50 pixels in our study).

Finally, the correlation between face aftereffects and recognition ability could plausibly have a non-face-space contribution. In our study, face aftereffects were dissociated from general visual memory and from nonface aftereffects, which ruled out the possibility that general attentional factors played a role in the correlation between face aftereffects and face recognition. However, there could perhaps be individual differences in *face-specific* attentional strategies. For example, if some individuals pay greater attention to the mouth than to the eyes, relative to other individuals, this might result in their exhibiting smaller eye-height aftereffects and—if the eyes are more diagnostic than the mouth—poorer face recognition.[3] (Note, however, that no current evidence suggests that region-specific attention influences magnitude of the aftereffect.)

What additional factors might contribute to individual differences in face recognition? Although our results indicate that face-space tuning for eye height is important for face recognition ability, the observed correlation ($r = .23$) was well below the upper bound ($r = .86$, calculated as the square root of the product of the internal reliabilities of the two tasks). Thus, considerable variance must be accounted for by other factors, such as the following.

Within face-space, quality of coding for face attributes other than eye height (e.g., mouth width, cheek shape) would also be expected to contribute to face recognition ability. Our finding that aftereffects for eye height alone correlate significantly with face recognition suggests that the steepness of neural functions for eye height might generalize to other face attributes; that is, an individual with more sensitive coding for one face-space attribute might also have more sensitive coding for others.

Beyond face-space, individual differences in holistic processing (Wang, Li, Fang, Tian, & Liu, 2012) and general visual memory (Dennett et al., 2012; Wilmer et al., 2010) may also contribute to face recognition ability. We found that the CCMT and face aftereffects explained nonoverlapping variance in the CFMT, which suggests that general visual memory contributes to face recognition independently of face-space coding. Indeed, holistic processing, general visual memory, and face-space coding might all contribute independently to face recognition ability.

Acknowledgments

Jess Irons tested some participants.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by Australian Research Council Grant DP0984558.

Notes

1. Children with autism spectrum disorder (ASD), who also sometimes show poor face memory, do show reduced face aftereffects (Pellicano, Jeffery, Burr, & Rhodes, 2007). However, it is difficult to rule out the possibility that these children apply reduced attention to the adapting faces as a result of the lack of social interest that characterizes ASD, and attention affects the size of face aftereffects (Moradi, Koch, & Shimojo, 2005; Rhodes et al., 2011).

2. In the present study, incorrectly assuming that adaptor distance was +50 pixels for all participants made little difference to the absolute *r* value, but resulted in the correlation becoming only marginally significant, $r = .22$, $p = .05$, 95% CI: [0, .42].

3. We thank Mike Webster for this idea.

References

Afraz, S.-R., & Cavanagh, P. (2008). Retinotopy of the face aftereffect. *Vision Research*, *48*, 42–54.

Bowles, D. C., McKone, E., Dawel, A., Duchaine, B., Palermo, R., Schmalzl, L., . . . Yovel, G. (2009). Diagnosing prosopagnosia: Effects of ageing, sex, and participant-stimulus ethnic match on the Cambridge Face Memory Test and Cambridge Face Perception Test. *Cognitive Neuropsychology*, *26*, 423–455. doi:10.1080/02643290903343149

Dakin, S. C., & Omigie, D. (2009). Psychophysical evidence for a non-linear representation of facial identity. *Vision Research*, *49*, 2285–2296.

Dennett, H. W., McKone, E., Tavashmi, R., Hall, A., Pidcock, M., Edwards, M., & Duchaine, B. (2012). The Cambridge Car Memory Test: A task matched in format to the Cambridge Face Memory Test, with norms, reliability, sex differences, dissociations from face memory, and expertise effects. *Behavior Research Methods*, *44*, 587–605. doi:10.3758/s13428-011-0160-2

Dickinson, J. E., Almeida, R. A., Bell, J., & Badcock, D. R. (2010). Global shape aftereffects have a local substrate: A tilt aftereffect field. *Journal of Vision*, *10*(13), Article 5. doi:10.1167/10.13.5

Duchaine, B. C., & Nakayama, K. (2006). The Cambridge Face Memory Test: Results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia*, *44*, 576–585.

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*, 1187–1196.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*, 89–94.

Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience*, *8*, 1386–1390.

Maddess, T., McCourt, M. E., Blakeslee, B., & Cunningham, R. B. (1988). Factors governing the adaptation of cells in area-17 of the cat visual cortex. *Biological Cybernetics*, *59*, 229–236.

McKone, E., Hall, A., Pidcock, M., Palermo, R., Wilkinson, R. B., Rivolta, D., . . . O'Connor, K. B. (2011). Face ethnicity and measurement reliability affect face recognition performance in developmental prosopagnosia: Evidence from the Cambridge Face Memory Test-Australian. *Cognitive Neuropsychology*, *28*, 109–146. doi:10.1080/02643294.2011.616880

Moradi, F., Koch, C., & Shimojo, S. (2005). Face adaptation depends on seeing the face. *Neuron*, *45*, 169–175.

Movshon, J. A., & Lennie, P. (1979). Pattern-selective adaptation in visual cortical neurons. *Nature*, *278*, 850–852.

Nishimura, M., Doyle, J., Humphreys, K., & Behrmann, M. (2010). Probing the face-space of individuals with prosopagnosia. *Neuropsychologia*, *48*, 1828–1841.

Palermo, R., Rivolta, D., Wilson, C. E., & Jeffery, L. (2011). Adaptive face space coding in congenital prosopagnosia: Typical figural aftereffects but abnormal identity aftereffects. *Neuropsychologia*, *49*, 3801–3812. doi:10.1016/j.neuropsychologia.2011.09.039

Pellicano, E., Jeffery, L., Burr, D., & Rhodes, G. (2007). Abnormal adaptive face-coding mechanisms in children with autism spectrum disorder. *Current Biology*, *17*, 1508–1512.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research*, *46*, 2977–2987.

Rhodes, G., Jeffery, L., Evangelista, E., Ewing, L., Peters, M., & Taylor, L. (2011). Enhanced attention amplifies face adaptation. *Vision Research*, *51*, 1811–1819. doi:10.1016/j.visres.2011.06.008

Rhodes, G., & Leopold, D. A. (2011). Adaptive norm-based coding of face identity. In A. J. Calder, G. Rhodes, M. H. Johnson, & J. V. Haxby (Eds.), *The Oxford handbook of face perception* (pp. 263–286). Oxford, England: Oxford University Press.

Robbins, R., McKone, E., & Edwards, M. (2007). Aftereffects for face attributes with different natural variability: Adapter position effects and neural models. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 570–592.

Susilo, T., McKone, E., Dennett, H. W., Darke, H., Palermo, R., Hall, A., . . . Rhodes, G. (2011). Face recognition impairments

despite normal holistic processing and face space coding: Evidence from a case of developmental prosopagnosia. *Cognitive Neuropsychology*, *27*, 636–664. doi:10.1080/02643294.2011.613372

Susilo, T., McKone, E., & Edwards, M. (2010a). Solving the upside-down puzzle: Why do upright and inverted face aftereffects look alike? *Journal of Vision*, *10*(13), Article 1. doi:10.1167/10.13.1

Susilo, T., McKone, E., & Edwards, M. (2010b). What shape are the neural response functions underlying opponent coding in face space? A psychophysical investigation. *Vision Research*, *50*, 300–314. doi:10.1016/j.visres.2009.11.016

Tanaka, J. W., & Corneille, O. (2007). Typicality effects in face and object perception: Further evidence for the attractor field model. *Attention, Perception, & Psychophysics*, *69*, 619–627. doi:10.3758/bf03193919

Thomas, C., Avidan, G., Humphreys, K., Jung, K. J., Gao, F., & Behrmann, M. (2009). Reduced structural connectivity in ventral visual cortex in congenital prosopagnosia. *Nature Neuroscience*, *12*, 29–31.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *43*, 161–204.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science*, *23*, 169–177. doi:10.1177/0956797611420575

Webster, M. A., & MacLin, O. H. (1999). Figural aftereffects in the perception of faces. *Psychonomic Bulletin & Review*, *6*, 647–653.

Wilmer, J. B. (2008). How to use individual differences to isolate functional organization, biology, and utility of visual functions; with illustrative proposals for stereopsis. *Spatial Vision*, *21*, 561–579.

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., . . . Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences, USA*, *107*, 5238–5241. doi:10.1073/pnas.0913053107

Wilson, H. R., Loffler, G., & Wilkinson, F. (2002). Synthetic faces, face cubes, and the geometry of face space. *Vision Research*, *42*, 2909–2923.

NeuroImage 221 (2020) 117191

Contents lists available at ScienceDirect

NeuroImage

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

Processing communicative facial and vocal cues in the superior temporal sulcus

Ben Deen[a,b,*], Rebecca Saxe[a], Nancy Kanwisher[a]

[a] Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, United States
[b] Laboratory of Neural Systems, The Rockefeller University, New York, NY, United States

A B S T R A C T

Facial and vocal cues provide critical social information about other humans, including their emotional and attentional states and the content of their speech. Recent work has shown that the face-responsive region of posterior superior temporal sulcus ("fSTS") also responds strongly to vocal sounds. Here, we investigate the functional role of this region and the broader STS by measuring responses to a range of face movements, vocal sounds, and hand movements using fMRI. We find that the fSTS responds broadly to different types of audio and visual face action, including both richly social communicative actions, as well as minimally social noncommunicative actions, ruling out hypotheses of specialization for processing speech signals, or communicative signals more generally. Strikingly, however, responses to hand movements were very low, whether communicative or not, indicating a specific role in the analysis of face actions (facial and vocal), not a general role in the perception of any human action. Furthermore, spatial patterns of response in this region were able to decode communicative from noncommunicative face actions, both within and across modality (facial/vocal cues), indicating sensitivity to an abstract social dimension. These functional properties of the fSTS contrast with a region of middle STS that has a selective, largely unimodal auditory response to speech sounds over both communicative and noncommunicative vocal nonspeech sounds, and nonvocal sounds. Region of interest analyses were corroborated by a data-driven independent component analysis, identifying face-voice and auditory speech responses as dominant sources of voxelwise variance across the STS. These results suggest that the STS contains separate processing streams for the audiovisual analysis of face actions and auditory speech processing.

1. Introduction

We learn a great deal about the character, thoughts, and emotions of another person by watching their face and listening to their voice. In addition to explicit verbal information, face movements and vocal sounds convey rich nonverbal clues to others' internal states that are essential for normal social interaction. What brain mechanisms underlie the extraction and representation of these communicative signals?

A candidate locus of these processes is the superior temporal sulcus (STS), which is considered a convergence zone for diverse sources of social information. Many prior studies using fMRI and electrocorticography have found responses to human vocal sounds within the middle STS and superior temporal gyrus (Belin et al., 2002, 2000; Binder et al., 2000; Liebenthal et al., 2005; Mesgarani et al., 2014; Norman-Haignere et al., 2015; Overath et al., 2015; Scott et al., 2000; Shultz et al., 2012; Vouloumanos et al., 2001; Wright et al., 2003). These responses have been interpreted either to reflect specialization either for speech processing (Norman-Haignere et al., 2015; Overath et al., 2015; Scott et al., 2000; Vouloumanos et al., 2001), or processing of vocal sounds more generally (Belin et al., 2002, 2000; Deen et al., 2015; Fecteau et al., 2004; Shultz et al., 2012). Within the posterior STS (pSTS), neuroimaging studies have reliably observed visual responses to perceived face movements (Allison et al., 2000; Bernstein et al., 2018; Pelphrey et al., 2005; Pitcher et al., 2011; Puce et al., 1998; Schultz et al., 2013), and spatial patterns of response that discriminate types of face movement (Deen and Saxe, 2019; Said et al., 2010; Srinivasan et al., 2016). These observations have led to the hypothesis that the STS contains a dorsal stream for face processing, specialized for extracting dynamic information from face motion, and distinct from a static form pathway on the ventral surface (Bernstein and Yovel, 2015; Freiwald et al., 2016).

While the face-motion-responsive subregion of pSTS (here termed fSTS) has typically been described as a category-specific visual region (Bernstein and Yovel, 2015; Freiwald et al., 2016; Haxby et al., 2000; O'Toole et al., 2002; Pitcher et al., 2011; Schultz et al., 2013), the broader posterior STS is considered a zone of multimodal association cortex, with responses to both visual and auditory stimuli (Beauchamp et al., 2004, 2008; Hein et al., 2007; Noesselt et al., 2007; Van Atteveldt et al., 2004), and recent studies have found common responses to face movements and vocal sounds within the pSTS in individual human brains (Deen et al., 2015; Watson et al., 2014a; Zhu and Beauchamp, 2017). Our recent work found that fSTS, functionally defined as the maximally face-motion-sensitive subregion of pSTS, has an equally strong response to auditory vocal sounds as to face movements (Deen et al., 2015). These results suggest that fSTS should be considered fundamentally multimodal, and raise questions about the functional role of this face- and voice-specific response.

* Corresponding author at: The Rockefeller University, 1300 York Ave. New York, NY 10065, United States.
E-mail address: benjamin.deen@gmail.com (B. Deen).

https://doi.org/10.1016/j.neuroimage.2020.117191
Received 21 March 2020; Received in revised form 14 July 2020; Accepted 19 July 2020
Available online 23 July 2020

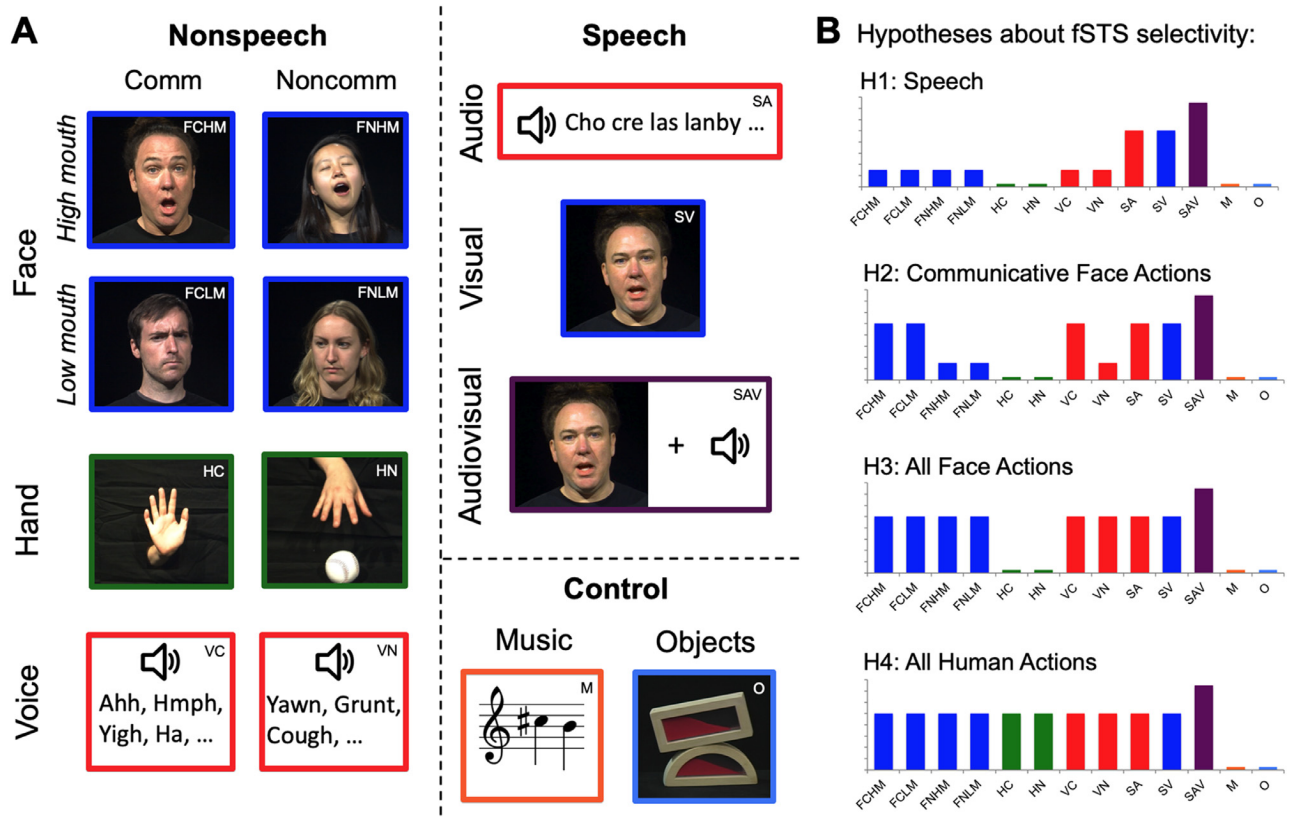


Fig. 1. (A) fMRI condition structure. Thirteen dynamic visual and auditory conditions were used, including face movements and vocal sounds categorized as speech, nonspeech communicative, and noncommunicative, as well as hand and object movement and music as controls. (B) Response profiles predicted by four hypotheses about the selectivity of face-motion-sensitive posterior superior temporal sulcus (fSTS).

Here, we consider four hypotheses regarding the functional role of the fSTS (Fig. 1). *1) The fSTS is specialized for processing audiovisual speech.* Speech is arguably the most ecologically relevant vocal sound we experience, and is well known to be processed audiovisually (McGurk and MacDonald, 1976; Reisberg et al., 1987; Sumby and Pollack, 1954). A face- and voice-responsive area would be well placed to support audiovisual speech processing, and prior studies have found that disrupting pSTS activity using transcranial magnetic or direct current stimulation impairs audiovisual speech perception (Beauchamp et al., 2010; Marques et al., 2014; Riedel et al., 2015). *2) The fSTS is specialized for processing communicative signals produced by faces.* Beyond speech, dynamic facial and vocal signals are used more broadly to communicate social cues via expressions and nonspeech vocalizations. The STS has been argued to play a role in social perception, the inference of abstract social information from perceptual cues (Allison et al., 2000; Brass et al., 2007; Pelphrey et al., 2004; Saxe et al., 2004), and in processing communicative actions in particular (Redcay, 2008; Redcay et al., 2016; Shultz et al., 2012). *3) The fSTS is involved in the perceptual processing of any dynamic audio or visual signal produced by a human face.* On this hypothesis, the fSTS is specialized for processing dynamic facial and vocal cues, but has a broad involvement in processing different actions within this category, including minimally socially relevant actions like a cough or neck stretch. *4) The fSTS is involved in the perceptual processing of any dynamic audio or visual signal produced by a human body.* On this hypothesis, the fSTS not only processes perceptual signals produced by others' faces, but by any body movement, including hand and full body movements. Prior research has found areas within pSTS responsive to both face movements and hand/body movements (Deen et al., 2015; Pelphrey et al., 2005; Thompson et al., 2007), but our recent work found that the strongest pSTS response to naturalistic face movement lies slightly anterior to body movement responses (Deen et al., 2015). The present study aimed to distinguish these hypotheses, and to test how the functional specialization of face-sensitive posterior STS compares to that of voice/speech-responsive middle STS.

To this end, we used fMRI to measure STS responses to a range of naturalistic face and hand movements, and vocal sounds (Fig. 1). These included speech signals, as well as richly communicative, socially relevant nonspeech signals (e.g., a surprised face, a vocal expression of disgust, a hand gesturing "stop"), and noncommunicative, less socially relevant stimuli (e.g., a chewing face, a throat-clearing sound, and a hand writing with a pen). While many prior fMRI studies have measured responses to a small number of conditions in a given set of participants, directly comparing responses to many stimulus conditions within individual participants can provide stronger constraints on theories of functional specialization (Deen et al., 2015; Fedorenko et al., 2013; Norman-Haignere et al., 2015; Poldrack, 2017). We compare responses across two STS regions-of-interest (ROIs), defined functionally in individual participants: fSTS, defined by a visual dynamic faces > dynamic objects contrast, and vSTS, defined by an auditory voices > music contrast. Additionally, we use a data-driven voxel decomposition method (independent component analysis) to identify dominant sources of variance in responses across the STS.

We find that the fSTS responds broadly to different types of face movements and vocal sounds, including speech, nonspeech communicative, and noncommunicative signals, but does not respond strongly to hand movements or non-social control stimuli (object movements or musical sounds). Although the mean response of the fSTS did not discriminate between communicative and noncommunicative signals, patterns of response in the region could be used to decode this distinction, both within and across input domains (faces and voices). This response

profile is consistent with a mid-level representation of face actions that is not restricted to socially relevant input, but begins to make abstract social dimensions explicit, and to generalize across input domains. The vSTS, in contrast, responded most strongly to auditory speech signals, over nonspeech vocal sounds, visual stimuli, and nonsocial controls. ROI-based responses were corroborated by a data-driven independent component analysis, demonstrating that voxelwise responses across the STS are well modeled as a linear combination of four component response profiles: responses to visual stimuli, auditory stimuli, faces and voices, and speech. These results suggest that the STS is organized into separate processing streams, one for audiovisual face actions and another for speech sounds.

2. Methods

2.1. Participants

Fifteen adults participated in the study (age 18–34 years, nine female). Participants had no history of neurological or psychiatric impairment, and normal or corrected vision. All participants provided written, informed consent.

2.2. Stimuli and paradigm

Participants viewed a set of video and audio clips depicting various face and hand movements and vocal sounds, as well as nonsocial controls, broadly sampling the space of human social perceptual inputs (Fig. 1). Among nonspeech stimuli, we included both richly social communicative actions and minimally social noncommunicative actions in each modality, and orthogonally manipulated the presence of mouth motion in face movements. For our purposes, a "communicative" action is defined as one produced to intentionally communicate information to another agent. Communicative hand movements consisted of gestures, while noncommunicative hand movements consisted of hand-object interactions. We additionally included audio, visual, and audiovisual speech stimuli, consisting of speakers uttering lists of nonsense words with English phonology. Lastly, we included audio clips of instrumental music as an auditory control, and video clips of moving objects as a visual control. This led to thirteen total conditions (Fig. 1A): 1) communicative, high-mouth-motion face movements (FCHM); 2) communicative, low-mouth-motion face movements (FCLM); 3) noncommunicative, high-mouth-motion face movements (FNHM); 4) noncommunicative, low-mouth-motion face movements (FNLM); 5) communicative hand movements (HC); 6) noncommunicative hand movements (HN); 7) communicative nonspeech vocal sounds (VC); 8) noncommunicative nonspeech vocal sounds (VN); 9) audio nonword speech (SA); 10) visual nonword speech (SV); 11) audiovisual nonword speech (SAV); 12) music (M); 13) objects (O).

Human stimuli were recorded in a television studio using a professional-grade HD video camera and microphone. Face movements, vocal sounds, and speech acts were performed by four actors (two female), wearing black shirts, with a black matte backdrop. Hand movements were performed by three actresses (all female), with their right hand protruding from a black sheet, such that only their hand and upper arm were visible. All actors were unfamiliar to participants in the study.

Among nonspeech stimuli, there were 8–11 specific actions (or tokens) for each condition; each actor performed each action 3–13 times. These tokens were as follows: 1) FCHM: disgusted expression, exhausted exhale, intrigued expression, uncertain expression, uncertain head shake and expression, tongue stick, surprised expression (with mouth open), disapproving head shake and expression ("tsk-tsk"), "yeesh" expression; 2) FCLM: concerned brow raise, confused brow furrow, eye roll, disappointed head hang, head nod ("yes"), head shake ("no"), single head nod ("hi"), skeptical expression, suggestive expression, surprised expression (with mouth closed), wink; 3) FNHM: blow air, puff cheeks, chew food, cough, move lower jaw left/right, lick lips, pick at teeth with tongue, yawn; 4) FNLM: blink, falling asleep motion (head falling), gaze shift to the lower left, gaze shift to the lower right, gaze shift to the upper left, gaze shift to the upper right, neck stretch (side to side), neck stretch (rotating 180°), shiver, smooth pursuit eye movement, sniff; 5) HC: air quotes, "come here" wave, finger wag, money sign, finger gun gesture, finger point, "so-so" gesture, thumbs down, thumbs up, wave hello, dismissive wave; 6) HN: flip coin, grasp ball (with all fingers), grasp ball (with pointer finger and thumb), shake a bottle, sprinkle seasoning, toss a ball, tug a cord, turn a book page, twist a bottle cap, type on a keyboard, write with a pen; 7) VC: relaxed ahh, sad aww, cute aww, amused ha, hmph, flirtatious rrr, ugh, uh-huh, uh-uh, yigh; 8) VN: ahh (as if opening mouth for a doctor), wretching sound (as if being choked), cough, gargle, grunt, hiccup, throat clear with mouth closed, throat clear with mouth open, yawn. Among speech stimuli, there were 6 tokens (specific lists of nonwords; e.g. "cho cre las lanby caldet raldence cre paments cotlessy ploo"); each actor spoke each list 3–13 times.

From the resulting set of 1323 video and audio clips of nonspeech actions, we then chose a subset to use for the experiment, such that clip duration was controlled within modality (faces, hands, or voices), and such that balanced proportions of stimuli from each token and actor were included for each condition. Likewise, from the resulting set of 184 speech clips, we chose a subset such that duration of all clips was near 5 s, and such that balanced properties of stimuli from each token and actor were included. This resulted in 128 FCHM clips (mean duration 2.23 s), 128 FCLM clips (2.22 s), 128 FNHM clips (2.28 s), 128 FNLM clips (2.31 s), 144 HC clips (1.98 s), 144 HN clips (1.97 s), 157 VC clips (1.32 s), 168 VN clips (1.48 s), and 46 speech clips (5.07 s).

As a nonsocial auditory control condition, we used 150 instrumental music clips from a range of genres (e.g. classical, jazz, rock), cut in duration to 1.5 s to match the length of VN stimuli. Music clips were chosen from a larger set of 724 clips, as the subset of 150 clips that best matched vocal stimuli in frequency spectra (details on the computation of frequency spectra and other acoustic properties are included in Supplementary Information). All audio stimuli were root-mean-square amplitude normed and ramped with a 50 ms linear ramp at the beginning and end of the clip. As a nonsocial visual control condition, we used 60 video clips of dynamic objects, used in a prior experiment (Pitcher et al., 2011), cut to 2.27 s to match the duration of face motion clips.

In the fMRI experiment, stimuli were presented in a blocked design, with separate blocks for each of the thirteen conditions. A fixed number of clips were presented in each block; because stimulus durations differed across modalities, this number varied across modalities such that the total stimulus duration for blocks of each condition was roughly 20 s (9 stimuli for faces and objects, 10 for hands, 13 for nonspeech vocal sounds and music, and 4 for speech clips). The inter trial interval between clips in a block was chosen such that total block length was 22 s for each block. In each run, 26 blocks (2 per condition) were presented, in palindromic order, with specific block order counterbalanced across runs and participants. Blocks were separated by 6 s of a baseline condition, consisting of a black screen with a white central fixation cross. There was an additional 10 s of baseline at the beginning of the experiment, 16 s in the middle, and 10 s at the end, such that each run lasted 12:32 min. Each participant received eight runs of the experiment during a scan session. To maintain attention, participants performed a 1-back task during the experiment, pressing a button when an individual clip within a block repeated itself (one repeat per block). 1-back behavioral performance was high (mean accuracy 93.3%, hit rate 74.1%, false alarm rate 4.3%) and consistent across runs (Supplementary Information, Figure S3).

2.3. Stimulus ratings

To verify that our communicativeness manipulation was effective, we collected behavioral ratings on the stimuli using Amazon Mechanical Turk. For each video or audio clip from the communica-

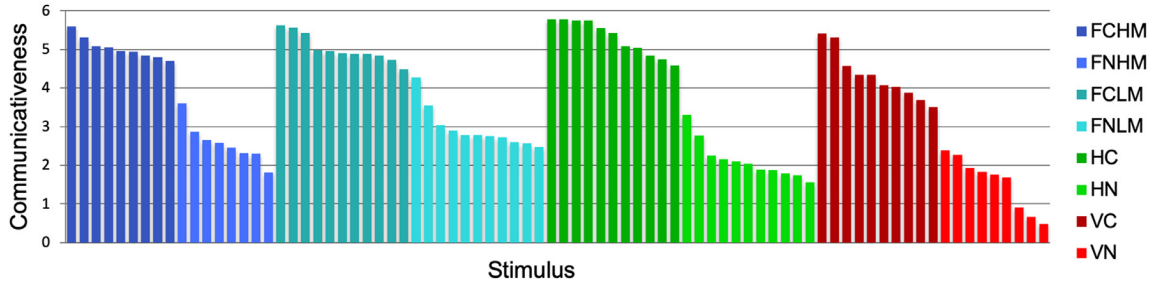


Fig. 2. Behavioral ratings of communicativeness, across the 80 specific actions used in the study, categorized by condition. Condition labels: FC = communicative face movement, FN = noncommunicative face movement, HM = high mouth motion, LM = low mouth motion, HC = communicative hand movement, HN = non-communicative hand movement, VC = communicative vocal sound, VN = noncommunicative vocal sound.

tive/noncommunicative conditions (FCHM, FCLM, FNHM, FNLM, HC, HN, VC, VN), 20 participants viewed or listened to the clip and answered questions in a brief survey. To assess communicativeness, we asked, "To what extent is this (sound/action) communicative (i.e., produced to intentionally communicate information to another human)?" Participants responded on a scale of 0 (not communicative at all) to 6 (highly communicative). Other questions were asked for separate purposes and are not reported here. Participants were limited to users in the USA, and with a task approval rating of at least 95%, and at least 50 tasks performed previously. The surveys included a catch question with an objective answer (e.g., "what is the gender of the actor/actress?" for face movement videos). Only responses with a correct answer to the catch question were accepted, to ensure that participants watched or listened to the clip and weren't responding randomly. Responses were averaged across participants, actors, and specific clips for each token (with an average of 281 responses per token), and statistics were performed across tokens.

Communicativeness ratings across all tokens are shown in Fig. 2. To assess the reliability of these responses, we split responses across two subsets of ten participants, and computed the split-half correlation across tokens. This correlation was very high ($r = 0.99$, $P \approx 0$), indicating highly reliable responses. We next used a one-way ANOVA to assess the effect of category (treating all eight categories as distinct) on responses, and observed a highly significant effect of category on communicativeness ratings ($F(7,72) = 84.14$, $P < 10^{-31}$, $R^2 = 0.89$). In particular, communicativeness was significantly higher for FCHM relative to FNHM ($t(15) = 12.42$, $P < 10^{-8}$), FCLM relative to FNLM ($t(20) = 10.84$, $P < 10^{-9}$), HC relative to HN ($t(20) = 15.47$, $P < 10^{-11}$), and VC relative to VN ($t(17) = 9.09$, $P < 10^{-7}$). Within each modality (faces, voices, hands), all tokens in the communicative condition were rated as more communicative than tokens in the noncommunicative condition. All communicative tokens were rated higher than middle score of 3, and all but 5 of the 39 noncommunicative tokens were rated lower than 3. These results demonstrate that our manipulation of communicativeness had the desired effect.

2.4. Data acquisition

MRI data were acquired using a Siemens 3T MAGNETOM Tim Trio scanner (Siemens AG, Healthcare, Erlangen, Germany). High-resolution T1-weighted anatomical images were collected using a multi-echo MPRAGE pulse sequence (repetition time [TR] = 2.53 s; echo time [TE] = 1.64 ms, 3.5 ms, 5.36 ms, 7.22 ms, flip angle $\alpha = 7^\circ$, field of view [FOV] = 256 mm, matrix = 256 × 256, slice thickness = 1 mm, 176 near-axial slices, acceleration factor = 3, 32 reference lines). Functional data were collected using a T2*-weighted echo planar imaging (EPI) pulse sequence sensitive to blood-oxygen-level-dependent (BOLD) contrast (TR = 2 s, TE=30 ms, $\alpha = 90^\circ$, FOV = 192 mm, matrix = 64 × 64, slice thickness = 3 mm, slice gap = 0.6 mm, 32 near-axial slices, near-whole-brain coverage).

2.5. Data preprocessing and modeling

Data were processed using the FMRIB Software Library (FSL), version 4.1.8, supplemented by custom MATLAB scripts. Anatomical and functional images were skull-stripped using FSL's brain extraction tool. Functional data were motion corrected using rigid-body transformations to the middle image of each run, corrected for interleaved slice acquisition using sinc interpolation, spatially smoothed using an isotropic Gaussian kernel (5 mm FWHM), and high-pass filtered (Gaussian-weighted least squares fit straight line subtraction, with $\sigma = 50$ s (Marchini and Ripley, 2000)). Although all analyses were performed in native functional space for each participant, normalization was required for combining results of certain analyses across participants. Functional images were registered to anatomical images using a rigid-body transformation determined by Freesurfer's bbregister (Greve and Fischl, 2009). Anatomical images were in turn normalized to the Montreal Neurological Instititute-152 template brain (MNI space), using FMRIB's nonlinear registration tool (FNIRT).

Whole-brain general linear model (GLM)-based analyses were performed for each participant and run. Regressors were defined as boxcar functions including each block from a given condition, convolved with a canonical double-gamma hemodynamic response function. Temporal derivatives of each regressor were included in the models, and all regressors were temporally high-pass filtered. FMRIB's improved linear model (FILM) was used to correct for residual autocorrelation (Woolrich et al., 2001). Lastly, data were combined across runs for each participant using 2nd-level fixed effects analyses, after registering beta maps from each run to a template image in native functional space (the middle image from the first run). Data were also combined across even runs and odd runs, for split-half analyses.

2.6. Region-of-interest analysis

How do face- and voice-sensitive subregions of the STS respond to communicative and noncommunicative face motions, hand motions, and vocal sounds? To address this question, we performed a region-of-interest (ROI) analysis, defining regions with face and voice contrasts. The face contrast compared the four face movement conditions to the dynamic object condition. The voice contrast compared the three vocal conditions (communicative/noncommunicative vocal sounds and audio speech) to the music condition. ROIs were defined in individual participants using the face and voice contrasts from the odd runs of the task. To spatially constrain ROI locations, we used search spaces defined based on a prior study, which identified a posterior STS face-sensitive region and a middle STS voice-sensitive region (Deen et al., 2015). Search spaces were defined as the set of active voxels (at the group level) within a 15mm-radius sphere around a peak coordinate, and registered from MNI space to each current participant's native functional space. For each participant, hemisphere, and contrast, we defined an ROI as the set of active voxels ($P < 10^{-3}$ voxelwise) within a 7.5mm-radius sphere around

Table 1
Mean coordinates of ROI centers-of-gravity, in MNI space.

ROI	*x*	*y*	*z*
lfSTS	−54.6	−36.9	3.9
rfSTS	54.3	−36.1	5.8
lvSTS	−60.0	−15.8	−0.9
RvSTS	57.5	−15.7	−5.3

the peak coordinate within the search space. Participants with no active voxels were excluded from the corresponding analysis; we identified right fSTS in 15/15 participants, left fSTS in 10, right vSTS in 13, and left vSTS in 11 participants. Mean ROI center-of-gravity coordinates are given in Table 1.

While we used a relatively strict statistical threshold to identify focal regions with particularly strong responses, and for consistency with our prior work (Deen et al., 2015), this method has the disadvantage of excluding participants without ROIs defined. An additional ROI analysis is described in the supplement, which assesses face-responsive regions within posterior/middle/anterior STS search spaces, defining ROIs using a top-N-voxel criterion. This analysis includes all participants by design, and enables us to ask whether significant responses to both faces and voices exist elsewhere along the length of the STS. The results corroborate the presence of face and voice responses in face-motion-sensitive posterior STS observed in the main ROI analysis.

For each ROI in the main analysis (left and right fSTS and vSTS), we extracted responses (percent signal change) across all thirteen conditions, in independent data from even runs of the experiment. Percent signal change was extracted by averaging beta values across each ROI and dividing by mean BOLD signal in the ROI. We then performed several statistical tests to characterize the response profiles of these regions. All tests were performed as mixed effects ANOVAs across conditions and participants, with participant included as a random effect, using MATLAB's fitlme function.

We first assessed selectivity profiles by comparing faces to objects, hands to objects, and vocal sounds (including speech) to music, using a separate ANOVA for each contrast and region. This analysis served to confirm that each region had a reliable effect of the contrast used to define it, and to replicate the pattern of selectivity we have observed previously (Deen et al., 2015). Second, we tested whether communicativeness modulated ROI responses, using a region by modality (face, voice, hand) by communicativeness ANOVA on all human non-speech conditions. Third, we tested whether speech content modulated responses, using a region by modality (face, voice) by speech content (speech, non-speech) ANOVA across all face and voice conditions. These ANOVAs were followed up with post-hoc tests to characterize the effects observed. Lastly, to test whether responses to face motion were modulated by the presence of mouth motion, we compared responses to high mouth motion versus low mouth motion videos.

2.7. Multivariate pattern analysis

The ROI analysis revealed that the fSTS responded similarly to communicative and noncommunicative face movements and vocal sounds. We next asked: would spatial patterns of response in these regions discriminate communicative from noncommunicative stimuli? Multivoxel pattern analysis (MVPA) provides a more sensitive measure of whether a brain region discriminates between two stimulus conditions, indicating that this distinction is represented in the region.

Specifically, we used the Haxby correlation method (Haxby et al., 2001). For each participant, we first split the data into two halves, and computed patterns of response for communicative and noncommunicative stimuli (for a given modality) in each half. We constructed a 2×2 matrix of Fisher-transformed correlations between patterns from the first and second halves, and used this to compute a difference score or "discrimination index": the mean within-condition correlation minus the mean between-condition correlation (i.e., the diagonal elements minus the off-diagonal elements of this matrix). Lastly, a one-tailed *t*-test was performed across participants, to test whether the discrimination index was significantly greater than zero, indicating that patterns in this region reliably discriminated between communicative and noncommunicative conditions.

In each ROI, defined as described above, we performed seven specific comparisons, testing discrimination of communicativeness within and across modalities: 1) within face movements; 2) within vocal sounds; 3) within hand movements; 4) within face movements, generalizing from low to high mouth movements; 5) face movements to vocal sounds; 6) face movements to hand movements; and 7) vocal sounds to hand movements. For the first three analyses, data were split across even and odd runs; for the fourth, across high and low mouth motion conditions; and for the last three, across the relevant modalities.

We next asked whether other regions could discriminate communicative and noncommunicative stimuli. We first tested the vSTS, using the same tests described above. Additionally, we ran a whole-brain searchlight analysis, focusing on the crossmodal face-to-voice analysis. Using a crossmodal comparison guarantees that decoding is not driven by low-level stimulus confounds. At each voxel in a gray matter mask, we placed an 8mm-radius sphere around the voxel, intersected this with the gray matter mask, and computed a discrimination index for this region. The mask was defined using the MNI gray matter atlas, thresholded at 0%, registered to each participant's native functional space, and intersected with their brain mask. Maps of discrimination indices for each participant were registered to MNI space, and inference was performed across participants, by performing a one-tailed *t*-test on values at each voxel. The resulting statistical maps were thresholded at $P < .01$ voxelwise, to form contiguous clusters of activation (where two voxels are considered contiguous if they share a vertex). To correct for multiple comparisons across voxels, we used a permutation test to generate a null distribution for cluster sizes, and used this to threshold clusters of activation at $P < .05$.

2.8. Independent component analysis

While ROI-based analyses provide a detailed characterization of responses in STS subregions of interest, the STS is a large and functionally diverse area, and response profiles of interest may be missed by restricting focus to specific functional ROIs. We next asked: what are the dominant response profiles to dynamic faces and voices across the entire STS? To this end, we analyzed our data using independent component analysis (ICA), which models voxelwise responses as a linear combination of underlying response profiles, such that the weightings of each profile across voxels are maximally statistically independent. This approach complements the ROI analysis in two ways: 1) it is data-driven, allowing the dominant features of STS functional organization to be revealed by our data; 2) it assesses responses across the full STS, rather than in a set of predefined ROI locations.

Methods used for ICA are depicted in Fig. 5. The input data for our implementation of ICA consisted of a condition-by-voxel matrix. We first defined an STS mask by manually drawing gray matter in the STS bilaterally in MNI space, and registered this to each participant's native functional space. Within this bilateral STS mask, we selected voxels that responded to a task > rest contrast at a liberal threshold ($P < .01$ voxelwise) within each individual participant. Beta values from each of the thirteen conditions were extracted from each selected voxel, to construct a condition-by-voxel data matrix for a given participant. For each participant, we then removed the mean of this matrix across voxels, and divided by the standard deviation across voxels and conditions, to ensure that each participant contributed similarly to the overall matrix. These within-participant data matrices were concatenated across participants in the voxel dimension to define a group-level data matrix. This approach to combining data across participants doesn't rely on normal-

ization, and thus doesn't require an assumption that voxels in similar locations across participants are functionally similar, and allows for voxel selection in each participant (Norman-Haignere et al., 2015).

Prior to performing ICA, we performed dimensionality reduction using principal components analysis (PCA), to restrict our attention to dimensions capturing reliable variance. To this end, we used a leave-one-participant-out approach. For each participant, we ran PCA on a data matrix from the other 14 participants, to obtain a set of 13 principal component vectors in 13-dimensional condition space. We then split the left-out participant's data in half by even and odd runs, and computed a condition-by-voxel data matrix separately for each half. For each potential number of components D (between 1 and 13), we projected the first-half data matrix onto the subspace spanned by the first D components, and computed the extent to which the resulting projected data could explain the second-half data matrix, by computing explained variance across voxels and conditions. Principal component dimensions capturing reliable variance should increase variance explained in second-half data, while dimensions capturing unreliable variation should decrease it as a result of overfitting the first-half data. Averaging across left-out participants, we found that split-half variance explained was maximized with four components (Fig. 5). Identified principal components were highly consistent across left-out participants: the mean normalized dot product between the first four PC vectors across PCA solutions from different left-out participants was 0.99.

Having identified the number of principal component dimensions capturing reliable variance in our data, we next ran PCA on our full data matrix, reduced our data to values along the first four principal component dimensions, and prewhitened the data by dividing by the standard deviation along each dimension. After prewhitening, performing ICA corresponds to finding an orthogonal basis or rotation that minimizes statistical dependence between values along each axis (Fig. 5). By the Central Limit Theorem, linear combinations of independent random variables will tend toward Gaussian distributions. Thus, identifying underlying independent components from observed linear combinations is equivalent to finding axes with minimally Gaussian data distributions (Hyvärinen and Oja, 2000). We obtained this basis using an algorithm that minimizes entropy along a set of orthogonal axes (Norman-Haignere et al., 2015, nonparametric algorithm, https://github.com/snormanhaignere/nonparametric-ica). For prewhitened data, minimizing entropy is equivalent to minimizing mutual information, a measure of statistical dependence. Minimizing entropy is also equivalent to maximizing non-Gaussianity, because the Gaussian distribution has maximum entropy for a given variance. This procedure yielded a set of four 13-dimensional independent component (IC) vectors, corresponding to response profiles capturing maximally independent sources of variance. In addition to reporting these profiles, we assessed spatial maps of voxel weights. Each voxel's response profile was modeled as a linear combination of IC vectors, where the coefficient for each component constituted a weight. These values were normalized to MNI space and averaged across participants to compute spatial maps of voxel weights for each component. To test whether IC weights were lateralized, we computed a laterality index—the difference between the mean voxel weight in left and right hemispheres. This index was tested against the null hypothesis of zero using a one-sample, two-tailed *t*-test across participants.

Our ICA method can only find meaningful independent components if data distributions along these dimensions are non-Gaussian. We tested this assumption by measuring statistical properties of voxel weight distributions—skewness and kurtosis—in each participant. These statistics were tested against the null hypothesis of values from a Gaussian distribution (skewness=0, kurtosis=3) using a nonparametric bootstrap test, resampling from the distribution of statistics across participants (10,000 samples).

Are spatial patterns of IC voxel weights consistent across participants? We next assessed spatial correlations of weight maps from pairs of participants. Correlations were computed between maps in MNI space, restricted to voxels that were used as input for both participants. To assess significance, we compared within-component and between-component correlations using a permutation test. We formed a null distribution for the difference between within- and between-component correlations, by permuting pairs of components (1–1, 1–2, 3–4, etc.), which are exchangeable under the null hypothesis of no difference between within- and between-condition correlations (10 choose 4 = 210 permutations).

Lastly, to evaluate the geometry of IC response profiles in 13-dimensional condition space, we computed normalized dot products between each component's response profile (corresponding to the cosine of the angle between response vectors). For illustration, these were compared to normalized dot products of principal component vectors, which are constrained to be orthogonal.

3. Results

3.1. Region-of-interest analysis

What role do face- and voice-responsive subregions of the STS play in interpreting social communicative signals? Here we ask this question by measuring fMRI responses in these regions to a range of dynamic visual and auditory social stimuli, including communicative and non-communicative face and hand movements and vocal sounds, as well as nonword speech stimuli. All tests were performed as mixed-effects ANOVAs across conditions and participants, with participant included as a random effect.

Responses in each ROI across all conditions are shown in Fig. 3. We first tested the selectivity profile of face-sensitive posterior STS (fSTS) and voice-sensitive middle STS (vSTS) by comparing responses to faces versus objects, hands versus objects, and voices versus music in independent data. The fSTS had a strong response to face versus object movements (left: $t(48) = 6.58$, $P < 10^{-7}$; right: $t(73) = 12.07$, $P < 10^{-18}$) and vocal sounds versus music (left: $t(38) = 4.09$, $P < 10^{-3}$; right: $t(58) = 3.86$, $P < 10^{-3}$), and a small but significant response to hand versus object movements (left: $t(28) = 2.92$, $P < .01$; right: $t(43) = 4.57$, $P < 10^{-4}$). The vSTS bilaterally responded to vocal sounds over music (left: $t(42) = 2.87$, $P < .01$; right: $t(50) = 4.36$, $P < 10^{-4}$). Additionally, there was an effect of faces versus objects in the right vSTS ($t(63) = 4.28$, $P < 10^{-4}$), although this reflected a response below baseline to the object condition, not a response above baseline to faces. These results indicate that the fSTS responds strongly to both faces and vocal sounds, while the vSTS responds specifically to vocal sounds, consistent with our prior findings (Deen et al., 2015).

Are STS responses to social stimuli modulated by communicative content, and does this modulation vary by modality (faces, voices, hands) and region? We tested this using a region by modality by communicativeness ANOVA. Although the regions differed in their overall response (main effect of ROI, $F(3368) = 37.43$, $P < 10^{-20}$) and in their selectivity across modality (ROI by modality interaction, $F(6368) = 4.06$, $P < 10^{-3}$), the communicativeness of the stimuli did not influence the response (main effect and interaction terms involving this factor, all P's > 0.7). This result indicates that communicative content had little influence on mean responses in bilateral fSTS and vSTS.

Because this ANOVA combines data across regions and modalities, it could potentially miss a subtle effect specific to a given region and modality. To address this possibility, we next performed post-hoc tests comparing responses to communicative versus noncommunicative stimuli, within each region and modality. Of these twelve tests, ten yielded null results. We did observe, however, an effect of communicativeness on left vSTS responses for vocal sounds ($t(20) = 3.50$, $P = .002$) and marginally for face movements ($t(42) = 2.56$, $P = .014$); the former effect would survive Bonferroni multiple comparisons correction across the twelve tests. These results largely corroborate the above ANOVA, indicating that communicative content has little influence on fSTS and vSTS responses, with the exception of an increased response to communicative vocal sounds in the left vSTS.

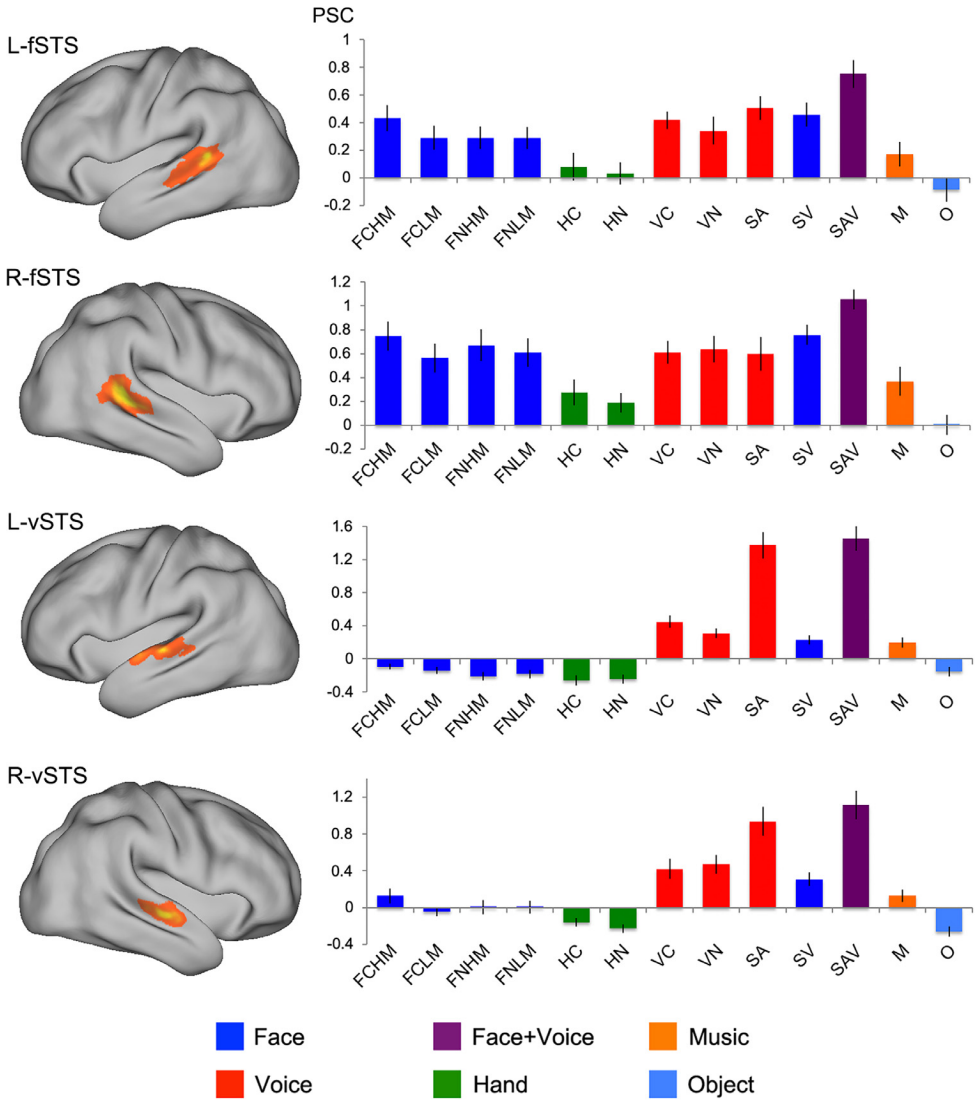


Fig. 3. Face-responsive posterior STS (fSTS) respond strongly to all face movements and vocal sounds, while voice-responsive middle STS (vSTS) responds selectively to speech sounds. Regions were defined using a faces > objects contrast (fSTS) and a voices > music contrast (vSTS). Left: heat maps of region-of-interest locations across participants. Right: responses of these regions (in percent signal change, PSC) across the thirteen experimental conditions, extracted from data independent from those used to define the regions. Condition labels: FC = communicative face movement, FN = noncommunicative face movement, HM = high mouth motion, LM = low mouth motion, HC = communicative hand movement, HN = noncommunicative hand movement, VC = communicative vocal sound, VN = noncommunicative vocal sound, SA = audio speech, SV = visual speech, SAV = audiovisual speech, *M* = music, *O* = objects.

We next asked whether STS responses to face movements and vocal sounds are modulated by speech content. A region by modality by speech content ANOVA again revealed that regions differed in their overall response (main effect of region, $F(3376) = 6.41$, $P < 10^{-3}$), and their relative response to faces and voices (region by modality interaction, $F(3376) = 18.40$, $P < 10^{-10}$). We also observed a region- and modality-specific modulation by speech content (region by modality by speech content interaction, $F(3368) = 4.03$, $P < .01$). Post-hoc tests revealed that these effects were driven by the presence of modality and speech effects in the vSTS bilaterally, and the absence of these effects in the fSTS. In particular, the vSTS responded more strongly to audio speech over vocal nonspeech sounds (left: $t(31) = 11.47$, $P < 10^{-11}$; right: $t(37) = 5.05$, $P < 10^{-4}$) and to visual speech over nonspeech face movements (left: $t(53) = 8.94$, $P < 10^{-11}$; right: $t(63) = 5.49$, $P < 10^{-6}$). The vSTS additionally responded more strongly overall to vocal than to face movement stimuli (left: $t(86) = 9.07$, $P < 10^{-13}$; right: $t(102) = 7.88$, $P < 10^{-11}$). In contrast, fSTS responses were not modulated by speech content or modality, with the exception of a marginally stronger response to visual speech over nonspeech in the left fSTS ($t(48) = 2.17$, $P = .035$).

Lastly, we compared the response of each region to nonspeech face movements with and without a mouth motion component (HM versus LM), to ask whether common responses to face movements and vocal sounds are driven by the presence of mouth movement (Zhu and Beauchamp, 2017). While both right and left fSTS responded strongly to face movements with or without a mouth component, responses in the right hemisphere were modulated by the presence of mouth movement (HM > LM, right: $t(58) = 3.06$, $P < .01$; left: $t(38) = 1.49$, $P = .15$). vSTS did not respond strongly to nonspeech face movements, but a marginal effect of mouth movement was observed in the right hemisphere (right: $t(50) = 2.28$, $P < .05$; left: $t(42) = 0.29$, $P = .77$). Thus, face- and voice-sensitive fSTS responded both to movements with and without mouth motion, but had a slight preference for movements with a mouth component in the right hemisphere.

Do face and voice responses, as observed in fSTS, exist in middle and anterior parts of the STS? A supplementary ROI analysis assessed face-motion-responsive ROIs within posterior, middle, and anterior STS, and found that while face and voice responses were most prominent posteriorly, such responses can be found along the length of the STS bilaterally (Fig. S4). This demonstrates that face-motion-responsive regions throughout middle and anterior the STS also have responses to vocal sounds, and shows that the face/voice response observed in posterior STS is robust across multiple methods for defining ROIs.

To summarize, we found that face-sensitive posterior STS (fSTS) responds strongly to a range of different face movements and vocal sounds, but does not respond strongly to hand movements or nonsocial audio or visual controls. This region responded similarly to various types of face movement and vocal sound, across differences in modality, communica-

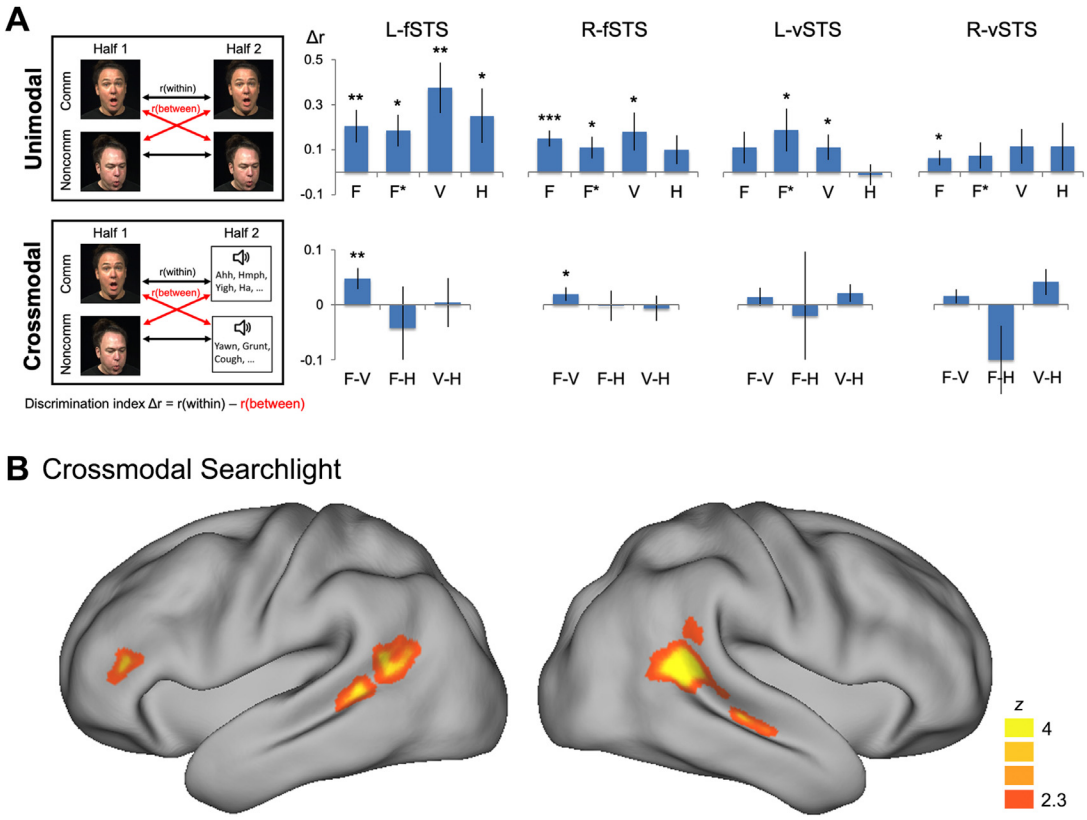


Fig. 4. Multivoxel pattern analysis results: decoding communicativeness from spatial patterns of response, both within and across modality. (A) Region-of-interest-based results, for fSTS and vSTS. Discrimination indices (correlation difference scores) for comparing patterns of response to communicative and noncommunicative stimuli. Within modality effects for faces (F); faces, generalizing from high to low mouth motion (F*); voices (V); and hands (H). Crossmodal effects for faces to voices (F-V), faces to hands (F-H), and voices to hands (V-H). * denotes $P < .05$, ** $P < .01$, *** $P < .001$. (B) Searchlight results for decoding communicativeness across modality (faces to voices). Whole-brain statistical map thresholded at $P < .01$ voxelwise, followed by a $P < .05$ permutation-based clusterwise threshold to correct for multiple comparisons.

tive content, and speech content. In contrast, the response profile of voice-sensitive middle STS (vSTS) indicates that this region is largely speech-selective, with a much stronger response to audio speech than to vocal nonspeech sounds and other conditions.

3.2. Multivoxel pattern analysis

While the ROI analysis showed similar mean responses in fSTS to communicative and noncommunicative face actions, it remains possible that patterns of activity in this region contain information about communicativeness. We next ask whether spatial patterns of response across voxels in fSTS differed between communicative and noncommunicative stimuli, both within and across modalities (faces, voices, hands).

MVPA results are shown in Fig. 4. Patterns in the fSTS were able to discriminate communicative from noncommunicative face movements (left: $t(9) = 2.83$, $P < .01$; right: $t(14) = 4.17$, $P < 10^{-3}$), even when requiring generalization across high and low mouth motion conditions (left: $t(9) = 2.64$, $P < .05$; right: $t(14) = 2.27$, $P < .05$). fSTS patterns were also able to discriminate between communicative and noncommunicative vocal sounds (left: $t(9) = 3.33$, $P < 10^{-3}$; right: $t(14) = 2.17$, $P < .05$), and the left but not right fSTS was able to discriminate between communicative and noncommunicative hand movements (left: $t(9) = 2.07$, $P < .05$; right: $t(14) = 1.56$, $P = .07$).

Are common patterns of fSTS response evoked by communicative and noncommunicative stimuli from different modalities? Indeed, these patterns could discriminate communicativeness when generalizing across face movements and vocal sounds (left: $t(9) = 2.95$, $P < .01$; right: $t(14) = 2.32$, $P < .05$), but not generalizing across hand movements and face movements or vocal sounds (P's > 0.45). This result indicates that fSTS responses differentiate communicative and noncommunicative stimuli in a manner that is to some extent consistent across audio and visual face actions, but does not generalize to hand movements. Furthermore, this crossmodal decoding result cannot be explained in terms of low-level visual or acoustic properties that differ across communicative and noncommunicative conditions within either modality.

Can patterns of response differentiating communicative from noncommunicative face actions be observed in other brain regions? We first tested these effects in the vSTS. Patterns in left vSTS were able to discriminate communicativeness for face movements, generalizing across high to low mouth movement conditions ($t(10) = 1.99$, $P < .05$) and for vocal sounds ($t(10) = 1.99$, $P < .05$), while patterns in right vSTS were able to discriminate communicativeness for face movements ($t(12) = 1.88$, $P < .05$). Other unimodal effects, and all crossmodal effects, were not significant (P's > 0.05). Thus, while spatial patterns of response in the vSTS show some sensitivity to communicative content, the effects were relatively weak and inconsistent across hemispheres, and neither region showed evidence for crossmodal decoding.

We next performed a whole-brain searchlight analysis. We focused on crossmodal decoding of communicativeness from facial to vocal stimuli, because this comparison is impervious to low-level confounds. The results from this searchlight are shown in Fig. 4B. Regions with significant decoding ability were found in the left posterior STS and right posterior and middle STS, overlapping with but extending posteriorly beyond face-responsive regions. We also observed a region of left infe-

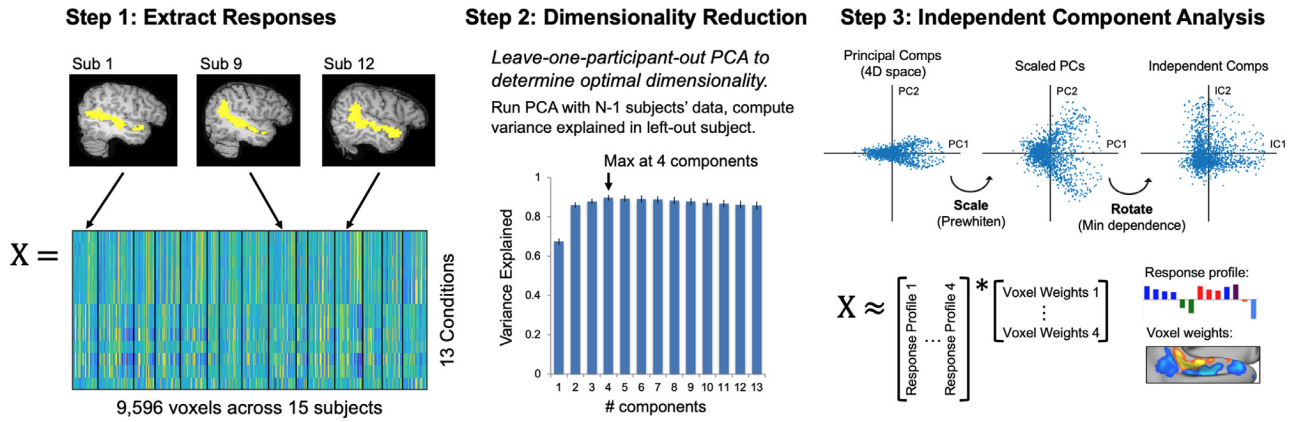


Fig. 5. Independent component analysis methods. Left, step 1: responses (beta values) were extracted across STS voxels and conditions, and concatenated across participants to form a data matrix. Middle, step 2: leave-one-participant-out principal component analysis (PCA) was used to determine the dimensionality for which the PC-spanning subspace explained maximum variance in left-out participants. Right, step 3: independent component analysis (ICA) was performed within the subspace spanned by the first four principal components, by first scaling data to have equal variance along each dimension (prewhitening), and then finding a rotation that minimizes statistical dependence between dimensions. Step 3 is visualized using synthetic data in two dimensions.

rior frontal gyrus. These results indicate that fMRI-decodable information about the communicativeness of face movements and vocal sounds is not strictly limited to the fSTS, but circumscribed to a set of focal regions within the STS and frontal cortex.

3.3. Independent component analysis

Are parts of the STS beyond functionally-defined fSTS and vSTS involved in processesing dynamic facial and vocal stimuli? While ROI-based analyses provide a detailed characterization of specific functional subregions, they don't assess responses in other parts of the STS, and require a priori assumptions about which regions are involved in processing our stimuli. We next complemented this approach with a data-driven independent component analysis, to ask more broadly, what are the dominant response profiles to dynamic social stimuli across the STS?

An initial PCA-based dimensionality reduction technique revealed that the split-half reliable sources of variance in response profiles across voxels could be captured by a 4-dimensional subspace of the 13-dimensional space of possible response vectors (Fig. 5). This subspace captured 95.3% of the total variance across voxels. Running ICA then yielded four response profiles spanning this subspace, with minimal statistical dependence of voxels' responses along each dimension. These response profiles, as well as spatial maps of voxel weights, are shown in Fig. 6A. Note that they are arbitrarily ordered and named based on a post-hoc assessment of their response profile.

The first two components had straightforward modality-specific response profiles. The first component had a positive response to all visual conditions, and roughly zero response to auditory conditions, and thus was termed the visual component. The voxel weights for this component followed a posterior-to-anterior spatial organization, with positive weights posteriorly (adjacent to early visual cortex) and decreasing weights moving anteriorly along the STS. The second component had a positive response to all auditory conditions, and roughly zero response to visual conditions, and thus was termed the auditory component. The voxel weights for this component were strongest near the upper bank of the middle STS (near early auditory cortex), and decreased moving ventrally, anteriorly, and posteriorly from this region.

The third component had a positive response to all face movement and vocal sound conditions, including communicative and noncommunicative conditions, and speech and nonspeech conditions, but had a negative response to hand movement, music, and object conditions. Much like the response profile of the fSTS ROI described above, this profile captures the discrimination between facial/vocal and other stimuli, and was thus termed the face+voice component. The voxel weights for this component were strongest around the posterior STS, with positive weights extending into middle and anterior STS.

The fourth component had a strong response to audio and audiovisual speech conditions, weak response to the visual speech, vocal nonspeech, and music conditions, and a negative response to the remaining face, hand, and object visual conditions. Similar to the response profile of the vSTS ROI described above, the dominant feature of this profile was audio speech selectivity, with a much stronger weight on audio/audiovisual speech than other conditions, as well as weaker effects of audio over visual stimuli and visual speech over nonspeech face motion. This component was thus termed the speech component. Similar to the auditory component, voxelwise weights were strongest in the upper bank of the middle STS, and decreased moving ventrally, anteriorly, and posteriorly.

Are the STS response profiles captured by these independent components dominant in a particular hemisphere? We computed a laterality index—the difference between mean voxel weights in the left and right hemispheres—and tested this index across participants (Fig. 6B). This index was only significant for component 3, the face+voice component ($P < .05$, two-tailed t-test), which had stronger weights in the right hemisphere.

Do our data satisfy the key underlying assumption of ICA—that distributions along IC dimensions are non-Gaussian? To assess non-Gaussianity, we measured the skewness and kurtosis of voxel weight distributions (Fig. 6B). We then tested the distributions of these statistics across participants against the null hypothesis of Gaussian values (skewness=0, kurtosis=3) using a nonparametric bootstrap test. Skewness was significantly greater than zero for components 1, 2, and 4 ($P \approx 0$, i.e. no bootstrap samples were less than 0), but not for component 3 ($P \approx 0.09$). Kurtosis was significantly greater than 3 for all components ($P \approx 0$). This demonstrates that components were non-Gaussian, demonstrating sparsity (high kurtosis) and a bias toward positive values (right-skew), which validates the non-Gaussianity assumption of our ICA method. Sparse, right-skewed weight distributions may result from anatomical clustering of neural populations with similar response profiles, yielding a small number of voxels with particularly high weights (Norman-Haignere et al., 2015). Notably, the face+voice component was sparse but not significantly skewed, reflecting the presence of large positive and negative weights across voxels and conditions.

Are spatial maps of voxel weights consistent across individual participants? Maps of voxel weights from a representative set of participants are shown in Fig. 7. These maps showed a consistent spatial structure across participants, despite the IC analysis having no information

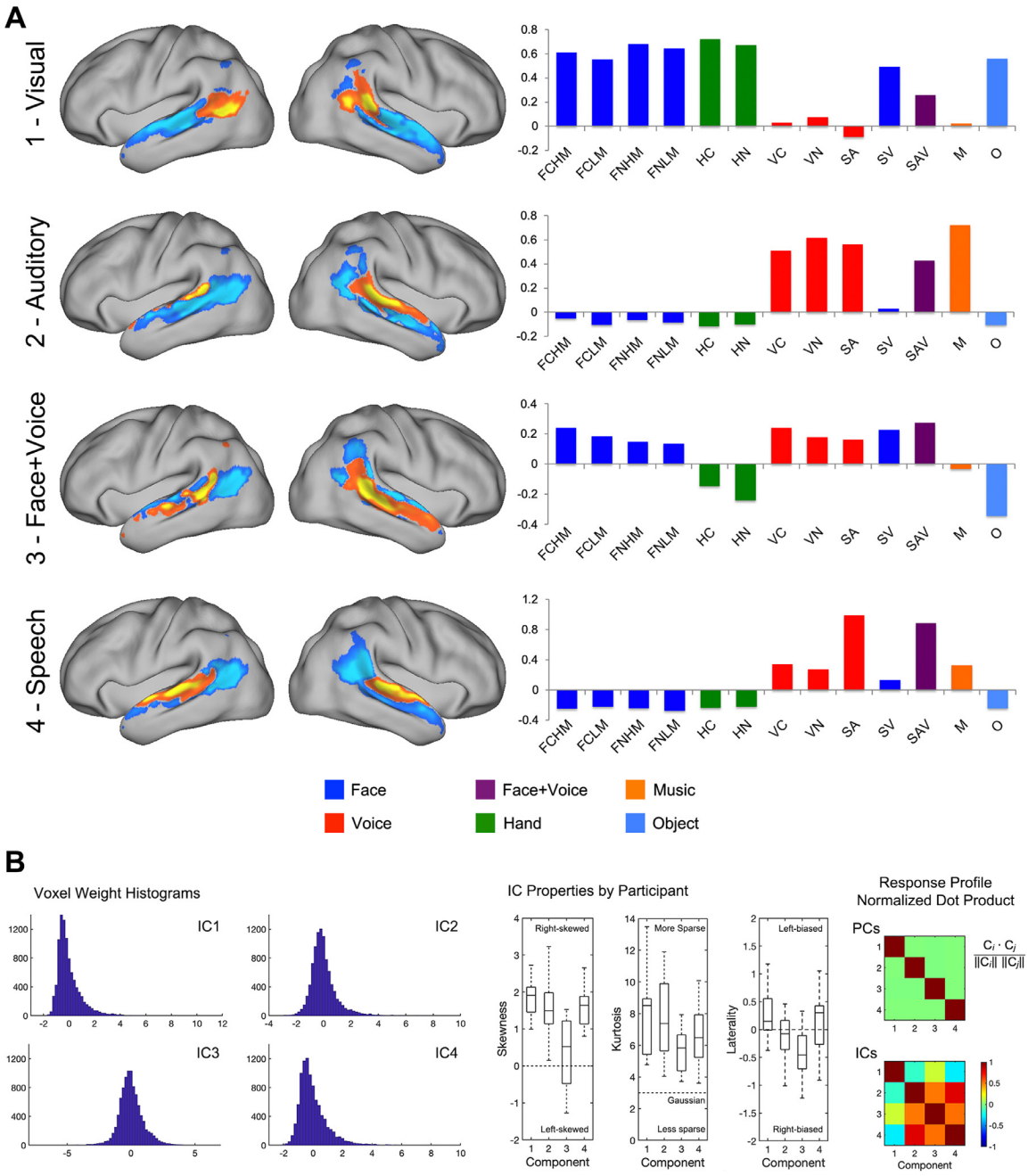


Fig. 6. Independent component analysis identifies face-voice and speech responses as dominant response profiles across the STS. (A) Right: response profiles for four independent components, which together explained ~95% of voxelwise variance in STS responses. Left: maps of voxel weights—the contribution of each component to a given voxel's response profile. Components are ordered arbitrarily and named based on post-hoc assessment of their response profiles. Condition labels: FC = communicative face movement, FN = noncommunicative face movement, HM = high mouth motion, LM = low mouth motion, HC = communicative hand movement, HN = noncommunicative hand movement, VC = communicative vocal sound, VN = noncommunicative vocal sound, SA = audio speech, SV = visual speech, SAV = audiovisual speech, *M* = music, *O* = objects. (B) Left: histograms of voxel weights for each component. Middle: properties of voxel weight distributions—skewness, kurtosis, and laterality index—shown as box and whisker plots of the distribution across participants. Boxes show the 25th percentile, median, and 75th percentile of the distribution, and whiskers show the range. Right: matrices showing normalized dot products between pairs of response profiles. Principle components (PCs) are constrained to be orthogonal, while independent components (ICs) are not.

about voxels' spatial location. To quantify this consistency, we compared within- and between-component correlations of spatial maps across participants. Within-component correlations were significantly larger than between-component correlations (mean correlation difference = 0.246, $P < .01$, permutation test).

How does the geometry of response profile vectors differ between principal components and independent components of our data? While PC response profiles are constrained to be orthogonal, IC response profiles do not have this constraint. To compare geometries of PC and IC response profiles, we computed normalized dot products between response profile vectors from each component, equal to cosine of the angle between response profiles in 13-dimensional condition space. These dot products were equal to zero for PCs, but were nonzero for ICs, with normalized dot products >0.5 for components 2, 3, and 4. Thus,

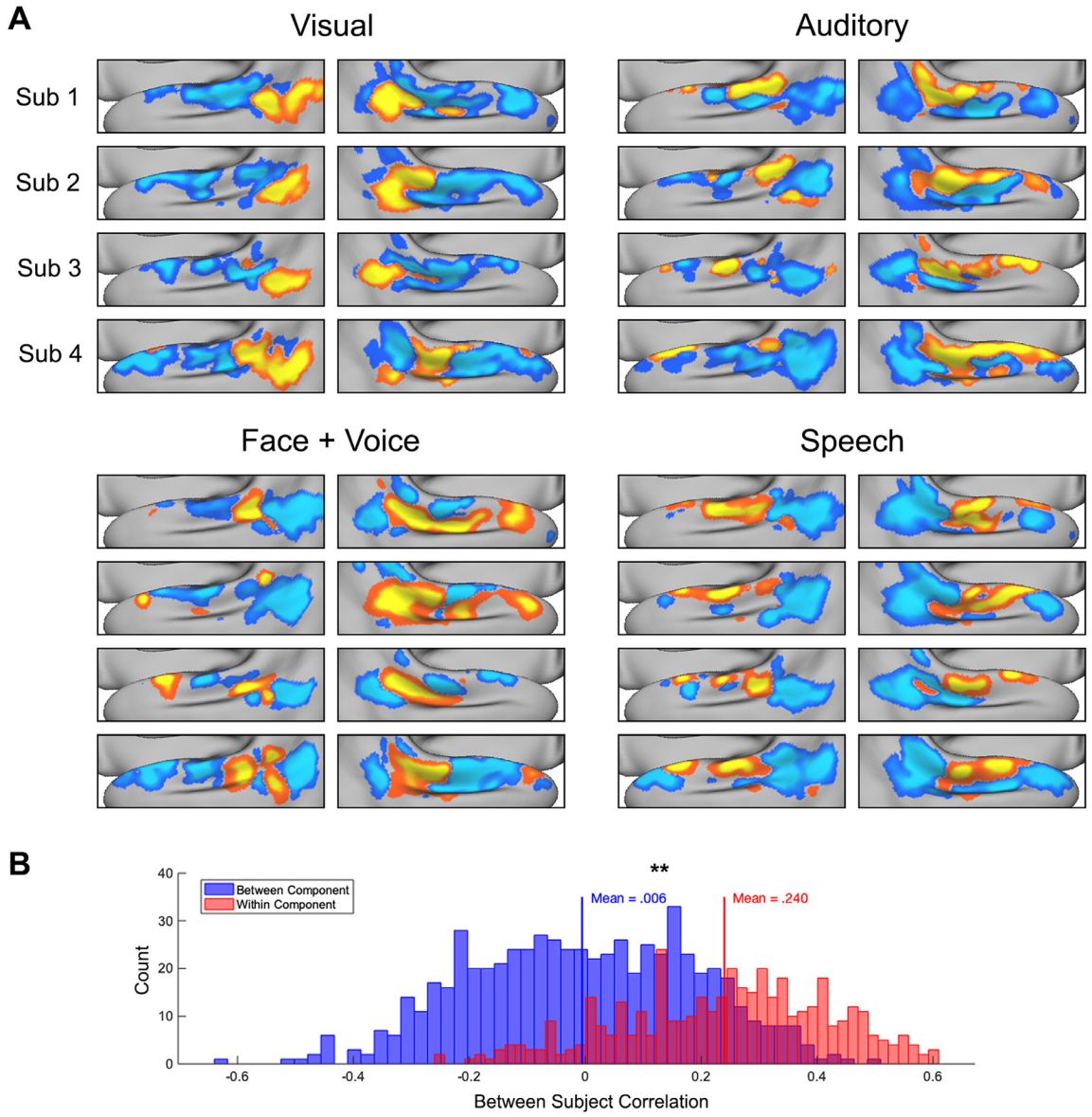


Fig. 7. Independent component voxel weight maps are consistent across participants. (A) Voxel weight maps for four representative participants. (B) Histograms of between-participant correlations in voxel weight maps, either within-component, or between-component. ** denotes $P < .01$.

PCA and ICA yielded components with a rather different geometric structure.

What proportion of unique variance in STS responses is explained by each component? Because IC vectors are nonorthogonal, they do not provide an orthogonal decomposition of voxelwise variance, as PC vectors would. However, we can assess the variance uniquely explained by each component by measuring the increase in explained variance from adding each IC vector to the subspace spanned by the other IC vectors. These measures were: 30% unique variance explained by the visual component, 7% by auditory, 8% by speech, and 5% by face-voice. Thus, each component uniquely explained an appreciable proportion of STS response variance.

In sum, a large portion of the voxelwise variance in response to the dynamic visual and auditory stimuli used in this experiment can be captured by a linear combination of four components: visual responses, auditory responses, responses to facial and vocal stimuli, and responses to auditory speech. Thus, face/voice- and speech-related response profiles identified in the ROI analysis are not merely idiosyncratic properties of the focal ROIs we chose, but are dominant profiles that capture variance in responses across the STS and emerge from a data-driven analysis.

4. Discussion

The present study measured STS responses to a range of visual and auditory social stimuli, in order to characterize the function of face- and voice-responsive STS subregions, fSTS and vSTS. We found that the fSTS responded strongly to both face movements and vocal sounds, but weakly to hand movements or nonsocial control stimuli. These findings are consistent with our prior results showing strong responses to faces and voices but weak responses to whole-body movements (Deen et al., 2015), and suggest a specific role of this region in processing audio and visual signals from the face. The fSTS had a similar mean response to a range of types of face movement and vocal sound, including communicative and noncommunicative stimuli and speech and nonspeech stimuli, in both modalities, pointing to a broad representation of dynamic face actions. These findings argue against hypotheses that fSTS is specialized for processing audiovisual speech, or communicative sig-

nals more generally. However, spatial patterns of response in this region could discriminate communicative and noncommunicative face actions, both within and across modality (faces/voices), demonstrating that this region encodes an abstract social feature crossmodally. The response profile of the fSTS contrasted with that of the adjacent vSTS, which had a selective response to auditory speech.

While prior work has documented overlapping posterior STS responses to faces and voices (Kreifelts et al., 2009; Watson et al., 2014a; Wright et al., 2003), the present result is striking in that the fSTS was defined as the maximally face-sensitive subregion of posterior STS in individual participants, and nevertheless it responded as strongly to vocal sounds as to faces. Face-responsive regions of middle and anterior STS were also found to have voice responses (Figure S3). Furthermore, a data-driven independent component analysis identified responses to faces and voices as a dominant source of voxelwise variance across the STS, with strongest voxel weights in posterior STS, and positive weights extending along the length of the STS in some participants. These results argue that "face regions" of the human STS (Haxby et al., 2000; Pitcher et al., 2011) are better characterized as "face-voice" regions, responsive to dynamic visual or auditory signals from human face, but minimally to nonfacial controls, including hand, body, and object movements, as well as nonvocal music and environmental sounds (see also Deen et al., 2015). This conclusion suggests a straightforward update to existing models of the human brain's face perception system (Bernstein and Yovel, 2015): the dorsal (STS) face processing stream is specialized not just for dynamic visual information from faces, but also dynamic auditory information from faces (see also Yovel and O'Toole, 2016).

What does the response profile of fSTS across multiple types of face and hand movement and vocal sound tell us about the functional role of this region? This region responded weakly to hand movements, even when communicative, suggesting against a role in processing any body movement. Among dynamic facial and vocal stimuli, however, the fSTS responded strongly to all stimulus categories presented—including speech and nonspeech, communicative and noncommunicative—and a similar pattern of response was observed for the face+voice component identified by ICA. This result argues against a strict specialization of this region for speech processing or social perceptual inference, instead pointing to a more general role in the multimodal perceptual processing of signals from faces. Such a region could plausibly contribute to a range of functions relying on audiovisual perceptual representations of face actions, including speech perception, social perception, and person identification. The broad response profile observed also suggests against the claim that voice responses within pSTS are specifically linked to mouth movement responses (Zhu and Beauchamp, 2017). While a small preference for stimuli with mouth movement was observed in the right hemisphere, the fSTS bilaterally responded strongly to both movements with and without a mouth component, and our prior work has found that a similarly defined region contains information about both eye and mouth movement type (Deen and Saxe, 2019). While our results don't contradict prior findings of subregions within posterior STS with preferences for eye or mouth movements (Pelphrey et al., 2005; Zhu and Beauchamp, 2017), they demonstrate that activations to any face movement or vocal sound constitute a dominant response profile across the STS.

While the fSTS responded strongly to both communicative and noncommunicative actions, spatial patterns of response in the fSTS were able to discriminate these two categories. This result held both within modality for faces and voices, as well as across these two modalities (e.g., training on faces and testing on voices, or vice versa), indicating that this distinction is encoded in an abstract, crossmodal manner. This finding demonstrates that this region encodes an abstract social dimension, and that representations in this region are to some extent audiovisual, with facial and vocal stimuli organized around a common dimension. In providing evidence for crossmodal coding of a socially relevant dimension, these results are broadly consistent with findings of crossmodal emotional state decoding in a region of pSTS/middle temporal gyrus (Peelen et al., 2010), and crossmodal adaptation for emotional state information in a region of pSTS (Watson et al., 2014b). However, we note that the regions assessed in these two studies likely differ slightly from the area studied here: e.g., the region found by Peelen et al. was not face-selective, and the region found by Watson et al. was not voice-selective.

What do these results tell us about the role of the fSTS in social perception, the process of inferring abstract social properties from perceptual input? As in the problem of transformation-invariant object recognition (DiCarlo et al., 2012), extracting social meaning from visual and auditory stimuli entails detecting cues that bear a highly nonlinear relationship to raw stimulus features, and thus might benefit from a hierarchical processing architecture. Brain regions positioned "lower" in the hierarchy would contain representations tied to lower-level stimulus features, potentially limited to certain domains of social information (face, hand or body motion, or vocal sounds). In contrast, brain regions situated "higher" in the hierarchy would contain explicit representations of communicated mental states and/or propositional content, abstracted across a range of stimulus features and input domains (Skerry and Saxe, 2014). On this view, social perceptual inference involves an interplay of regions across the hierarchy, with feedforward connections transmitting updated sensory input, and feedback connections conveying predictions driven by high-level representations (Koster-Hale and Saxe, 2013).

Where is the fSTS situated in this putative hierarchy? The properties reported here have some signatures of a low-level representation: the region responds similarly to highly and minimally socially relevant actions, and is specific to facial and vocal signals, not generalizing to socially relevant hand movements. However, other properties are more consistent with a high-level representation: fSTS responds to stimuli across multiple modalities (visual faces and auditory voices), and pattern analysis indicates that this region represents an abstract social property, in a manner that generalizes across modalities. Taken together, these results suggest that the fSTS plays a mid-level role in social perceptual inference, containing a representation of audiovisual face actions that is not restricted to socially relevant inputs, but which begins to make explicit abstract, social features across modalities.

What parts of the brain constitute "higher" regions in this hierarchy, with more explicit representations of abstract social information? Areas within higher-order association cortex implicated in high-level social cognition and theory of mind provide a plausible candidate network (Fletcher et al., 1995; Saxe and Kanwisher, 2003). These regions fall within the default mode network or apex network, situated at the top of the cortical sensory/motor processing hierarchy (DiNicola et al., 2020; Margulies et al., 2016), and have been found to contain abstract representations of features of others' internal states, including emotional states (Skerry and Saxe, 2014, 2015) and beliefs (Koster-Hale et al., 2014, 2017). This network contains a component in the anterior STS, and our prior work has found partial overlap between face movement and theory of mind responses within anterior STS (Deen et al., 2015). Thus, socially-sensitive subregions of the anterior STS could plausibly constitute a route through which information about dynamic face/voice signals is relayed from fSTS to areas involved in high-level social cognition. Future work should explore this possibility.

Beyond social perception, our results are consistent with prior studies implicating the posterior STS in the use of audiovisual information for speech perception and person identification. Studies using transcranial magnetic or direct current stimulation have found that disrupting the pSTS can disrupt audiovisual processing of speech content (Beauchamp et al., 2010; Marques et al., 2014; Riedel et al., 2015). fMRI studies have found sensitivity of pSTS responses to vocal identity (von Kriegstein et al., 2007, 2010), sensitivity to dynamic facial information relevant to identity has been hypothesized (Bernstein and Yovel, 2015; O'Toole et al., 2002), and recent studies have found ev-

idence for crossmodal identity representations (Anzellotti and Caramazza, 2017; Hasan et al., 2016; Tsantani et al., 2019). Furthermore, pSTS responses have been linked to benefits in auditory speech processing resulting from face-voice learning (Blank and von Kriegstein, 2013; von Kriegstein et al., 2008). However, we note that it is difficult to establish whether the studies mentioned above are in fact studying a common region of pSTS, or nearby but functionally distinct regions. Given differences in the precise anatomical location of functional regions across human participants, and differences in analysis and registration strategies across studies, finding responses in similar stereotaxic coordinates across studies does not demonstrate that these studies are assessing the same region (Brett et al., 2002); in fact, our prior work has demonstrated that nearby and even overlapping pSTS areas can have rather different response profiles (Deen et al., 2015). Using functional criteria to define regions in a consistent manner across studies provides one way to resolve this issue (Saxe et al., 2006).

In contrast to the broad response profile of the fSTS, the vSTS had a strikingly selective response profile, responding specifically to auditory speech stimuli over all other categories. While prior studies have argued that a similar region of the upper middle STS plays a role in processing speech sounds (Binder et al., 2000; Liebenthal et al., 2005; Scott et al., 2000; Vouloumanos et al., 2001; Wright et al., 2003), or vocal sounds more generally (Belin et al., 2002, 2000; Deen et al., 2015; Fecteau et al., 2004; Shultz et al., 2012), the present results suggest that this region is primarily specialized for speech processing. This result is consistent with a recent study assessing responses to a broad set of natural sounds, which found a response component localized to middle STS/STG with a much stronger response to speech than a variety of other sound categories, including nonspeech vocal sounds (Norman-Haignere et al., 2015; see also Pernet et al., 2015). Particularly striking here was the strong selectivity of vSTS for speech sounds over communicative nonspeech sounds, which were somewhat speech-like and typically involved one or multiple English phonemes. A potential explanation for this difference is that this region is sensitive to features of speech at longer timescales than individual phonemes, such as sequences of phonemes or prosodic contours (Overath et al., 2015). Although our results suggest that vSTS is specialized for processing speech over arbitrary vocal sounds, this doesn't rule out a potential role for this region for voice identification, given that speech sounds are the primary cue humans use to determine voice identity (Latinus et al., 2013).

The vSTS also responded more strongly to visually presented speech over other types of face movement, suggesting a potential role in the visual processing of speech signals as well. This finding is consistent with prior studies finding mid-STS responses to visual speech (Callan et al., 2004; Calvert et al., 1997; Capek et al., 2008), and extends these studies by including a number of meaningful face movement controls, including communicative nonspeech mouth movements.

Considering the response profiles of the fSTS and vSTS together, our results indicate that the STS contains distinct pathways for 1) processing of facial and vocal signals in general (corresponding to the dorsal face processing pathway), and 2) processing of speech signals. This conclusion contrasts with the common notion that the STS is subdivided into areas for processing faces (Haxby et al., 2000) and vocal sounds (Belin et al., 2000). This view of STS functional organization was further supported by data-driven ICA results, in which face/voice-responsive and speech-selective components emerged as dominant response profiles, contributing largely independent sources of variance in voxelwise responses across the STS. While we designate the regions studied here as fSTS and vSTS based on the functional criteria used to define them (face and voice responses), these results suggest that fvSTS and spSTS would be more appropriate names.

How do these findings relate to our understanding of systems for face and voice processing in nonhuman primates? The dorsal face processing stream in humans has been argued to relate to a dorsal stream within the upper bank of the macaque STS, which contains regions that respond selectively to face motion (Fisher and Freiwald, 2015; Freiwald et al., 2016). The upper bank of the macaque STS primarily comprises a polysensory region, the superior temporal polysensory area (STP, also termed TPO; Bruce et al., 1981; Seltzer and Pandya, 1978), which contains neurons responsive to faces and vocal sounds, some of which show multimodal interactions (Barraclough et al., 2005; Ghazanfar et al., 2008; Perrodin et al., 2014). Thus, the claim that the dorsal face processing stream is multimodal is generally consistent with the anatomical positioning of macaque face motion areas. However, macaque fMRI studies on responses to vocal sounds have yielded mixed results within the STP (Gil-da-Costa et al., 2006; Joly et al., 2012; Petkov et al., 2008), with responses observed primarily within the superior temporal plane and posterior STP, not consistent in location with face-motion responses. Thus, while an evolutionary relationship between face-motion-sensitive areas of macaque STP and human STS remains plausible, it is not clear whether the macaque STP contains subregions with selective, fMRI-detectable responses to both face motion and vocal sounds, as we observe here in humans. Future studies should test this by directly measuring responses to face movements and vocal sounds within individual macaques.

Can the response profiles reported here be accounted for by differences across categories in low-level visual or acoustic features? Face motion videos had lower motion energy than hand movement or object videos, suggesting against the possibility that face responses were driven by motion per se (Fig. S1). While different categories of auditory stimuli were reasonably well matched on frequency content, categories differed somewhat in spectrotemporal modulation, with stronger 2–4 Hz modulation for speech stimuli (Fig. S2). Thus, we can't rule out the possibility that responses were influenced by differences in acoustic properties. However, the response profile of the fSTS across multiple categories—a strong response to speech, nonspeech communicative, and noncommunicative vocal sounds, and weak response to music and nonvocal environmental sounds (Deen et al., 2015)—is not easily accounted for in terms of responses to spectral or temporal modulation. Furthermore, decoding of communicativeness from fSTS patterns generalized across auditory and visual modalities, and thus can't be explained by low-level features. Could the heightened vSTS response to speech over nonspeech vocal sounds simply reflect the spectrotemporal complexity of speech? Recent work has found that speech responses in middle STS/STG are substantially reduced to synthetic sounds matched in spectrotemporal modulation statistics, suggesting against this explanation (Norman-Haignere and McDermott, 2018).

Could effects attributed here to communicativeness relate to a different high-level factor? The distinction between communicative and noncommunicative stimuli overlaps with several other distinctions, such as social relevance and emotionality, which are difficult to dissociate. Thus, while we describe our results in terms of effects of communicativeness, they could equally well reflect another of these high-level distinctions. This point is particularly relevant for our MVPA results, where the distinction drives a difference in responses. Importantly, this does not diminish the claim that the fSTS represents an abstract social dimension crossmodally.

Are the fSTS responses reported here contingent on the behavioral task used in the scanner? Here, we used a task that is unrelated to the stimulus distinctions of interest—a 1-back task on individual video/audio clips—to ensure that differences in response across categories cannot be explained by task effects. However, prior studies have found a modest influence of task on pSTS responses to visually presented faces, with stronger responses when participants attend to gaze direction or facial expression than to identity (Bernstein et al., 2018; Hoffman and Haxby, 2000). Future studies should investigate fSTS responses to audiovisual social stimuli in during tasks involving social perceptual inference.

Lastly, we note that while our ICA results show that face/voice and speech responses constitute dominant response profiles across the STS, they of course don't rule out the possibility that other meaningful response profiles exist within this large region. Response profiles that ac-

count for a small amount of variance in STS-wide responses, or that don't satisfy the model's assumption of spatial orthogonality of voxel weights among components, could have been missed by this method. Furthermore, our ability to identify dominant sources of response variance is intrinsically constrained by the stimulus set chosen: there could be features driving STS variance that don't vary across the particular stimuli used here. Thus, the current results shouldn't be considered a full characterization of response variability to audiovisual face actions within the STS, but rather an assessment of dominant response profiles to a set of broad categories that capture multiple theoretically relevant dimensions.

In sum, we find that the face-responsive region of posterior STS responds to a range of face movements and vocal sounds, while the voice-responsive region of middle STS responds selectively to speech sounds. Spatial patterns of response in the fSTS differentiated communicative and noncommunicative stimuli across modalities (faces and voices), demonstrating that this region encodes an abstract social feature cross-modally. Future research should further detail the nature of representations of dynamic facial and vocal signals in these regions.

CRediT authorship contribution statement

Ben Deen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Rebecca Saxe:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Nancy Kanwisher:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Funding acquisition.

Acknowledgements

This research was funded by the NSF Center for Brains, Minds, and Machines (CCF- 1231216). B.D. was supported by an NSF graduate research fellowship and the Helen Hay Whitney Fellowship. The authors declare no competing financial interests.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2020.117191.

References

Allison, T., Puce, A., McCarthy, G., 2000. Social perception from visual cues: role of the STS region. Trends Cogn. Sci. (Regul. Ed.) 4, 267–278.
Anzellotti, S., Caramazza, A., 2017. Multimodal representations of person identity individuated with fMRI. Cortex 89, 85–97.
Barraclough, N.E., Xiao, D., Baker, C.I., Oram, M.W., Perrett, D.I., 2005. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. J. Cogn. Neurosci. 17, 377–391.
Beauchamp, M.S., Lee, K.E., Argall, B.D., Martin, A., 2004. Integration of auditory and visual information about objects in superior temporal sulcus. Neuron 41, 809–824.
Beauchamp, M.S., Nath, A.R., Pasalar, S., 2010. fMRI-guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. J. Neurosci. 30, 2414–2417.
Beauchamp, M.S., Yasar, N.E., Frye, R.E., Ro, T., 2008. Touch, sound and vision in human superior temporal sulcus. Neuroimage 41, 1011–1020.
Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal-lobe response to vocal sounds. Cognitive Brain Res. 13, 17–26.
Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.
Bernstein, M., Erez, Y., Blank, I., Yovel, G., 2018. An integrated neural framework for dynamic and static face processing. Sci. Rep. 8, 1–10.
Bernstein, M., Yovel, G., 2015. Two neural pathways of face processing: a critical evaluation of current models. Neurosci. Biobehav. Rev. 55, 536–546.
Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S., Springer, J.A., Kaufman, J.N., Possing, E.T., 2000. Human temporal lobe activation by speech and nonspeech sounds. Cereb. Cortex 10, 512–528.
Blank, H., von Kriegstein, K., 2013. Mechanisms of enhancing visual–speech recognition by prior auditory information. Neuroimage 65, 109–118.
Brass, M., Schmitt, R.M., Spengler, S., Gergely, G., 2007. Investigating action understanding: inferential processes versus action simulation. Curr. Biol. 17, 2117–2121.
Brett, M., Johnsrude, I.S., Owen, A.M., 2002. The problem of functional localization in the human brain. Nat. Rev. Neurosci. 3, 243–249.
Bruce, C., Desimone, R., Gross, C.G., 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J. Neurophysiol. 46, 369–384.
Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M., Vatikiotis-Bateson, E., 2004. Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. J. Cogn. Neurosci. 16, 805–816.
Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D., David, A.S., 1997. Activation of auditory cortex during silent lipreading. Science 276, 593–596.
Capek, C.M., MacSweeney, M., Woll, B., Waters, D., McGuire, P.K., David, A.S., Brammer, M.J., Campbell, R., 2008. Cortical circuits for silent speechreading in deaf and hearing people. Neuropsychologia 46, 1233–1241.
Deen, B., Koldewyn, K., Kanwisher, N., Saxe, R., 2015. Functional organization of social perception and cognition in the superior temporal sulcus. Cereb. Cortex 25, 4596–4609.
Deen, B., Saxe, R., 2019. Parts-based representations of perceived face movements in the superior temporal sulcus. Hum. Brain Mapp. 40, 2499–2510.
DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73, 415–434.
DiNicola, L.M., Braga, R.M., Buckner, R.L., 2020. Parallel distributed networks dissociate episodic and social functions within the individual. J. Neurophysiol. 123, 1144–1179.
Fecteau, S., Armony, J.L., Joanette, Y., Belin, P., 2004. Is voice processing species-specific in human auditory cortex? An fMRI study. Neuroimage 23, 840–848.
Fedorenko, E., Duncan, J., Kanwisher, N., 2013. Broad domain generality in focal regions of frontal and parietal cortex. Proc. Natl. Acad. Sci. 110, 16616–16621.
Fisher, C., Freiwald, W.A., 2015. Contrasting specializations for facial motion within the macaque face-processing system. Curr. Biol. 25, 261–266.
Fletcher, P.C., Happe, F., Frith, U., Baker, S.C., Dolan, R.J., Frackowiak, R.S., Frith, C.D., 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. Cognition 57, 109–128.
Freiwald, W., Duchaine, B., Yovel, G., 2016. Face processing systems: from neurons to real-world social perception. Annu. Rev. Neurosci. 39, 325–346.
Ghazanfar, A.A., Chandrasekaran, C., Logothetis, N.K., 2008. Interactions between the superior temporal sulcus and auditory cortex mediate dynamic face/voice integration in rhesus monkeys. J. Neurosci. 28, 4457–4469.
Gil-da-Costa, R., Martin, A., Lopes, M.A., Munoz, M., Fritz, J.B., Braun, A.R., 2006. Species-specific calls activate homologs of Broca's and Wernicke's areas in the macaque. Nat. Neurosci. 9, 1064–1070.
Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48, 63.
Hasan, B.A.S., Valdes-Sosa, M., Gross, J., Belin, P., 2016. "Hearing faces and seeing voices": amodal coding of person identity in the human brain. Sci Rep 6, 37494.
Haxby, J., Gobbini, M., Furey, M., Ishai, A., Shouten, J., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Haxby, J.V., Hoffman, E.A., Gobbini, M.I., 2000. The distributed human neural system for face perception. Trends Cogn. Sci. (Regul. Ed.) 4, 223–233.
Hein, G., Doehrmann, O., Müller, N.G., Kaiser, J., Muckli, L., Naumer, M.J., 2007. Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. J. Neurosci. 27, 7881–7887.
Hoffman, E.A., Haxby, J.V., 2000. Distinct representations of eye gaze and identity in the distributed human neural system for face perception. Nat. Neurosci. 3, 80–84.
Hyvärinen, A., Oja, E., 2000. Independent component analysis: algorithms and applications. Neural Netw. 13, 411–430.
Joly, O., Ramus, F., Pressnitzer, D., Vanduffel, W., Orban, G.A., 2012. Interhemispheric differences in auditory processing revealed by fMRI in awake rhesus monkeys. Cereb. Cortex 22, 838–853.
Koster-Hale, J., Bedny, M., Saxe, R., 2014. Thinking about seeing: perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. Cognition 133, 65–78.
Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., Saxe, R., 2017. Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. Neuroimage 161, 9–18.
Koster-Hale, J., Saxe, R., 2013. Theory of mind: a neural prediction problem. Neuron 79, 836–848.
Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., Wildgruber, D., 2009. Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice-and face-sensitive regions in the superior temporal sulcus. Neuropsychologia 47, 3059–3066.
Latinus, M., McAleer, P., Bestelmeyer, P.E., Belin, P., 2013. Norm-based coding of voice identity in human auditory cortex. Curr. Biol. 23, 1075–1080.
Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A., 2005. Neural substrates of phonemic perception. Cereb. Cortex 15, 1621–1631.
Marchini, J.L., Ripley, B.D., 2000. A new statistical approach to detecting significant activation in functional MRI. Neuroimage 12, 366–380.
Margulies, D.S., Ghosh, S.S., Goulas, A., Falkiewicz, M., Huntenburg, J.M., Langs, G., Bezgin, G., Eickhoff, S.B., Castellanos, F.X., Petrides, M., 2016. Situating the default-mode network along a principal gradient of macroscale cortical organization. Proc. Natl. Acad. Sci. 113, 12574–12579.
Marques, L.M., Lapenta, O.M., Merabet, L.B., Bolognini, N., Boggio, P.S., 2014. Tuning and disrupting the brain—Modulating the McGurk illusion with electrical stimulation. Front. Hum. Neurosci. 8, 533.
McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.
Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F., 2014. Phonetic feature encoding in human superior temporal gyrus. Science 343, 1006–1010.

Noesselt, T., Rieger, J.W., Schoenfeld, M.A., Kanowski, M., Hinrichs, H., Heinze, H.-J., Driver, J., 2007. Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. J. Neurosci. 27, 11431–11441.

Norman-Haignere, S., Kanwisher, N.G., McDermott, J.H., 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. Neuron 88, 1281–1296.

Norman-Haignere, S.V., McDermott, J.H., 2018. Neural responses to natural and model-matched stimuli reveal distinct computations in primary and nonprimary auditory cortex. PLoS Biol. 16, e2005127.

O'Toole, A.J., Roark, D.A., Abdi, H., 2002. Recognizing moving faces: a psychological and neural synthesis. Trends Cogn. Sci. (Regul. Ed.) 6, 261–266.

Overath, T., McDermott, J.H., Zarate, J.M., Poeppel, D., 2015. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat. Neurosci. 18, 903–911.

Peelen, M.V., Atkinson, A.P., Vuilleumier, P., 2010. Supramodal representations of perceived emotions in the human brain. J. Neurosci. 30, 10127–10134.

Pelphrey, K.A., Morris, J.P., Mccarthy, G., 2004. Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. J. Cogn. Neurosci. 16, 1706–1716.

Pelphrey, K.A., Morris, J.P., Michelich, C.R., Allison, T., McCarthy, G., 2005. Functional anatomy of biological motion perception in posterior temporal cortex: an fMRI study of eye, mouth and hand movements. Cereb. Cortex 15, 1866–1876.

Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E., Watson, R.H., Fleming, D., Crabbe, F., Valdes-Sosa, M., Belin, P., 2015. The human voice areas: spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage 119, 164–174.

Perrodin, C., Kayser, C., Logothetis, N.K., Petkov, C.I., 2014. Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. J. Neurosci. 34, 2524–2537.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nat. Neurosci. 11, 367–374.

Pitcher, D., Dilks, D.D., Saxe, R.R., Triantafyllou, C., Kanwisher, N., 2011. Differential selectivity for dynamic versus static information in face-selective cortical regions. Neuroimage 56, 2356–2363.

Poldrack, R.A., 2017. Precision neuroscience: dense sampling of individual brains. Neuron 95, 727–729.

Puce, A., Allison, T., Bentin, S., Gore, J.C., McCarthy, G., 1998. Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci. 18, 2188–2199.

Redcay, E., 2008. The superior temporal sulcus performs a common function for social and speech perception: implications for the emergence of autism. Neurosci. Biobehav. Rev. 32, 123–142.

Redcay, E., Velnoskey, K.R., Rowe, M.L., 2016. Perceived communicative intent in gesture and language modulates the superior temporal sulcus. Hum Brain Mapp 37, 3444–3461.

Reisberg, D., Mclean, J., Goldfield, A., 1987. Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In: Dodd, B., Campbell, R. (Eds.), Hearing By eye: The psychology of Lip-Reading. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, pp. 97–113.

Riedel, P., Ragert, P., Schelinski, S., Kiebel, S.J., von Kriegstein, K., 2015. Visual face–movement sensitive cortex is relevant for auditory-only speech recognition. Cortex 68, 86–99.

Said, C.P., Moore, C.D., Engell, A.D., Todorov, A., Haxby, J.V., 2010. Distributed representations of dynamic facial expressions in the superior temporal sulcus. J Vis 10, 11.

Saxe, R., Brett, M., Kanwisher, N., 2006. Divide and conquer: a defense of functional localizers. Neuroimage 30, 1088–1096.

Saxe, R., Kanwisher, N., 2003. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". Neuroimage 19, 1835–1842.

Saxe, R., Xiao, D.-K., Kovacs, G., Perrett, D., Kanwisher, N., 2004. A region of right posterior superior temporal sulcus responds to observed intentional actions. NeuropsychologiaNeuropsychologia 42, 1435–1446.

Schultz, J., Brockhaus, M., Bülthoff, H.H., Pilz, K.S., 2013. What the human brain likes about facial motion. Cereb. Cortex 23, 1167–1178.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J., 2000. Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123, 2400–2406.

Seltzer, B., Pandya, D.N., 1978. Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. Brain Res. 149, 1–24.

Shultz, S., Vouloumanos, A., Pelphrey, K., 2012. The superior temporal sulcus differentiates communicative and noncommunicative auditory signals. J. Cogn. Neurosci. 24, 1224–1232.

Skerry, A.E., Saxe, R., 2014. A Common Neural Code for Perceived and Inferred Emotion. J. Neurosci. 34, 15997–16008.

Skerry, A.E., Saxe, R., 2015. Neural representations of emotion are organized around abstract event features. Curr. Biol. 25, 1945–1954.

Srinivasan, R., Golomb, J.D., Martinez, A.M., 2016. A neural basis of facial action recognition in humans. J. Neurosci. 36, 4434–4442.

Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am. 26, 212–215.

Thompson, J.C., Hardee, J.E., Panayiotou, A., Crewther, D., Puce, A., 2007. Common and distinct brain activation to viewing dynamic sequences of face and hand movements. Neuroimage 37, 966–973.

Tsantani, M., Kriegeskorte, N., McGettigan, C., Garrido, L., 2019. Faces and voices in the brain: a modality-general person-identity representation in superior temporal sulcus. Neuroimage 201, 116004.

Van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L., 2004. Integration of letters and speech sounds in the human brain. Neuron 43, 271–282.

von Kriegstein, K., Dogan, Ö., Grüter, M., Giraud, A.-L., Kell, C.A., Grüter, T., Kleinschmidt, A., Kiebel, S.J., 2008. Simulation of talking faces in the human brain improves auditory speech recognition. Proc. Natl. Acad. Sci. 105, 6747–6752.

von Kriegstein, K., Smith, D.R., Patterson, R.D., Ives, D.T., Griffiths, T.D., 2007. Neural representation of auditory size in the human voice and in sounds from other resonant sources. Curr. Biol. 17, 1123–1128.

von Kriegstein, K., Smith, D.R., Patterson, R.D., Kiebel, S.J., Griffiths, T.D., 2010. How the human brain recognizes speech in the context of changing speakers. J. Neurosci. 30, 629–638.

Vouloumanos, A., Kiehl, K.A., Werker, J.F., Liddle, P.F., 2001. Detection of sounds in the auditory stream: event-related fMRI evidence for differential activation to speech and nonspeech. J. Cogn. Neurosci. 13, 994–1005.

Watson, R., Latinus, M., Charest, I., Crabbe, F., Belin, P., 2014a. People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. Cortex 50, 125–136.

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., Belin, P., 2014b. Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face-Voice Emotional Integration. J. Neurosci. 34, 6813–6821.

Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. Neuroimage 14, 1370–1386.

Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., McCarthy, G., 2003. Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb. Cortex 13, 1034–1043.

Yovel, G., O'Toole, A.J., 2016. Recognizing people in motion. Trends Cogn. Sci. (Regul. Ed.) 20, 383–395.

Zhu, L.L., Beauchamp, M.S., 2017. Mouth and voice: a relationship between visual and auditory preference in the human superior temporal sulcus. J. Neurosci. 37, 2697–2708.

Organization of high-level visual cortex in human infants

Article *in* Nature Communications · January 2017

DOI: 10.1038/ncomms13995

CITATIONS
112

READS
261

8 authors, including:

Ben Deen
The Rockefeller University
24 PUBLICATIONS **2,349** CITATIONS
SEE PROFILE

Hilary Richardson
Massachusetts Institute of Technology
21 PUBLICATIONS **400** CITATIONS
SEE PROFILE

Daniel D Dilks
Emory University
39 PUBLICATIONS **1,670** CITATIONS
SEE PROFILE

Boris Keil
Technische Hochschule Mittelhessen
110 PUBLICATIONS **2,976** CITATIONS
SEE PROFILE

nature COMMUNICATIONS

ARTICLE

Received 2 Aug 2016 | Accepted 18 Nov 2016 | Published 10 Jan 2017

DOI: 10.1038/ncomms13995 OPEN

Organization of high-level visual cortex in human infants

Ben Deen[1], Hilary Richardson[1], Daniel D. Dilks[1,2], Atsushi Takahashi[1], Boris Keil[3,4], Lawrence L. Wald[3,5], Nancy Kanwisher[1] & Rebecca Saxe[1]

How much of the structure of the human mind and brain is already specified at birth, and how much arises from experience? In this article, we consider the test case of extrastriate visual cortex, where a highly systematic functional organization is present in virtually every normal adult, including regions preferring behaviourally significant stimulus categories, such as faces, bodies, and scenes. Novel methods were developed to scan awake infants with fMRI, while they viewed multiple categories of visual stimuli. Here we report that the visual cortex of 4–6-month-old infants contains regions that respond preferentially to abstract categories (faces and scenes), with a spatial organization similar to adults. However, precise response profiles and patterns of activity across multiple visual categories differ between infants and adults. These results demonstrate that the large-scale organization of category preferences in visual cortex is adult-like within a few months after birth, but is subsequently refined through development.

[1] Department of Brain and Cognitive Sciences and McGovern Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [2] Department of Psychology, Emory University, Atlanta, Georgia 30322, USA. [3] Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Harvard Medical School, Massachusetts General Hospital, Charlestown, Massachusetts 02129, USA. [4] Institute of Medical Physics and Radiation Protection, Department of Life Science Engineering, Mittelhessen University of Applied Science, Giessen 35390, Germany. [5] Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. Correspondence and requests for materials should be addressed to B.D. (email: benjamin.deen@gmail.com).

In human adults, the cortex shows a systematic spatial and functional organization. Responses in visual cortex are driven by high-level, behaviourally relevant categories, including human faces, bodies, objects and natural scenes, both within circumscribed, highly selective regions[1–4], and in graded response patterns across larger swaths of cortex[5–7]. The origins of these responses have been the topic of intense debate: are they learned, reflecting a gradual accrual of expertise, or do they reflect innate predispositions?

A key constraint on theories of cortical development would be evidence of when these responses emerge in cortex. However, the functional organization of high-level responses in visual cortex has never been tested in infants, and existing indirect evidence makes contradictory predictions. Slow, hierarchical development of visual functions over years is suggested by late developmental change in children aged 4–10 years[8,9], slow and staggered time courses of myelination[10] and cortical thinning[11], and late developmental change in juvenile macaques[12,13]. By contrast, early functional maturation of cortex in infancy is consistent with high-level responses measured by electroencephalography (EEG)[14,15] and near-infrared spectroscopy (NIRS)[16,17], rare electrophysiological recordings from infant macaques[18], and the sophisticated cognition of pre-verbal infants revealed by the modern developmental psychology[19].

The main obstacle to resolving this debate is the difficulty of neuroimaging awake infants. The imaging techniques most commonly used in human infants (EEG and fNIRS) lack the coverage and resolution needed to measure the spatial organization of cortex. Only two prior studies have collected functional magnetic resonance imaging (fMRI) data from awake infants, and because of infants' limited tolerance, it has been difficult to collect sufficient data to test replicability or functional profiles of response[20,21]. Here we implement novel methods for awake infant fMRI to study the early development of high-level visual responses in cortex. We employ a number of technical advances to increase participant comfort, optimize signal strength and minimize head motion artefacts: (1) infant-sized MR head coils; (2) quiet pulse sequences; (3) dynamic and engaging visual stimuli; and (4) a combination of extant and novel data analysis techniques for minimizing motion artefacts.

Our data demonstrate that by 4–6 months of age, human infants have category-sensitive visual responses to faces and scenes, with a spatial organization mimicking that observed in adults. However, we also observe differences: both in response profiles across multiple categories (which were less selective in infants), and in patterns of response across cortex. Thus, the overall functional organization of high-level visual cortex develops very early, and is subsequently refined.

Results

fMRI findings. We obtained low-motion fMRI data from 9 infants (of 17 tested; age 3–8 months; Supplementary Table 1), while they viewed engaging, brightly coloured, infant-friendly movies of faces, natural scenes, scrambled scenes, human bodies and objects (Supplementary Fig. 1). We first compared responses to faces versus scenes, because in adults this comparison yields the most robust differential responses, and delineates a large-scale spatial organization of extrastriate cortex[22,23]. Face- or scene-preferring regions in occipitotemporal cortex were observed in eight of nine infants, with a similar spatial organization as in adults (Fig. 1; Supplementary Figs 2 and 3). In individual infants, face-preferring regions were observed in the fusiform gyrus, lateral occipital cortex, superior temporal sulcus (STS) and medial prefrontal cortex; scene-preferring regions were observed in the parahippocampal gyrus and lateral occipital cortex. Many of these regions showed reliable responses in a group analysis, demonstrating generalization across infants (Fig. 1). Region-of-interest (ROI) analyses corroborated whole-brain results, demonstrating reliable face and scene preferences in data independent from those used to define ROIs, in all regions tested (Fig. 2; Supplementary Figs 4–7; Expt. 1, $n=9$, permutation test; ventral face region, $z=2.85$, $P=2.2\times10^{-3}$ lateral face region, $z=3.27$, $P=5.4\times10^{-4}$; STS face region, $z=4.74$, $P=1.1\times10^{-6}$; ventral scene region, $z=6.41$, $P=7.3\times10^{-11}$; and lateral scene region, $z=3.43$, $P=3.0\times10^{-4}$). In six infants who participated in more than one experiment, these preferences were also replicated using distinct face and scene movies (Expts. 2–8, $n=6$, permutation test; ventral face region, $z=2.22$, $P=0.013$; lateral face region, $z=2.51$, $P=6.0\times10^{-3}$; STS face region, $z=4.19$, $P=1.4\times10^{-5}$; ventral scene region, $z=5.64$, $P=8.5\times10^{-9}$; and lateral scene region, $z=5.00$, $P=2.9\times10^{-7}$).

These results demonstrate that the spatial organization of preferential responses to faces versus scenes is similar in 4–6-month-old infants and in adults, extending throughout the ventral visual stream and even into prefrontal cortex. In subsequent analyses, we sought to constrain the functional interpretation of these responses. Are cortical regions in infants responding to highly specific visual categories[1,2,4], to broader visual or semantic dimensions[5,6], or to lower-level visual features that co-vary with high-level categories[24–27]? Do large-scale patterns of response to categories other than faces and scenes change over development? Measuring responses to multiple visual categories enabled us to ask these questions.

Do preferential responses to faces and scenes in infants reflect a high-level category preference, or a bias toward lower-level visual features, such as eccentricity, spatial frequency or rectilinearity (the presence of 90° angles)[24–27]? We tested whether cortical responses in infants were better predicted by these lower-level visual features than by high-level categories. In Experiment 2, scenes and scrambled scenes were reduced to 80% the size of face and body movies, but category preferences were unaffected, suggesting that these responses were not driven by eccentricity (Supplementary Fig. 7B). Across all experiments, rectilinearity and spatial frequency content of the movies predicted responses no better, and in scene regions significantly worse than modulation by visual category (Fig. 3). The visual category model was particularly better for scene-preferring regions because the control condition in most experiments (scrambled scenes) had high spatial frequency and high rectilinearity, most clearly differentiating the predictions of the lower-level features from the visual category model. For face-preferring regions, category and low-level feature models made similar predictions for these stimuli; future experiments including a low-spatial-frequency, highly-curvilinear control condition will clarify the responses of these regions. Overall, however, our data suggest that by 4–6 months, category-sensitive cortical responses are not primarily driven by lower-level visual features.

Another outstanding question is whether responses to faces and scenes in infants reflect regions with highly selective responses to specific categories, or weaker-graded preferences across multiple categories. In adults, for example, the fusiform face area and parahippocampal place area have highly selective response profiles, preferring faces or scenes to any other visual category (for example, objects, bodies, animals, foods and so on)[2,4], whereas broad areas around these regions have graded preferences predicted by coarser semantic dimensions[5,6]. We searched for highly selective regions by contrasting faces (or scenes) to objects. In adults, these contrasts revealed the predicted spatially focal, strongly selective regions: each region

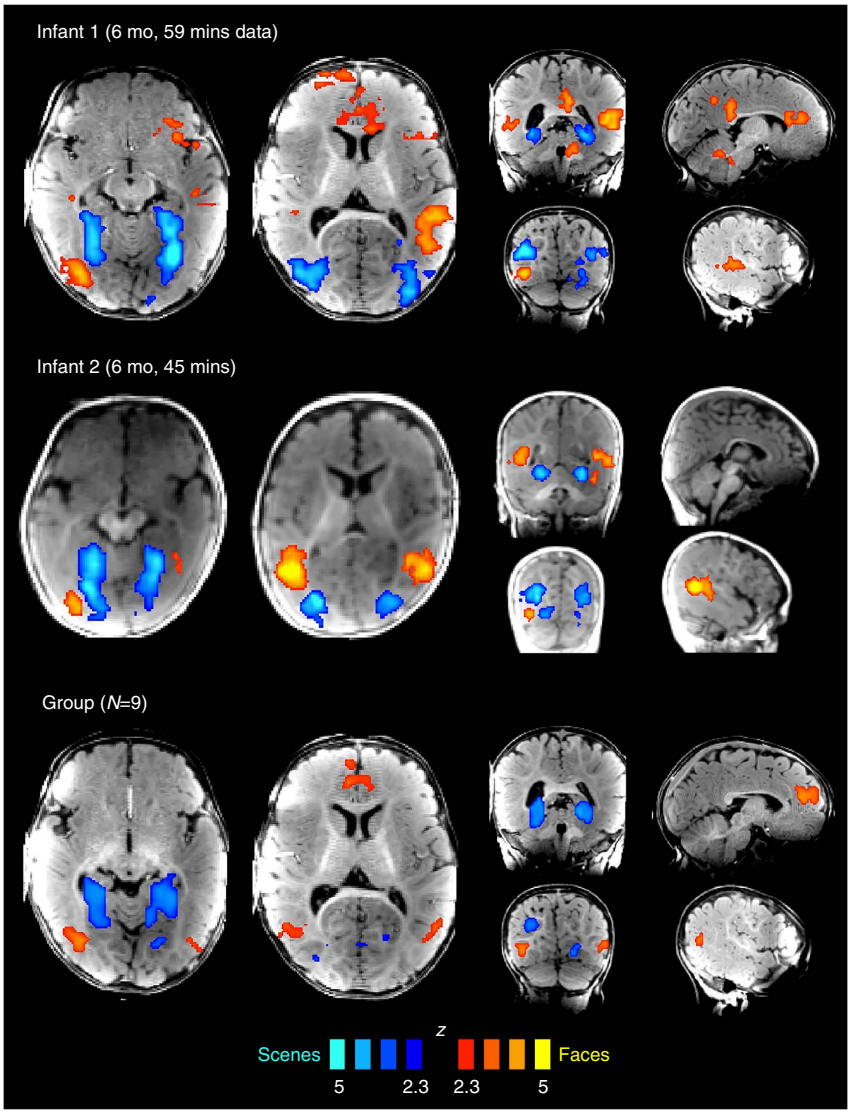


Figure 1 | Category-sensitive responses to faces and scenes in infants show adult-like spatial organization. Regions preferring faces over scenes are reported in red/yellow, and regions preferring scenes over faces in blue. The top two rows of whole-brain activation maps show results from the two individual infants with the largest amount of usable data, while the third shows a group map with statistics across infants. Maps are thresholded at $P<0.01$ voxelwise, and corrected for multiple comparisons using a clusterwise threshold of $P<0.05$.

showed a higher response to its preferred category than to all three other categories (Fig. 4; Supplementary Fig. 3; permutation test comparing faces or scenes to objects, $n=3$; ventral face region, $z=5.54$, $P=1.5\times10^{-8}$; lateral face region, $z=4.35$, $P=6.8\times10^{-6}$; STS face region, $z=7.13$, $P=5.0\times10^{-13}$; ventral scene region, $z=6.10$, $P=5.3\times10^{-10}$; and lateral scene region, $z=5.12$, $P=1.5\times10^{-7}$). In infants, however, no region showed a higher response to faces or scenes over objects (permutation test, $n=6$; ventral face region, $z=-0.75$, $P=0.77$; lateral face region, $z=0.91$, $P=0.18$; STS face region, $z=1.40$, $P=0.08$; ventral scene region, $z=-1.36$, $P=0.91$; and lateral scene region, $z=0.81$, $P=0.21$). Similar results were obtained for a range of ROI sizes (Fig. 4): adults showed a significant response to faces (or scenes) over objects for all regions and ROI sizes (permutation test, $n=3$, all P's<0.05), while infants did not show a significant response for any region or ROI size, including ROIs as small as 0.8 cm^3 (permutation test, $n=6$, all P's>0.05). Thus, within the spatial resolution of our methods, we find no evidence that the difference between groups reflects a change in the size of selective regions.

Could these null findings result simply from poor data quality in infants? Several observations argue against this interpretation. First, standard errors did not differ substantially across infants and adults, and when a reduced subset of adult data was analysed to inflate standard errors, the same results were obtained (Supplementary Fig. 8). Second, although no region preferred faces (or scenes) to objects in infants, the reverse contrast in exactly the same data revealed robust responses to objects, compared with either faces or scenes, with adult-like spatial organization in temporal and parietal cortex (Supplementary Fig. 9). Thus, while the large-scale spatial organization of responses to faces versus scenes is present in infants and remains a principal dimension of cortical organization into adulthood, highly selective regions for particular categories apparently emerge later in development, perhaps requiring more extensive visual experience.

In addition to the absence of category-selective regions, we found evidence for developmental change in the large-scale patterns of functional response across multiple categories. To summarize and quantify the spatial structure of responses

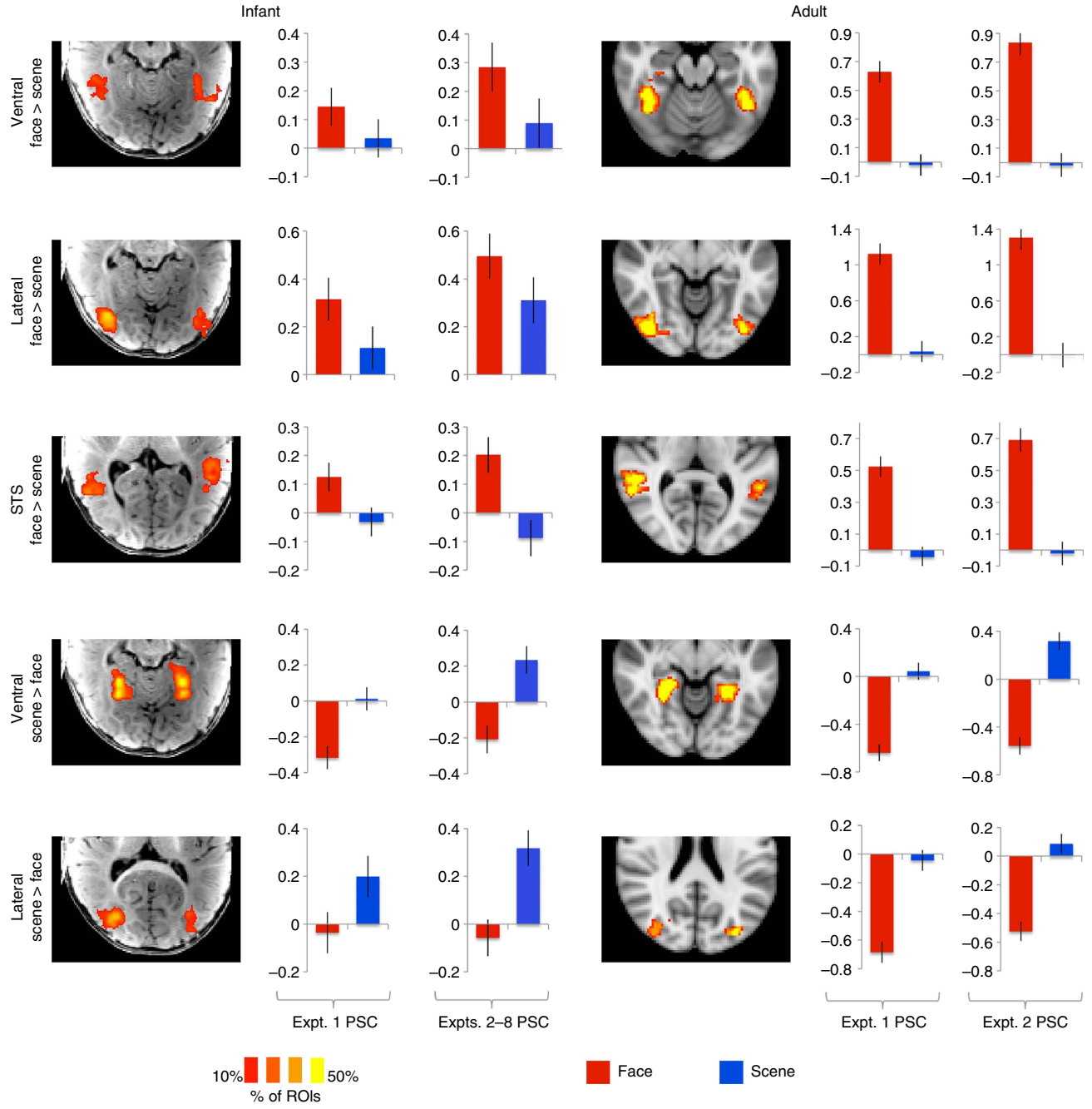


Figure 2 | The location and reliability of responses to faces and scenes is consistent across infants and adults. Brain images show heat maps of region-of-interest (ROI) locations across participants (% of ROIs that included a given voxel), with ROIs defined as the top 5% of voxels responding to faces over scenes (or vice versa) within an anatomical region. Bar plots show each ROI's response (per cent signal change, PSC) to faces and scenes in independent data, separately for Expts. 1, 2–8. Error bars show the standard deviation of a permutation-based null distribution for the corresponding value. Baseline corresponds to the response to scrambled scenes (Expts. 1–3, 7–8) or scrambled objects (Expts. 4–6). Statistics for infant data are presented in the main text; as expected, face and scene preferences were highly significant in adults for all regions (permutation test, $n=3$; all P's $<10^{-15}$).

to multiple categories, we computed representational similarity matrices, capturing the similarity of spatial patterns of response across categories[28]. While face and scene responses were dissimilar in both groups, consistent with the results above, the pattern of similarity across all categories differed between infants and adults (Fig. 5; Supplementary Fig. 10). Representational similarity matrices across the four categories were highly similar within adults ($n=3$, mean Kendall's tau $=0.91$), and moderately similar within infants ($n=6$, mean Kendall's tau $=0.41$), but dissimilar between groups (mean Kendall's tau $=0.14$; significantly lower than within group similarity for both infants, $P=0.024$, and adults, $P=0.012$, permutation test). Thus visual responses to multiple categories differ in infants and adults, as measured both by response profiles of focal regions, and distributed patterns of response across cortex.

Discussion

Using novel methods to acquire and analyse fMRI data from awake human infants, this study demonstrates that the cortex of

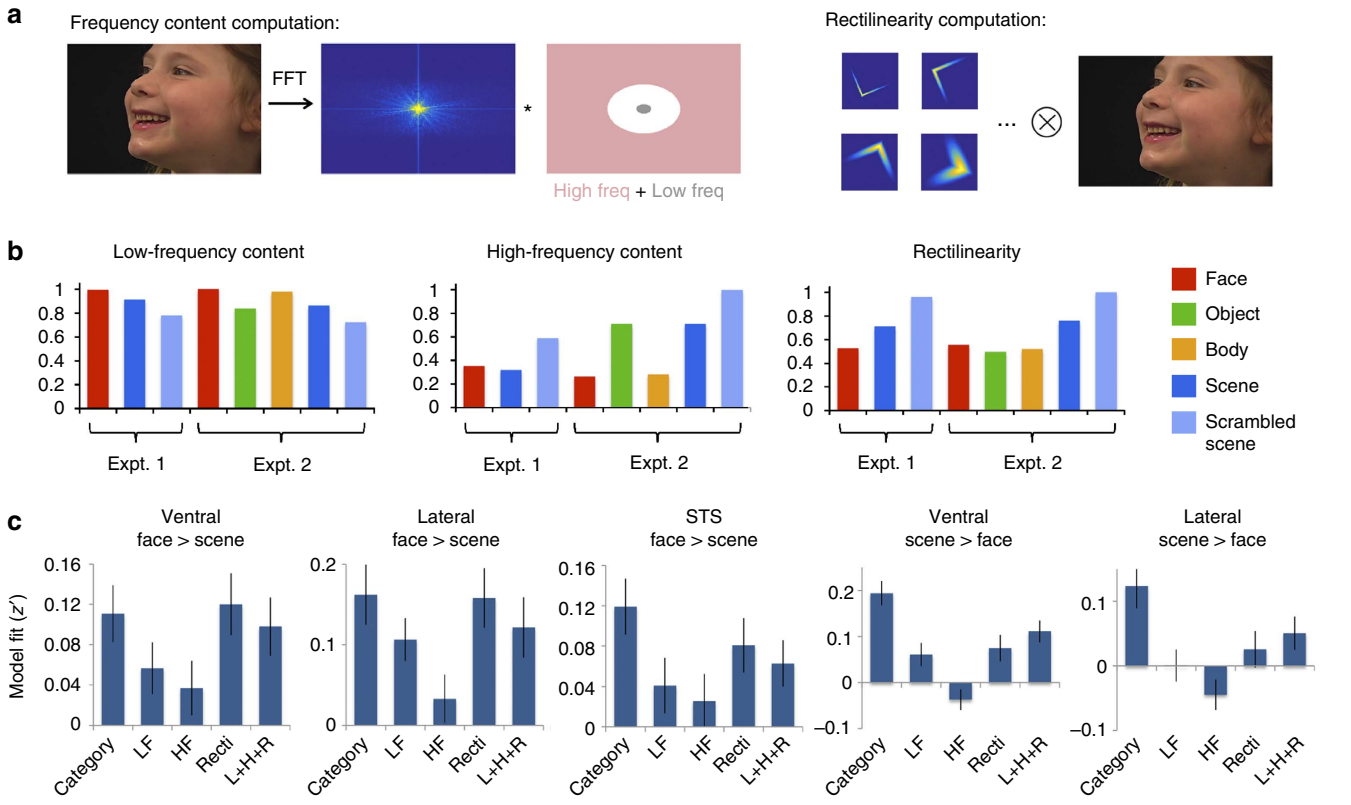


Figure 3 | Comparison of categorical and visual feature-based models of region-of-interest (ROI) responses. (**a**) Schematic showing how high- and low-frequency content and rectilinearity were computed from movie frames. (**b**) Mean values of these visual features across the stimuli used in Expts. 1–2, normalized such that the maximum value across categories is set to 1. (**c**) Model fits of category and visual feature models to ROI responses, with error bars specifying standard error. In all three face-preferring regions, there was no significant difference between the category model and the best-performing visual feature model (ventral face region, $t(54) = -0.48$, $P = 0.64$; lateral face region, $t(54) = 0.25$, $P = 0.80$; STS face region, $t(54) = 1.55$, $P = 0.13$). In these regions, the category model (including a high response to faces) and the rectinilinearity model (a low response to rectilinearity) made very similar predictions; other types of stimuli (such as curve-scrambled faces) may be needed to distinguish these hypotheses. In contrast, for scene regions, the category model and low-level feature models made distinct predictions due to the inclusion of a highly rectilinear non-scene condition (grid-scrambled movies). For the two scene-preferring regions, the category model significantly outperformed all visual feature models. For brevity, we report statistics only for the comparison with the best-performing model (ventral scene region, $t(54) = 3.56$, $P = 7.8 \times 10^{-4}$; lateral scene region, $t(54) = 2.56$, $P = 0.013$). HF, high-frequency content; L + H + R = low-frequency content, high-frequency content and rectilinearity; LF, low-frequency content; Recti, rectilinearity.

4–6-month-old human infants is already spatially organized, with distinct regions responding preferentially to human faces versus natural scenes. The spatial structure of these responses is very similar to that observed in adults, and extends throughout cortex, including occipital, temporal, parietal and frontal regions. Thus, while the anatomical maturation of human cortex is slow and asynchronous, basic aspects of functional organization are present across cortex from a very early age.

Prior fMRI studies have observed category-sensitive responses in high-level visual cortex in children as young as 4 years[8]. By demonstrating that these responses exist by 4–6 months of age, the current study provides a stronger constraint on theories of cortical development: this functional organization must either be determined innately, without any need for visual experience, or develop within the first few months of life. A limited role for visual experience in the development of category-sensitive responses is consistent with evidence that in congenitally blind adults, category-sensitive responses in visual cortex develop in the absence of any visual input[29,30].

The observation of face-sensitive functional responses in human infants is also consistent with prior evidence from EEG and NIRS[14–17]. Using fMRI, our results go beyond those prior studies because we are able to assess the precise spatial organization of category-sensitive responses, and to measure responses in non-superficial regions, such as ventral temporal cortex. This novel evidence of the functional organization of cortex in infancy can be directly related to the extensive fMRI literature on visual responses in adults. In addition to providing spatial resolution, the current data provide better functional characterization of cortical responses in infants. By acquiring a large amount of high-quality data within individual infants, we are able to measure responses to multiple categories, and to internally replicate our finding of face and scene responses, across experiments that used different specific movie stimuli. We also provide initial evidence that infants' responses to high-level, behaviourally significant categories cannot be explained in terms of responses to simple lower-level visual features.

While our data indicate that the spatial organization of responses to faces and scenes is remarkably adult-like, we additionally observed that both the fine-grained selectivity and spatial pattern of activity across multiple categories change with age. In particular, and in contrast to adults, infants did not have strongly category-selective regions, that is, circumscribed regions showing a robustly stronger response to one category than to any other. Differences between infants and adults must be interpreted with caution, given the marked differences in brain size and general visual and cognitive function.

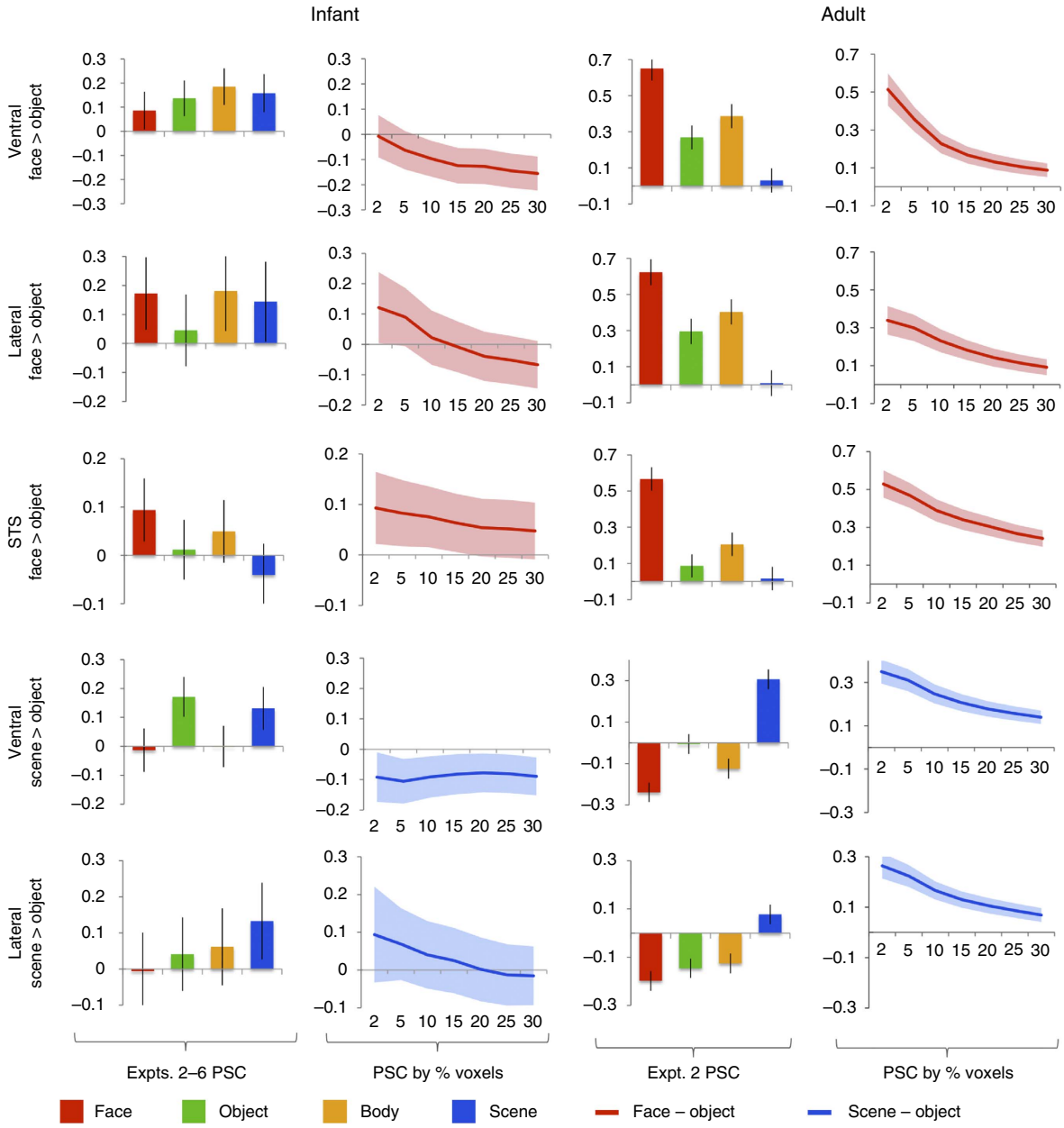


Figure 4 | Infants lack strongly category-selective regions. Region-of-interest (ROI) responses (per cent signal change, PSC, in independent data) in regions defined by comparing faces to objects and scenes to objects, in infants and adults. Bar plots show responses of ROIs defined as the top 5% of voxels within an anatomical region, while line graphs show how the difference between face and object or scene and object responses varies as a function of ROI size. Adults show strongly selective responses, substantially higher to the preferred category than any other category, while infants do not show a reliable or selective response at any ROI size. Error bars show the standard deviation of a permutation-based null distribution for the corresponding PSC value or PSC difference. Baseline corresponds to the response to scrambled scenes (Expts. 2–3) or scrambled objects (Expts. 4–6).

For instance, one possibility is that in adults, category-selective responses are enhanced by top-down feedback and selective attention, which are not yet mature in infants. Nevertheless, these data are consistent with the hypothesis that the early-developing large-scale functional organization of category preferences in cortex provides a scaffolding for subsequent refinement of responses, leading ultimately to the strongly category-selective regions observed in adults[31]. The process of refinement likely depends on both physiological maturation (for example, myelination of long-range connections between brain regions) and visual experience. For example, the visual word form area develops as a result of experience with a specific orthography[32], but is guided by pre-existing patterns of anatomical connectivity[33]. Similarly, extensive training with novel symbols can generate selective responses in a cortical region in macaques; the location of this region is consistent across animals, suggesting refinement based on a pre-existing scaffold[12,13].

These results point to myriad future questions, including: what is the time course of the development of category-selective visual regions during and after the first year of life? How do maturation and visual experience interact to drive this development? And does a similar principle (an initial preference that is subsequently refined) apply to the development of functionally specific regions in other perceptual and cognitive domains? We hope that the methods introduced here will aid in future investigations of these questions.

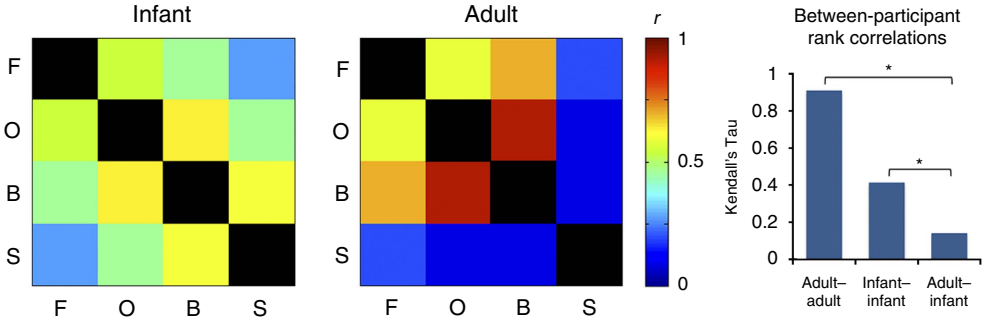


Figure 5 | Distinct representational similarity for multiple visual categories in infants and adults. Left two images show representational similarity matrices: correlations between spatial patterns of response across extrastriate visual cortex, to faces (F), objects (O), bodies (B) and scenes (S). Bar graph on the right shows rank correlations (Kendall's tau) between similarity measures from pairs of participants. Within group (adult–adult and infant–infant) rank correlations are significantly higher than between group (adult–infant) rank correlations, indicating a reliably distinct similarity structure across groups. *denotes $P<0.05$.

Methods

Participants. We scanned 17 infants (age 2.3–8.6 months, three female) and acquired useable (low-motion) data from nine infants (age 3.0–8.0 months, one female). We also scanned three adults (age 27–34 years, one female) for comparison (Supplementary Table 1). Because low-motion data from infants was relatively rare, whenever possible we scanned infants in multiple sessions (between 1 and 16 scan sessions per infant, for a total of 63 sessions). Sessions occurring within a month were analysed together as a single data set; sessions separated by more than a month were analysed as separate data sets (this occurred for five infants; only one data set per infant was used for group analyses). Adult participants and parents of infant participants provided written, informed consent, as approved by the Committee on the Use of Humans as Experimental Subjects at MIT.

Paradigm. Stimuli were infant-friendly dynamic movie clips depicting faces, objects, bodies, and scenes (Supplementary Fig. 1). Participants initially viewed Experiment 1 (Expt. 1), a two-condition (face, scene) experiment, with grid-scrambled scenes included as a baseline (pilot testing indicated that infants would not tolerate a baseline with less visual structure). When time permitted, we additionally ran Experiment 2 (Expt. 2), a four-condition (face, object, body and scene) version, with distinct face and scene movies, a scrambled scene baseline, and both scene and scrambled scene movies presented at 80% size, to minimize the possibility of a retinotopic confound in the scene versus face comparison. In certain cases, experiments (Expts. 3–8) using different movies of the same categories were used, to further test generalization of responses across specific movies; these experiments, as well as more detail on stimuli, are further described below (Further paradigm details). Stimuli were presented in 18 s-long blocks, typically comprising six 3 s-long movie clips. Baseline blocks occurred every seven blocks (Expt. 1) or five blocks (Expt. 2); experimental blocks were ordered pseudorandomly between baseline blocks. During infant functional scans, an experimenter or parent lay in the scanner bore to monitor the infant, and told the experimenters if the infant closed his or her eyes, fell asleep or fussed out. For infants, individual runs were not fixed in duration, but instead ended whenever the infant fussed out or fell asleep. For adults, runs lasted 22 blocks (Expt. 1) or 21 blocks (Expt. 2), with a baseline block at the start and end of each run. Adults received five runs each of Expt. 1 and Expt. 2. Parents of actors in stimulus videos provided written, informed consent for the publication of images in Figure 3 and Supplementary Figure 1.

Data acquisition. MRI data were acquired using a Siemens 3T MAGNETOM Tim Trio scanner (Siemens AG, Healthcare, Erlangen, Germany). We used a standard 32-channel head coil for adult participants, and a custom-built infant-sized 32-channel head coil for infants[34]. The latter was shaped like a reclined car seat to increase comfort, and had coil elements close to the infant's head, to reduce head motion and increase signal-to-noise ratio. For infants whose heads did not fit in this coil, a 32-channel head coil designed for 5 year olds was used instead. To further increase infant comfort, we acquired data using a quiet (70–72 dB sound pressure level) T2*-weighted pulse sequence[35], sensitive to blood-oxygen-level-dependent contrast (repetition time (TR) = 3 s, echo time (TE) = 43 ms, $\alpha = 90°$, field of view (FOV) = 192 mm, matrix = 64 × 64, slice thickness = 3 mm, slice gap = 0.6 mm). For infants, we used 18–24 near-axial slices, using the minimum number of slices required to cover occipitotemporal cortex for a given head size, because pulse sequence audio volume scaled with number of slices; for adults, we used 36 near-axial slices for whole-brain coverage. Infants were swaddled during all scans to reduce body movement.

Anatomical images were only collected in certain cases, because our focus was normally to collect as much awake functional data as possible, and because collecting a high-quality anatomical typically required the infant to be asleep to reduce motion. When anatomicals were collected, we used one of three T1-weighted pulse sequences of varying length, using briefer, lower-quality sequences when an infant would only stay still for a short duration. These included a 24 s sequence (TR = 283 ms, TE = 1.8 ms, flip angle $\alpha = 9°$, FOV = 159 mm, matrix = 106 × 106, slice thickness = 1.5 mm, 96 sagittal slices), a 2.2-min sequence (TR = 800 ms, TE = 3.43 ms, flip angle $\alpha = 9°$, FOV = 160 mm, matrix = 160 × 160, slice thickness = 1 mm, 144 sagittal slices), and a 6.5-min sequence (TR = 2530 ms, TE = 1.64 ms, flip angle $\alpha = 7°$, FOV = 256 mm, matrix = 256 × 256, slice thickness = 1 mm, 176 sagittal slices, acceleration factor = 2, 24 reference lines). In adults anatomicals were acquired using the 6.5-min sequence.

Data selection. Data were processed primarily using custom scripts, with tools from the FMRIB Software Library (FSL) version 4.1.8 and Freesurfer additionally used for registration and motion correction. Because some of our infant data contained a substantial amount of head movement, and because head motion causes highly deleterious artefacts in fMRI data[36], we first aimed to discard high-motion data that could corrupt our results and lead to false negatives. Each run was first motion corrected by registering each volume to the middle volume, using rigid transformations determined by FSL's MCFLIRT. Using the motion parameters estimated by this correction, we applied a technique known as scrubbing[37,38], removing pairs of adjacent volumes with >0.5 mm of total translation or 0.5° of total rotation between them. We also removed volumes where the participant's eyes were closed, and the first three volumes of each run (to allow the MR system to equilibrate).

While this technique is effective in removing artefactual spikes of response that occur at high-motion time points, it can still leave large baseline shifts in voxels' time courses that occur when a participant's head moves substantially and remains in a new location relative to the head coil and external magnetic field. We thus instituted a second cutoff on scrubbed data, at pairs of adjacent volumes with >2 mm of total translation or 2° of total rotation between them. At these cutoff points, we temporally split runs to form 'pseudoruns' of scrubbed data, where the head was in a relatively consistent position. These pseudoruns were subsequently analysed as one would normally analyse a full run. Pseudoruns were kept for analysis if they contained at least 24 time points, as well as six time points per condition for all conditions (where condition timing was lagged by 6 s to account for hemodynamic delay), such that responses to each condition could be estimated. Last, participants were included in analyses if they had at least 5 min of data saved after this procedure, across experiments.

Supplementary Table 1 shows the amount of data acquired and saved, across participants. We initially acquired 23.06 h of data across 17 infants, and were left with 4.26 h of data across 9 infants after motion screening. Resulting pseudoruns in infants ranged in length from 1.2–17.5 min (mean 4.3 min). While this procedure led to a substantial reduction in data quantity, it drastically reduced the amount of head motion present in the resulting data, reducing mean volume-to-volume translation from 1.11 to 0.13 mm, and mean rotation from 1.69° to 0.17°. In adults, neither scrubbing nor pseudorun selection resulted in any volumes being removed, such that pseudoruns were equivalent to the original runs. Adult data had mean volume-to-volume translation of 0.04 mm, and mean rotation of 0.02°.

Data preprocessing. Pseudoruns were first motion-corrected by registering each volume to the middle volume, using rigid transformations determined by FSL's MCFLIRT. Data were skull-stripped using FSL's Brain Extraction Tool, and spatially smoothed using a 3 mm-full-width at half-maximum Gaussian kernel.

Data registration. To combine data across pseudoruns, middle volumes from each pseudorun for a given participant were all registered to a common target middle volume, chosen to have minimal distortion. All registration was performed using FSL's FLIRT, unless otherwise noted. In infant data, head motion across pseudoruns posed challenges for this registration: different volumes could have different positions within the bounding box, and different types of nonrigid distortion. To optimize registration, we thus adopted the following procedure: (1) middle volumes were algorithmically registered to target volumes using both a rigid transformation and a general affine transformation; (2) translation and rotation parameters for both of these transformations were hand-tuned to improve registration quality; and (3) we selected whichever resulting transformation (hand-tuned rigid or hand-tuned affine) provided a more accurate registration based on visual inspection of anatomical landmarks. For adult data, middle volumes were registered to the target using a rigid transformation.

For infant data, in cases where anatomicals were collected, target functional volumes were registered to anatomical images using a rigid transformation, with translation and rotation parameters subsequently hand-tuned. For adult data, because surface reconstructions could be obtained, target functionals were registered to anatomicals with a rigid transformation determined by Freesurfer's bbregister. Anatomical images in adults were in turn registered to the Montreal Neurological Institute (MNI) 152 template brain using a nonlinear transformation determined by FSL's FNIRT.

Last, we aimed to register data across infants, for the purposes of registering search spaces for ROI analyses (described below), and to compute group-level whole-brain statistical maps. To this end, target functional volumes from each infant were registered to the target functional of infant 1, data set 3 (the infant and data set with the most useable data) using an affine transformation, with translation and rotation parameters subsequently hand-tuned. While these transformations were not perfect, insofar as linear registration cannot perfectly align different brains, they were primarily used for the registration of large search spaces, which should be tolerant to minor inaccuracies in registration. Lastly, to transform search spaces across infants and adults, this target functional volume was registered to the MNI brain using an affine transformation, with translation and rotation parameters subsequently hand-tuned.

Data modelling. For each pseudorun, whole-brain voxelwise linear models were performed to estimate the blood-oxygen-level-dependent response to visual stimuli. Regressors for each condition (excluding the baseline) were defined as boxcar functions with value 1 during blocks of that condition, convolved with a canonical double-gamma hemodynamic response function. Twelve nuisance regressors were additionally included to reduce the influence of potential artefacts. A linear trend regressor was included to account for signal drift. Motion parameter regressors (three translation parameters and three rotation parameters determined by motion correction) were used to minimize effects of head motion. Last, five principal component analysis (PCA)-based noise regressors were used to account for other noise sources (a method similar to GLMDenoise[39]). PCA-based regressors were defined by: (1) choosing a 'noise pool' of voxels with $<1\%$ of variance explained by the task regressors; (2) running PCA on time series from these voxels; and (3) choosing the top five principal components as regressors. For both task and nuisance regressors, time points that were scrubbed in data selection were removed after the regressors were defined (with the exception of PCA-based regressors, which were defined using scrubbed data).

This analysis provided beta values for task regressors corresponding to the magnitude of response to each condition, and contrast values corresponding to differences across conditions. To combine the resulting contrast values across pseudoruns for a given participant and data set, we computed a weighted average of contrast maps registered to a common functional space, using weights corresponding to the amount of data contributed by each pseudorun. Weights were proportional to $(c^T(X^TX)^{-1}c)^{-1}$, where c is the contrast vector and X is the design matrix for a given pseudorun. For a given contrast (for example, faces versus scenes), we combined data across all experiments containing that contrast.

We next statistically assessed these average contrast values for each participant. Because fMRI time series are temporally autocorrelated, within-participant statistics are typically computed using feasible generalized least squares, with an empirical estimate of the autocorrelation structure. However, the validity of extant methods for estimating the autocorrelation of fMRI data is not well established[40], and these methods have not been validated in infant data. To obviate the need for any assumptions about the autocorrelation structure in our data, we instead used a nonparametric permutation test[41]. Specifically, on each of 5,000 iterations, we randomly permuted the block order for each pseudorun, and computed a contrast value for each voxel. This procedure provided a null distribution that was used to threshold voxelwise contrast values at $P<0.01$, one-tailed. Estimated null distributions were fit with a Gaussian distribution, allowing us to estimate small P values that wouldn't be possible to estimate from the fraction of samples from the null distribution exceeding the observed statistic; for statistics with larger P values, the Gaussian fit gave very similar P values to those computed using the raw null distribution. For visualization and reporting purposes, voxelwise statistics were converted to z-values based on their computed P value. To correct for multiple comparisons across voxels, we additionally used a permutation test to build a null distribution for sizes of contiguous clusters of activation, and thresholded cluster sizes at $P<0.05$.

We additionally computed a group-level statistical map to perform inference across infants. Average contrast maps for each infant were registered to the target functional space of infant 1, data set 3, and voxelwise t-tests were performed across infants, comparing contrast values to zero, thresholded at $P<0.01$. For infants with multiple data sets acquired at different ages, we only used the data set with the largest amount of saved data. As above, voxelwise t-statistics were converted to z-values based on their computed P value for visualization purposes. To correct for multiple comparisons across voxels, a permutation test was used to build a null distribution for sizes of contiguous clusters of activation (where on each iteration, signs of contrast values for each infant were randomly flipped), and thresholded cluster sizes at $P<0.05$.

ROI analysis. To assess response profiles of brain regions identified in the whole-brain analysis, we performed ROI analyses. ROIs were defined as the set of voxels within a broad anatomical search space with the top N% of statistical values for a specific contrast, such as comparing faces to scenes or faces to objects. The value N was typically 5%, but was also varied from 2 to 30% to measure selectivity as a function of ROI size. Search spaces were hand-drawn on the anatomical image of one participant (infant 1, data set 3), and registered to other participants' functional images as described above (Data registration). They included (Supplementary Fig. 4): (1) lateral occipitotemporal cortex, covering the expected locations of the occipital face area and occipital place area (mean size 39.2 cm^3 in infants; 54.3 cm^3 in adults); (2) ventral temporal cortex, covering the expected locations of the fusiform face area and parahippocampal place area (38.4 cm^3 in infants; 54.0 cm^3 in adults); (3) STS, covering the expected location of the posterior STS face region (40.2 cm^3 in infants; 53.0 cm^3 in adults); and (4) medial prefrontal cortex (65.5 cm^3 in infants; 97.0 cm^3 in adults). To maximize the amount of data used to define regions, but still extract responses from data independent of those used to define the ROI[42], we used a leave-one-pseudorun-out analysis: ROIs were defined using data from all but one pseudorun, responses were extracted from the remaining pseudorun, and after iterating this process across all pseudoruns and participants, the resulting beta values and contrasts were combined using the weighted average described above (Data modelling). Beta values and contrasts were converted to per cent signal change values by dividing by mean signal strength within the ROI.

For most analyses, differences between conditions were statistically assessed using a permutation test, analogous to the procedure described above (Data modelling); these tests assess the significance of the observed effects within our sample. In addition, we tested whether the effects observed can be expected to generalize to the population. We compared responses to faces and scenes, because these conditions were observed by all infants, and combined data across all experiments to increase power within each participant. For each ROI (defined as described above, using the face versus scene contrast), mean per cent signal change values were computed for each participant, and the difference between responses to faces and scenes was statistically compared to zero using a one-tailed t-test across infants. As with the whole-brain group-level analysis, when infants yielded multiple data sets acquired at different ages, we only used the data set with the largest amount of usable data.

Visual feature analysis. We next asked whether responses in category-sensitive visual regions could be explained in terms of lower-level visual features. In particular, we focused on high- and low-frequency content and rectilinearity (the presence of 90° angles in an image), which have been argued previously to modulate responses in category-sensitive visual regions[24–26]. Frequency content and rectilinearity measures were computed on individual frames from each movie clip, and averaged across frames for a given clip. Frames were first converted to grayscale and normalized to have zero mean and unit standard deviation, to remove effects of overall luminance and contrast. We then computed the discrete Fourier transform of each frame, and defined low-frequency content as total power at frequencies less than one cycle per degree of visual angle, and high-frequency content as total power at frequencies greater than five cycles per degree of visual angle, following the cutoffs used by Razimehr *et al.*[25] Rectilinearity was computed using a procedure described by Nasr *et al.*[24]: frames were convolved with a bank of 90° angle Gabor filters at different scales and orientations, and magnitudes of convolved images were averaged across spatial position and filter to yield a single measure (Fig. 3).

We then assessed whether responses in category-sensitive ROIs were better predicted by category identity or by visual features. Regressors for visual features were defined by constructing time series of feature values for each individual movie in a given pseudorun, convolved with a canonical double-gamma hemodynamic response function. Categorical regressors were defined as described above (Data modelling). We compared five models: category (containing regressors for each visual category in an experiment), low-frequency content, high-frequency content, rectilinearity, and a model containing low-frequency, high-frequency and rectilinearity regressors. To eliminate the possibility that differences in model fit resulted from different degrees of freedom across models, model fit was assessed using leave-one-pseudorun-out cross-validation. For a given pseudorun, models were fit using data from all other pseudoruns with the same set of conditions from

that participant and data set (ROIs were also defined using data independent from the left-out pseudorun, as described in the ROI analysis section above). This provided a set of beta values that was used to define a predicted response for the left-out pseudorun, for each model. Model fit was assessed by computing the Fisher-transformed correlation (z'-value) between the time series in the left-out pseudorun and the predicted response. Linear trend and motion parameter nuisance regressors were included in all models. Model fit estimates were compared across models using paired, two-tailed t-tests across pseudoruns.

Representational similarity analysis. As an alternative method of comparing visual responses across infants and adults, we assessed the similarity structure of spatial patterns of response to different categories of stimuli[28]. Specifically, we computed correlations between spatial patterns of response (beta values comparing each condition to baseline) to the four conditions of Expts. 2–6, in infants and adults. Patterns were computed across voxels within extrastriate cortex, defined as the union of the three anatomical search spaces described above (ventral temporal cortex, lateral occipital cortex and the STS), with data combined across runs as described above (Data modelling). Correlation matrices (or representational similarity matrices, RSMs) were Fisher-transformed, averaged across participants and then inverse-Fisher-transformed for reporting.

To compare RSMs across groups, we next asked whether the ordering of correlation magnitudes across pairs or conditions (for example, face-object, face-body and so on) differed across infants and adults. We computed rank correlations (Kendall's tau) between correlation values from each pair of participants, either within infants, within adults, or between infants and adults, and asked whether orderings were more consistent (higher rank correlation) within group than between. To test whether the difference between within- and between-group rank correlations was significantly greater than zero, we performed an exact permutation test, building a null distribution for these values by computing them based on all possible group assignments of the six infants and three adults.

Further paradigm details. Across infants, eight slightly different experiments were run. Experiment 1 contained two categories (face and scene) and was run in every infant. Experiment 2 contained four categories (face, body, object and scene) and was run in a subset of $n = 4$ infants. Experiments 3–8 contained 3–4 categories and were each only run in a single infant. Experiments 3–7 used stimuli that are very similar to those used in Experiment 2, and were used in early scanning sessions before switching to Experiment 2. Experiment 8 contained distinct stimuli and was intended to provide additional evidence for generalization of category preferences across different specific videos. Because we did not acquire enough usable data with Experiments 3–8 to analyse them in isolation, they were ultimately only used in combination with other experiments, to increase power for various analyses. In particular, because all experiments contained face and scene categories, all were used for whole-brain face versus scene comparisons, and to define ROIs based on this contrast. Because Experiments 3–6 contained four categories, they additionally contributed to four-condition ROI responses.

Experiment 1 consisted of Filmed Faces and Baby Einstein Scenes conditions, as well as a baseline condition of spatially scrambled scenes (using 15×15 grid scrambling, as is the case for all scrambled stimuli). The Filmed Faces were 60 3 s-long close-up videos of children's faces on a black background, filmed by the experimenters, as used in a previous experiment in adults[43]. These videos did not contain parts of the body below the neck. The Baby Einstein Scenes were 36 3 s-long videos of scenes taken from the Baby Einstein video collection, which all depicted a three-dimensional (3D) spatial layout, and did not contain humans or animals.

Experiment 2 consisted of Filmed Front Faces, Filmed Objects*, Filmed Bodies, Filmed Scenes (presented at 80% size) and a baseline condition of spatially scrambled scenes (also presented at 80% size). The Filmed Front Faces were 30 3 s-long videos of front-view faces, similar to the Filmed Faces condition, but containing distinct specific videos. The Filmed Objects* were a set of 20 3 s-long close-up videos of children's toys on a black background (for example, rolling balls and moving gear toys), filmed by the experimenters. These 20 clips were selected from a larger set of 60 clips used in a previous experiment[43] (where the *denotes the subset), which were chosen to have virtually no information about 3D scene layout (for example, corners between walls or between a wall and a floor). The Filmed Bodies were a set of 60 3 s-long close-up videos of children's bodies or body parts (not showing faces) on a black background, as used in a previous experiment[43]. The Filmed Scenes were a set of 60 3 s-long videos filmed by the experimenters from a camera moving through an outdoor scene (for example, a road and a field), as used in a previous experiment[43]. These all depicted a 3D spatial layout, and did not contain humans or animals.

Experiment 3 consisted of Filmed Faces, Filmed Objects*, Filmed Bodies, Baby Einstein Scenes and a baseline condition of spatially scrambled scenes.

Experiment 4 consisted of Filmed Front Faces, Filmed Objects, Filmed Bodies, Filmed Scenes and a baseline condition of spatially scrambled objects. Filmed Objects were the full set of 60 filmed object videos from which the Filmed Objects* videos were selected.

Experiment 5 consisted of Filmed Faces, Filmed Objects, Filmed Bodies, Filmed Scenes and a baseline condition of spatially scrambled objects.

Experiment 6 consisted of Filmed Faces, Filmed Objects, Filmed Bodies, Baby Einstein Scenes and a baseline condition of spatially scrambled objects.

Experiment 7 consisted of Filmed Front Faces, Filmed Side Faces, Filmed Objects*, Baby Einstein Scenes (presented at 80% size) and a baseline condition of spatially scrambled scenes. The Filmed Side Faces were 35 3 s-long videos of side-view faces, similar to the Filmed Faces and Filmed Front Faces conditions but containing distinct specific videos.

Experiment 8 consisted of Baby Einstein Faces, Baby Einstein Objects, Animated Scenes and a baseline condition of spatially scrambled scenes. Baby Einstein Faces were three 18 s-long videos (containing multiple clips) of children's faces, taken from the Baby Einstein video collection. While these videos typically only contained faces, hands were occasionally presented in the vicinity of the face. Baby Einstein Objects were three 18 s-long videos (containing multiple clips) of children's toys and other objects in motion, taken from the Baby Einstein video collection. Animated Scenes were 18 6 s-long videos designed by having a camera move through an animated scene created using Blender 3D animation software. These all depicted a 3D spatial layout, and did not contain humans or animals.

Data availability. The stimuli, data and analysis code that support the findings of this study are available from the corresponding author on request.

References

1. Downing, P. E., Jiang, Y., Shuman, M. & Kanwisher, N. A cortical area selective for visual processing of the human body. *Science* **293,** 2470–2473 (2001).
2. Epstein, R. & Kanwisher, N. A cortical representation of the local visual environment. *Nature* **392,** 598–601 (1998).
3. Kanwisher, N. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc. Natl Acad. Sci.* **107,** 11163–11170 (2010).
4. Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17,** 4302–4311 (1997).
5. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76,** 1210–1224 (2012).
6. Konkle, T. & Caramazza, A. Tripartite organization of the ventral stream by animacy and object size. *J. Neurosci.* **33,** 10235–10242 (2013).
7. Kriegeskorte, N. *et al.* Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60,** 1126–1141 (2008).
8. Cantlon, J. F., Pinel, P., Dehaene, S. & Pelphrey, K. A. Cortical representations of symbols, objects, and faces are pruned back during early childhood. *Cereb. Cortex* **21,** bhq078 (2010).
9. Golarai, G., Liberman, A., Yoon, J. M. & Grill-Spector, K. Differential development of the ventral visual cortex extends through adolescence. *Front. Hum. Neurosci.* **3,** 80 (2009).
10. Deoni, S. C. *et al.* Mapping infant brain myelination with magnetic resonance imaging. *J. Neurosci.* **31,** 784–791 (2011).
11. Sowell, E. R. *et al.* Longitudinal mapping of cortical thickness and brain growth in normal children. *J. Neurosci.* **24,** 8223–8231 (2004).
12. Srihasam, K., Vincent, J. L. & Livingstone, M. S. Novel domain formation reveals proto-architecture in inferotemporal cortex. *Nat. Neurosci.* **17,** 1776–1783 (2014).
13. Srihasam, K., Mandeville, J. B., Morocz, I. A., Sullivan, K. J. & Livingstone, M. S. Behavioral and Anatomical Consequences of Early versus Late Symbol Training in Macaques. *Neuron* **73,** 608–619 (2012).
14. De Haan, M. & Nelson, C. A. Brain activity differentiates face and object processing in 6-month-old infants. *Dev. Psychol.* **35,** 1113 (1999).
15. De Haan, M., Pascalis, O. & Johnson, M. H. Specialization of neural mechanisms underlying face recognition in human infants. *J. Cogn. Neurosci.* **14,** 199–209 (2002).
16. Grossmann, T. *et al.* Early cortical specialization for face-to-face communication in human infants. *Proc. Roy. Soc. B* **275,** 2803–2811 (2008).
17. Lloyd-Fox, S. *et al.* Social perception in infancy -- a near infrared spectroscopy study. *Child Dev.* **80,** 986–999 (2009).
18. Rodman, H. R., Skelly, J. P. & Gross, C. G. Stimulus selectivity and state dependence of activity in inferior temporal cortex of infant monkeys. *Proc. Natl Acad. Sci. USA* **88,** 7572–7575 (1991).
19. McKone, E., Crookes, K., Jeffery, L. & Dilks, D. D. A critical review of the development of face recognition: experience is less important than previously believed. *Cogn. Neuropsychol.* **29,** 174–212 (2012).
20. Biagi, L., Crespi, S. A., Tosetti, M. & Morrone, M. C. BOLD response selective to flow-motion in very young infants. *PLoS Biol.* **13,** e1002260 (2015).
21. Dehaene-Lambertz, G., Dehaene, S. & Hertz-Pannier, L. Functional neuroimaging of speech perception in infants. *Science* **298,** 2013–2015 (2002).
22. Downing, P., Chan, A.-Y., Peelen, M., Dodds, C. & Kanwisher, N. Domain specificity in visual cortex. *Cereb. Cortex* **16,** 1453–1461 (2006).
23. Nasr, S. *et al.* Scene-selective cortical regions in human and nonhuman primates. *J. Neurosci.* **31,** 13771–13785 (2011).

24. Nasr, S., Echavarria, C. E. & Tootell, R. B. Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. *J. Neurosci.* **34,** 6721–6735 (2014).
25. Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C. & Tootell, R. B. The 'parahippocampal place area' responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biol* **9,** e1000608 (2011).
26. Yue, X., Pourladian, I. S., Tootell, R. B. & Ungerleider, L. G. Curvature-processing network in macaque visual cortex. *Proc. Natl Acad. Sci.* **111,** E3467–E3475 (2014).
27. Hasson, U., Levy, I., Behrmann, M., Hendler, T. & Malach, R. Eccentricity bias as an organizing principle for human high-order object areas. *Neuron* **34,** 479–490 (2002).
28. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2,** 4 (2008).
29. Striem-Amit, E., Cohen, L., Dehaene, S. & Amedi, A. Reading with sounds: sensory substitution selectively activates the visual word form area in the blind. *Neuron* **76,** 640–652 (2012).
30. Striem-Amit, E. & Amedi, A. Visual cortex extrastriate body-selective area activation in congenitally blind people 'seeing' by using sounds. *Curr. Biol.* **24,** 687–692 (2014).
31. Reid, R. C. From functional architecture to functional connectomics. *Neuron* **75,** 209–217 (2012).
32. Baker, C. I. *et al.* Visual word processing and experiential origins of functional selectivity in human extrastriate cortex. *Proc. Natl Acad. Sci.* **104,** 9087–9092 (2007).
33. Saygin, Z. M. *et al.* Connectivity precedes function in the development of the visual word form area. *Nat. Neurosci.* **19,** 1250–1255 (2016).
34. Keil, B. *et al.* Size-optimized 32-channel brain arrays for 3 T pediatric imaging. *Magn. Reson. Med.* **66,** 1777–1787 (2011).
35. Zapp, J., Schmitter, S. & Schad, L. R. Sinusoidal echo-planar imaging with parallel acquisition technique for reduced acoustic noise in auditory fMRI. *J. Magn. Reson. Imaging* **36,** 581–588 (2012).
36. Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. & Turner, R. Movement-related effects in fMRI time-series. *Magn. Reson. Med.* **35,** 346–355 (1996).
37. Siegel, J. S. *et al.* Statistical improvements in functional magnetic resonance imaging analyses produced by censoring high-motion data points. *Hum. Brain Mapp.* **35,** 1981–1996 (2014).
38. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59,** 2142–2154 (2012).
39. Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F. & Wandell, B. A. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* **7,** 247 (2013).
40. Eklund, A., Andersson, M., Josephson, C., Johannesson, M. & Knutsson, H. Does parametric fMRI analysis with SPM yield valid results? An empirical study of 1484 rest datasets. *Neuroimage* **61,** 565–578 (2012).
41. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15,** 1–25 (2002).
42. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.* **4,** 274–290 (2009).
43. Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C. & Kanwisher, N. Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage* **56,** 2356–2363 (2011).

Acknowledgements

We thank the Packard Foundation, Ellison Medical Foundation and NSF (graduate research fellowship to B.D., and the Center for Brains, Minds and Machines, CCF-1231216 to N.K. and R.S.) for funding this research; Anna Wexler for assistance in stimulus creation; Grace Lisandrelli for assistance with recruitment and data collection; Jorie Koster-Hale, Bob Desimone, Charles Jennings and Winrich Freiwald for useful feedback on the manuscript; and all of our infants and parents for participating.

Author contributions

B.D., N.K. and R.S. designed research; B.D., H.R., D.D.D. and R.S. collected data; B.K. and L.L.W. provided the infant head coil; A.T. provided technical assistance with data acquisition; B.D. and R.S. Analysed data; B.D., N.K. and R.S. wrote paper.

Additional information

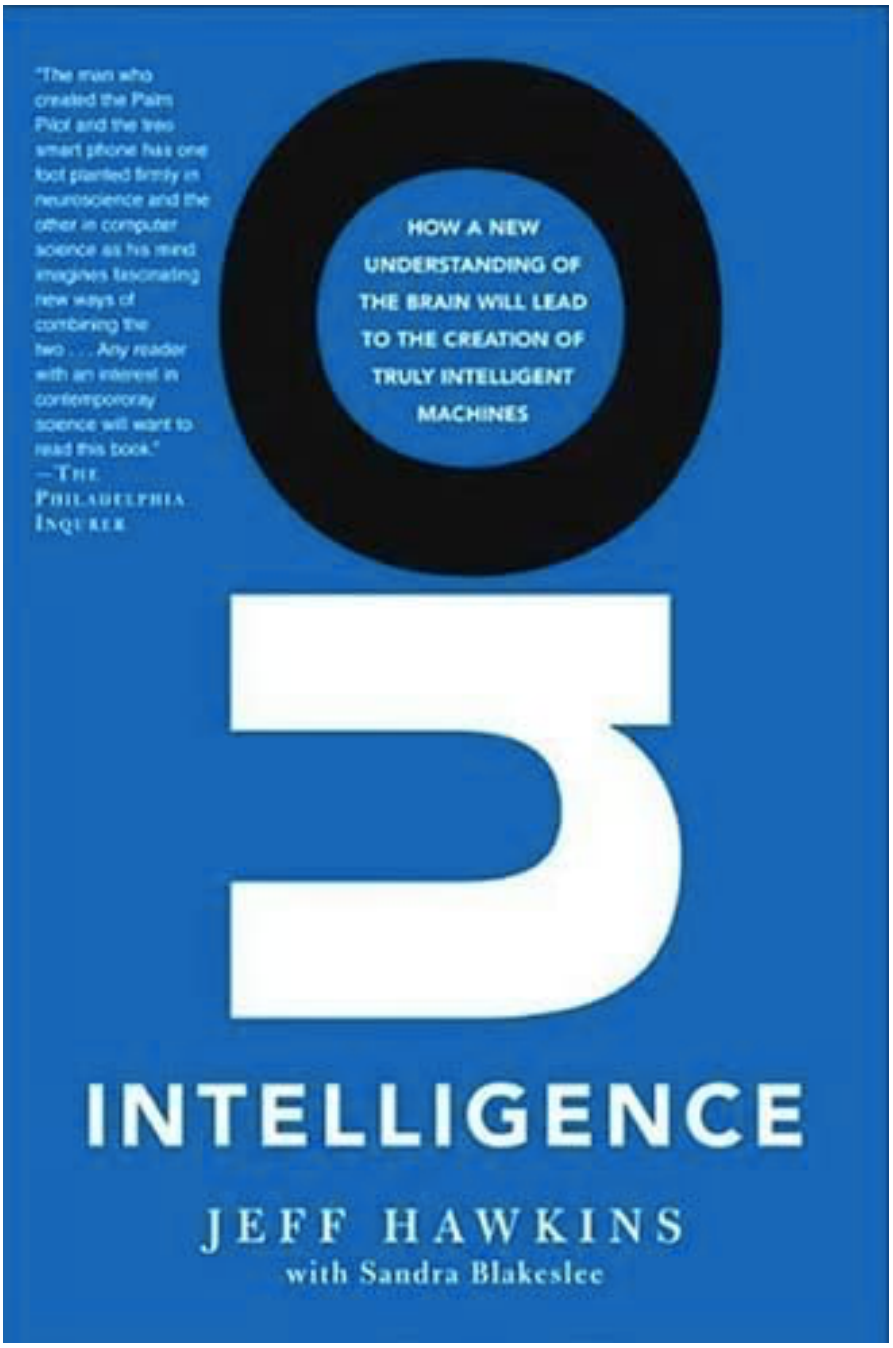
Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Deen, B. *et al.* Organization of high-level visual cortex in human infants. *Nat. Commun.* **8,** 13995 doi: 10.1038/ncomms13995 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.


"The man who created the Palm Pilot and the Treo smart phone has one foot planted firmly in neuroscience and the other in computer science as his mind imagines fascinating new ways of combining the two . . . Any reader with an interest in contemporary science will want to read this book."
—The Philadelphia Inquirer
HOW A NEW UNDERSTANDING OF THE BRAIN WILL LEAD TO THE CREATION OF TRULY INTELLIGENT MACHINES
INTELLIGENCE
JEFF HAWKINS
with Sandra Blakeslee