CORRECTED JULY 1, 2009; SEE LAST PAGE

Emotion
2009, Vol. 9, No. 2, 260–264

1528-3542/09/$12.00 DOI: 10.1037/a0014681

BRIEF REPORTS

Structural Resemblance to Emotional Expressions Predicts Evaluation of Emotionally Neutral Faces

Christopher P. Said
Princeton University

Nicu Sebe
University of Amsterdam, University of Trento, Italy

Alexander Todorov
Princeton University

People make trait inferences based on facial appearance despite little evidence that these inferences accurately reflect personality. The authors tested the hypothesis that these inferences are driven in part by structural resemblance to emotional expressions. The authors first had participants judge emotionally neutral faces on a set of trait dimensions. The authors then submitted the face images to a Bayesian network classifier trained to detect emotional expressions. By using a classifier, the authors can show that neutral faces perceived to possess various personality traits contain objective resemblance to emotional expression. In general, neutral faces that are perceived to have positive valence resemble happiness, faces that are perceived to have negative valence resemble disgust and fear, and faces that are perceived to be threatening resemble anger. These results support the idea that trait inferences are in part the result of an overgeneralization of emotion recognition systems. Under this hypothesis, emotion recognition systems, which typically extract accurate information about a person's emotional state, are engaged during the perception of neutral faces that bear subtle resemblance to emotional expressions. These emotions could then be misattributed as traits.

Keywords: social cognition, face perception, trait judgments, emotional expressions, computer vision

People evaluate neutral faces on multiple trait dimensions and these evaluations have social consequences (Hassin & Trope, 2000). For instance, political candidates whose faces are perceived as more competent are more likely to win elections (Ballew & Todorov, 2007; Todorov, Mandisodza, Goren, & Hall, 2005), and cadets whose faces are perceived as more dominant are more likely to be promoted to higher military ranks (Mazur, Mazur, & Keating, 1984).

Although inferences about traits based on facial appearance are made reliably across observers, there is little evidence that these inferences accurately reflect the personality of the observed face. Most correlations between perceived traits and actual traits are weak though positive (Bond, Berry, & Omar, 1994), some are inconsistent for men and women (Zebrowitz, Voinescu, & Collins, 1996), and some are negative (Zebrowitz, Andreoletti, Collins, Lee, & Blumenthal, 1998). It is therefore puzzling that people make reliable and rapid trait inferences from faces (Willis & Todorov, 2006) when only little accurate information, at best, is provided about personality. One intriguing explanation is that neutral faces may contain structural properties that cause them to resemble faces with more accurate and ecologically relevant information (Zebrowitz, 2004) such as emotional expressions (Knutson, 1996; Montepare & Dobish, 2003).

Under this hypothesis, the adaptive ability to recognize emotions overgeneralizes to neutral faces that merely bear a subtle resemblance to emotions. For example, although people can categorize faces as emotionally neutral, they can also agree that these faces vary on trait dimensions such as trustworthiness (Engell, Haxby, & Todorov, 2007). One possibility is that the source of consensus in judging faces on social dimensions is the similarity of the face to expressions corresponding to the dimension of trait judgment (e.g., aggressiveness and anger). When given the task of making a trait judgment from an emotionally neutral face, people could base their judgments on this similarity. Evidence for this hypothesis comes from research showing that the more a neutral face is rated as happy by one group of participants the higher it is rated on dominance and affiliation by another group of participants, and the more a face is rated as angry the higher it is rated on dominance and the lower on affiliation (Montepare & Dobish, 2003). One interpretation of these findings is that people misattribute similarity to emotional expressions, possibly detected by emotion recognition systems, to personal dispositions.

Christopher P. Said and Alexander Todorov, Department of Psychology and the Center for the Study of Brain, Mind and Behavior at Princeton University, Princeton; Nicu Sebe, Faculty of Science, University of Amsterdam, Amsterdam and University of Trento.

We thank Valerie Loehr for her assistance with the acquisition of trait ratings, and Nick Oosterhof for helpful discussions. This research was supported by National Science Foundation Grant BCS-0446846.

Correspondence should be addressed to Christopher P. Said, Department of Psychology, Princeton University, Princeton, NJ 08540. E-mail: csaid@princeton.edu

However, an alternative explanation is that these correlations are driven by semantic similarities between traits and emotions and not by visual similarities (Bruner & Tagiuri, 1954; Cronbach, 1955; Schneider, 1973). It is possible, for instance, that an affiliative face is rated as appearing happy not because of visual similarity, but because a semantic relation between affiliation and happiness triggers top-down control of perception. We avoided this potential confound by using a Bayesian network classifier trained only on facial emotions to detect the subtle presence of emotions in a set of 66 neutral faces. The network accepts as input a feature vector containing the displacements between automatically chosen landmarks and the same landmarks of a prototypical neutral face (Figure 1a). The output of the network is a set of probabilities corresponding to each basic emotion: happiness, surprise, anger, disgust, fear, and sadness. As expected, the output probabilities were very low, because the network was applied to neutral faces. Nevertheless, these probabilities could be analyzed and compared to trait ratings provided by human participants. To the extent that trait inferences are based on perceptual similarity to emotional expressions, the network classification probabilities should predict trait judgments from emotionally neutral faces.

Method

Participants

There were 301 undergraduate students took part in the study and were compensated with course credit.

Face Stimuli

We used a set of 66 standardized faces (Lundqvist, Flykt, & Ohman, 1998) with direct gaze and neutral expressions. Each face stimulus showed an amateur actor or actress between 20 and 30 years of age, with no facial hair, earrings, eyeglasses, or visible make-up, all wearing gray T-shirts. All of these faces were categorized as emotionally neutral in an expression categorization task in a previous study (Engell et al., 2007). An equal number of male and female faces were used.

Acquisition of Trait Ratings From Human Participants

Participants were asked to rate the set of 66 faces on 14 trait dimensions (see Table 1). Twelve of the trait dimensions were

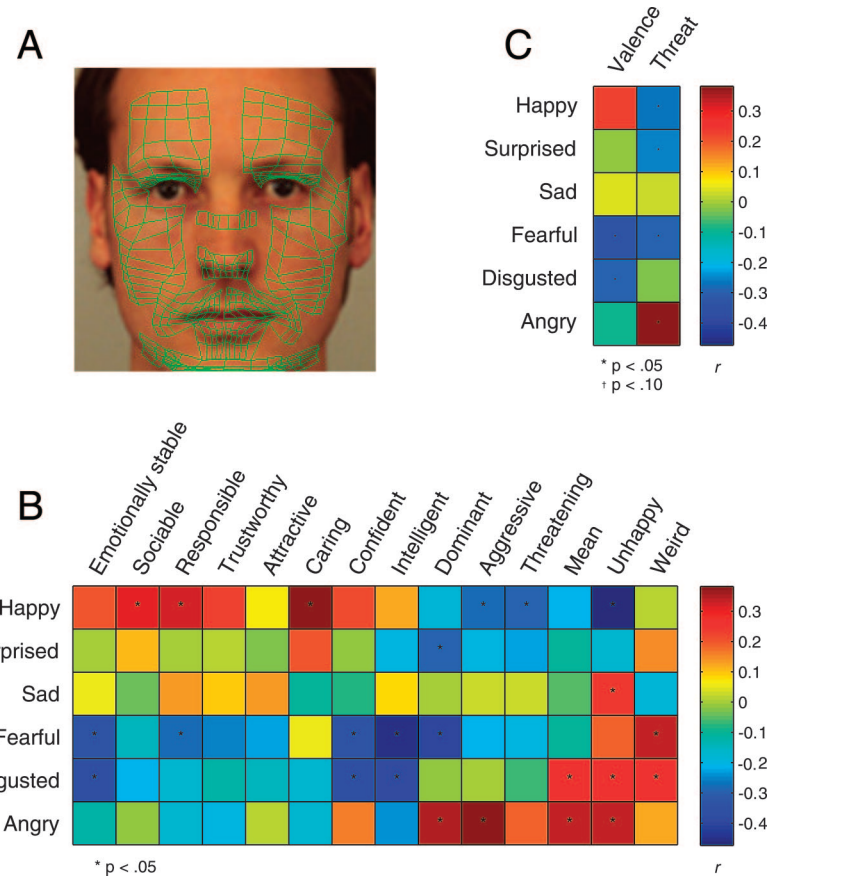

Figure 1. Relationships between human trait judgments and classifier emotion probabilities. (A) Example of neutral face with automatically placed landmarks. (B) Correlations between classifier emotion probabilities and human trait judgments. Traits are ordered by their loadings on the first component (valence; see Table 1). (C) Correlations between emotion probabilities and the first two components (valence and threat) derived from a principal components analysis of trait judgments with a subsequent oblimin rotation. Reprinted with permission from Arne Öhman, PhD, owner of the pictured figure.

Table 1
Loadings of Trait Judgments of Emotionally Neutral Faces on the First Two Components of a Principal Components Analysis With a Subsequent Oblimin Rotation

Trait judgment	Valence evaluation	Threat evaluation
Emotionally stable	.95	−.34
Sociable	.93	−.35
Responsible	.91	−.45
Trustworthy	.90	−.56
Attractive	.86	−.20
Caring	.80	−.74
Confident	.83	.16
Intelligent	.73	−.28
Dominant	−.03	.92
Aggressive	−.53	.94
Threatening	−.57	.89
Mean	−.62	.84
Unhappy	−.68	.44
Weird	−.91	.28

Note. The loadings represent the correlations of the trait judgments with the components. The third component had an eigenvalue smaller than 1.

chosen as a summary of unconstrained person descriptions of the faces by a separate group of 55 participants (Oosterhof & Todorov, 2008). Two other traits—dominance and threat—were added because of their importance for person perception (Bar et al., 2006; Wiggins et al., 1989). A new group of participants was then asked to rate the faces on a scale of 1 (not at all [trait name]) to 9 (extremely [trait name]) using their "gut instinct." Each face was presented at the center of the screen until the participant responded. The interstimulus interval was 1,000 ms. Each trait was rated by at least 18 participants, and each participant rated only one trait. To increase reliability, the faces were presented three times in three separate blocks. Within each block, their order was randomized. The reliability of the trait judgments was high ($\alpha > .90$ for each trait judgment). The procedures are described further in Oosterhof and Todorov (2008).

Emotion Classifier Training

A computer vision system was built to detect emotional expressions in faces by comparing the position of a set of automatically chosen landmarks to those of a prototypical neutral face. To create the prototypical neutral face, we used a database consisting of 1,000 frontal neutral faces with given annotation of facial features. The faces were normalized and registered using the Active Appearance Model (AAM) application programming interface (http://www2.imm.dtu.dk/~aam/).

For classification we used a Bayesian Network that classifies the feature vector x containing the displacements between automatically chosen landmarks and the same landmarks of the prototypical neutral face. The network is composed of a directed acyclic graph in which every node is associated with a variable X_i and with a conditional distribution $p(X_i|\Pi_i)$, where Π_i denotes the parents of X_i in the graph. The joint probability distribution is factored to the collection of conditional probability distributions of each node in the graph as: $p(X_i,\ldots,X_n) = \Pi_{i=1}^{n} p(X_i|\Pi_i)$. The directed acyclic graph is the structure, and the distributions $p(X_i|\Pi_i)$ represent the parameters of the network. The classifier receives x and generates a label $\hat{c}(x) \in C$, where in C is the set containing the seven basic emotions classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. An optimal classification rule can be obtained from the exact distribution $p(C,X)$ that represents the a posteriori probability of the class given the features. To learn the structure of the network from the training data we searched for the structure that minimized the probability of classification error directly using the stochastic search algorithm (Cohen, Cozman, Sebe, Cirelo, & Huang, 2004). To train the classifiers, we used a predetermined subset of the Cohn-Kanade database (Kanade, Tian, & Cohn, 2000) consisting of expression sequences of 53 FACS coded faces, starting from a neutral expression and ending in the peak of the facial expression. In a validation test of the remaining 61 faces the network performed well (mean hit rate = .885, SE = .013). Finally, the classifier was applied to the 66 neutral faces from which trait ratings by participants had been obtained. As noted above, these faces came from a database separate from the one used for training and validation.

Three faces were excluded from all subsequent analysis for having a probability of neutral less than 0.5. Because the distribution of emotion probabilities was positively skewed, we submitted the data to a cube root transform. Faces with a Cook's Distance of greater than the standard cutoff of 1.0 were excluded from any correlation described below. This procedure never excluded more than one face per correlation.

Results

Because the classifier was tested on neutral faces, the classifier probabilities for emotional expressions were all low, and the average classifier probability for the neutral category was high (M = .91, SD = .08). Nevertheless, 27 out of the 84 correlations of emotion probabilities and trait judgments were significant (Figure 1b). The probability of classifying faces as happy was positively correlated with all positive trait judgments and negatively correlated with all negative judgments (ps $< .05$ for caring, responsible, sociable, aggressive, threatening, and unhappy). The probability of classifying faces as surprised was negatively correlated with judgments of dominance ($p < .05$). The probability of classifying faces as angry was positively correlated with judgments of aggressiveness, meanness, unhappiness, and dominance (ps $< .05$). The probability of classifying faces as disgusted was positively correlated with judgments of meanness, unhappiness, and weirdness, and negatively correlated with judgments of intelligence, confidence, and emotional stability (ps $< .05$). The probability of classifying faces as sad was positively correlated with judgments of unhappiness ($p < .05$). The probability of classifying faces as fearful was positively correlated with judgments of weirdness and negatively correlated with judgments of intelligence, confidence, dominance, emotional stability, and responsibility (ps $<$ .05).

Because the trait judgments were highly correlated with each other, we submitted them to a principal components analysis (PCA)[1] with a subsequent oblimin rotation, which allows for the

[1] The first two (unrotated) principal components accounted for 82.2% of the variance. Their interpretation is described in Oosterhof and Todorov (2008). The third component had an eigenvalue smaller than 1 and accounted for less than 6% of the variance.

components to be correlated. In this research context, it is preferable to allow oblique rotation because there is no a priori reason to expect that latent variables underlying face evaluation are orthogonal. As shown in Table 1, all positive judgments had positive loadings and all negative judgments had negative loadings on the first component. Judgments of aggressiveness, dominance, threat, and meanness had the highest loadings on the second component. This pattern of correlations suggests that the components can be interpreted as valence and threat, respectively.[2] The two components were negatively correlated ($r = -.32$).

As shown in Figure 1c, the valence component (see Table 1) was positively correlated with the classifier probabilities of happiness, although the correlation was marginally significant ($p = .074$), and negatively correlated with the probabilities of disgust ($p < .02$) and fear ($p < .007$). The threat component was positively correlated with the classifier probabilities of anger ($p < .003$) and negatively correlated with the probabilities of happiness ($p < .038$) and surprise ($p < .045$).

Discussion

In a test of the hypothesis that trait inferences from emotionally neutral faces are based on perceptual similarity of the faces to emotional expressions, we compared multiple trait ratings of faces to emotion probabilities produced by a computer vision system trained to identify the basic emotional expressions. Consistent with this hypothesis, we found significant correlations between multiple trait ratings and resemblance of the faces to emotional expressions. Because the trait ratings were highly correlated with each other, we used PCA with a subsequent oblimin rotation to find a simple two-dimensional solution that underlies these ratings. This analysis identified two dimensions—valence and threat. In general, both dimensions spanned a range from positive to negative traits. However, the dimensions also captured different aspects of face evaluation. The threat dimension correlated with resemblance to expressions signaling avoidance (anger) behavior to the perceiver (Adams, Ambady, Macrae, & Kleck, 2006; Adams & Kleck, 2005; Marsh, Ambady, & Kleck, 2005). The valence dimension correlated negatively with resemblance to expressions that are not directly threatening but are not typically expected in neutral social interactions (disgust and fear).

These results suggest that trait judgments of neutral faces are driven in part by structural similarity to emotional expressions. In most cases, we would describe the process as an overgeneralization of emotion recognition systems in the brain. For instance, neural systems responsible for the detection of anger may be partially activated by a neutral face that resembles anger, thereby sending weak excitatory signals to representations of aggression, dominance, and all the other states that are typically associated with anger (Hess, Blairy, & Kleck, 2000; Knutson, 1996). Because the face is explicitly perceived as neutral, misattribution may occur, and these states may be interpreted by observers as indicating more permanent traits.

We believe that most of the correlations between trait ratings and resemblance of faces to expressions reflect a causal relationship between emotions and traits, where subtle resemblance to expressions causes trait inference. However, as with all correlational studies, we cannot rule out the possibility that the direction of causation is reversed. It is possible, in principle, that the perception of emotional expressions could be the results of overgeneralization of trait detection systems. We do not consider this to be plausible because it is far more likely for a system to detect veridical information (expressions) and then overgeneralize to less veridical information (traits), than for a system to detect nonveridical information and then generalize accurately to veridical information.

Although we argue that the main mechanism that can explain correlations between trait and emotion judgments is overgeneralization, not all such correlations are best described as overgeneralizations. Although trait ratings of intelligence were negatively correlated with classification probabilities of disgust and fear, it is difficult to think of any meaningful relationships between these concepts. We do not think that there is a direct causal relationship between the perception of fear and disgust and the perception of intelligence. Rather, there is evidence that the perception of intelligence may be caused in part by the overgeneralization of systems used to detect mental illness (Zebrowitz, Fellous, Mignault, & Andreoletti, 2003; Zebrowitz & Rhodes, 2004). We suspect that the facial properties associated with many forms of mental illness may have a coincidental relationship with facial properties associated with disgust and fear.

In this study, all faces were of amateur actors instructed to pose neutral faces. Even though the faces were all later rated as neutral in a forced choice task (Engell et al., 2007), it is possible that noncompliant actors may have engaged in muscle contractions to create microexpressions that influenced the classifier's performance. We cannot know with certainty whether the subtle resemblance to emotions detected in faces were due entirely to permanent structural features or whether they were also due in part to muscle contractions of noncompliant actors. It is possible that some people engage in persistent expressive habits requiring tonic muscle contraction at all times. A potential line of research is to study the extent to which trait inferences depend on tonic muscle contraction as opposed to the structure of other facial tissues. In any case, an important implication of the current work is that there are no really emotionally "neutral" faces. Faces are imbued with affect and one source of this affect is similarity to emotional expressions.

Because this study addresses the relationship between resemblance to emotional expressions and the traits of face identities, the results bear some relevance to research on the degree of overlap between neural systems responsible for the processing of facial expressions and those responsible for the processing of facial identity (Calder & Young, 2005; Haxby, Hoffman, & Gobbini, 2000). To the extent that one considers trait judgments a matter of facial identity, our results support the idea that emotion processing systems influence the processing of facial identity. Whether this influence is the result of shared bottom-up streams or the result of

[2] The first rotated component was highly correlated with the first unrotated component ($r = .94$), which was interpreted as valence by Oosterhof and Todorov (2008). The second rotated component was also correlated with the second unrotated component, which was interpreted as dominance by Oosterhof and Todorov (2008), although the correlation was not as high ($r = .79$). Forcing the components to be orthogonal seemed to remove the valence content of the second component.

high-level interaction between mostly separate streams cannot be determined by the present study.

Conclusions

In this study, we used an objective emotion classifier to demonstrate the relationship between a set of trait inferences and subtle resemblance to emotions in neutral faces. In general, positive traits are correlated with structural resemblance to happiness, whereas traits involving dominance and threat are correlated with structural resemblance to anger. Under the overgeneralization hypothesis, this is consistent with research on nonneutral faces showing an association between happiness and positive valence (Ekman, Friesen, & Ellsworth, 1972) and anger and dominance (Hess, Adams, & Kleck, 2005). Other sources of overgeneralization probably exist, including resemblance to other individuals with known traits (Zebrowitz & Montepare, 2008) and resemblance to babies (Zebrowitz et al., 2003). This last study used connectionist network modeling to test the overgeneralization effect for baby-faced individuals. Future research should continue to investigate the relative contributions of these and other sources using similarly objective measures.

References

Adams, R. B., Ambady, N., Macrae, C. N., & Kleck, R. E. (2006). Emotional expressions forecast approach-avoidance behavior. *Motivation and Emotion, 30,* 179–188.

Adams, R. B., & Kleck, R. E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion, 5,* 3–11.

Ballew, C. C., 2nd, & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 17948–17953.

Bar, M., Neta, M., & Linz, H. (2006). Very first impressions. *Emotion, 6(2),* 269–278.

Bond, C., Berry, D., & Omar, A. (1994). The kernel of truth in judgments of deceptiveness. *Basic and Applied Social Psychology, 15,* 523–534.

Bruner, J. S., & Tagiuri, R. (1954). The perception of people. In G. Lindzey (Ed.), *Handbook of social psychology* (Vol. 2). Cambridge, MA: Addison Wesley.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews. Neuroscience, 6,* 641–651.

Cohen, I., Cozman, F. G., Sebe, N., Cirelo, M. C., & Huang, T. S. (2004). Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(12).

Cronbach, L. J. (1955). Processes affecting scores on understanding of others and assumed similarity. *Psychological Bulletin, 52,* 177–193.

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

Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience, 19,* 1508–1519.

Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality Social Psychology, 78,* 837–852.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4,* 223–233.

Hess, U., Adams, R. B., & Kleck, R. E. (2005). Who may frown and who should smile? Dominance, affiliation, and the display of happiness and anger. *Cognition and Emotion, 19,* 515–536.

Hess, U., Blairy, S., & Kleck, R. E. (2000). The influence of facial emotion displays, gender, and ethnicity on judgments of dominance and affiliation. *Journal of Nonverbal Behavior, 24,* 265–283.

Kanade, T., Tian, Y., & Cohn, J. F. (2000). *Comprehensive database for facial expression analysis.* Paper presented at the Fourth IEEE international conference on automatic face and gesture recognition.

Knutson, B. (1996). Facial expressions of emotion influence interpersonal trait inferences. *Journal of Nonverbal Behavior, 20,* 165–182.

Lundqvist, D., Flykt, A., & Ohman, A. (1998). The Karolinska directed emotional faces (Publication from Psychology Section, Department of Clinical Neuroscience, Karolinska Hospital, S-171 76 Stockholm.

Marsh, A. A., Ambady, N., & Kleck, R. E. (2005). The effects of fear and anger facial expressions on approach- and avoidance-related behaviors. *Emotion, 5,* 119–124.

Mazur, A., Mazur, J., & Keating, C. (1984). Military rank attainment of a West Point Class: Effects of cadets physical features. *American Journal of Sociology, 90,* 125–150.

Montepare, J. M., & Dobish, H. (2003). The contribution of emotion perceptions and their overgeneralizations to trait impressions. *Journal of Nonverbal Behavior, 27,* 237–254.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, USA, 105,* 11087–11092.

Schneider, D. J. (1973). Implicit personality theory: Review. *Psychological Bulletin, 79,* 294–309.

Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science, 308,* 1623–1626.

Wiggins, J. S., Philips, N., Trapnell, P. (1989). Circular reasoning about interpersonal behavior: Evidence concerning some untested assumptions underlying diagnostic classification. *Journal of Personality and Social Psychology, 56,* 296–305.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychology Science, 17,* 592–598.

Zebrowitz, L. A. (2004). The origins of first impressions. *Journal of Cultural and Evolutionary Psychology, 2,* 93–108.

Zebrowitz, L. A., Andreoletti, C., Collins, M. A., Lee, S. Y., & Blumenthal, J. (1998). Bright, bad, babyfaced boys: Appearance stereotypes do not always yield self-fulfilling prophecy effects. *Journal of Personality and Social Psychology, 75,* 1300–1320.

Zebrowitz, L. A., Fellous, J. M., Mignault, A., & Andreoletti, C. (2003). Trait impressions as overgeneralized responses to adaptively significant facial qualities: Evidence from connectionist modeling. *Personality and Social Psychology Review, 7,* 194–215.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Compass, 2,* 1497–1517.

Zebrowitz, L. A., & Rhodes, G. (2004). Sensitivity to "bad genes" and the anomalous face overgeneralization effect: Accuracy, cue validity, and cue utilization in judging intelligence and health. *Journal of Nonverbal Behavior, 28,* 167–185.

Zebrowitz, L. A., Voinescu, L., & Collins, M. A. (1996). "Wide-eyed" and "crooked-faced": Determinants of perceived and real honesty across the life span. *Personality and Social Psychology Bulletin, 22,* 1258–1269.

Received June 16, 2008
Revision received November 4, 2008
Accepted November 10, 2008 ■

Correction to Said, Sebe and Todorov (2009)

In the article, "Structural Resemblance To Emotional Expressions Predicts Evaluation of Emotionally Neutral Faces" by Christopher Said, Nicu Sebe, and Alexander Todorov (*Emotion*, 2009, Vol. 9, No. 2, pp. 260-264) a symbol was incorrectly omitted from Figure 1, part C. To see the complete article with the corrected figure, please go to http://dx.doi.org/10.1037/a0014681