# Accuracy of inferring self- and other-preferences from spontaneous facial expressions

Michael S. North

Alex Todorov

Daniel N. Osherson

Word count (including Abstract and Footnote): 2499

References: 25

**Abstract**

Participants' faces were covertly recorded while they rated the attractiveness of people, the decorative appeal of paintings, and the cuteness of animals. Ratings employed a continuous scale. The same participants then returned and tried to guess ratings from 3-second videotapes of themselves and other targets. Performance was above chance in all three stimulus categories, thereby replicating the results of an earlier study (North, Todorov, & Osherson, 2010) but this time using a more sensitive rating procedure. Across conditions, accuracy in reading one's own face was not reliably better than other-accuracy. We discuss our findings in the context of "simulation" theories of face-based emotion recognition (Goldman, 2006) and the larger body of accuracy research.

*Keywords:* face perception, facial expressions, accuracy, self-accuracy, social cognition

How accurately can people read a casual emotion from a face whose owner does not suspect that s/he is under observation? Few quantitative studies have addressed this question. Although many valuable experiments have focused on the recognition of emotion, aside from North, Todorov, & Osherson (2010) none has involved low-level emotions decoded from dynamic, unwitting faces that are briefly encountered.

Research evaluating people's ability to infer emotions from the face typically relies on posed, still images. For example, studies incorporating Ekman's battery of basic emotions find reliable emotion categorization from photos of posing actors (e.g. Ekman, 1989; Hess, Blairy, & Kleck, 1997). This result emerges even when the stimuli depict the eye region alone (Baron-Cohen, Wheelwright, & Jolliffe, 1997). Such experiments reveal people's ability to infer enacted, exaggerated, basic emotions, but not the more mundane, dynamic facial information encountered on a daily basis.

Other studies have captured dynamic behavior using video recording. However, to our knowledge only Buck's (1979) "slide-viewing paradigm" focuses primarily on the face as a dynamic, *spontaneous* (i.e. naturally-occurring) source of information. In Buck's procedure, targets' faces are secretly filmed as they view and discuss a set of evocative images; perceivers then view this footage (with sound muted) and guess the relevant category and overall pleasantness of each image. The emotionality of the faces, however, is extreme due to the provocative character of the stimuli (e.g., sexualized images and mutilated bodies). Zaki, Bolger, & Ochsner (2008, 2009) similarly induced strong emotions by requesting that targets discuss some of the "most positive" and "most negative" events in their lives; observers then successfully detected the valence of the induced emotions. Ickes' influential (1997) accuracy paradigm fosters a less emotional,

unrehearsed dyadic interaction by filming participants as they chat in an alleged waiting room (gauging accuracy in guessing the other person's thoughts). Nevertheless, this also offers subsequent observers a wealth of information (e.g. auditory and gestural) besides nonverbal, facial behavior.

Other useful studies gauge accuracy from subtle, dynamic facial behavior, but not spontaneously. Such paradigms tend to utilize dynamic enactments of specific emotions (Ambadar, School, & Cohn, 2005; Gosselin, Kirouac, & Doré, 1995; Hess, Kappas, McHugo, Kleck, & Lanzetta, 1989) or else generate synthetic facial expressions via computer (Krumhuber & Kappas, 2005; Wehrle, Kaiser, Schmidt, & Scherer, 2000).

To isolate low-emotional, nonverbal, spontaneous facial cues, we utilized a procedure that clandestinely films participants while they view and rate a set of images. Judgments comprise various criteria, such as the decorative appeal of paintings. Viewing just the resulting facial behavior, a separate set of perceivers was above chance in guessing targets' image preferences (North et al., 2010). One limitation of this study is the coarse nature of the judgments asked of perceivers; a mere binary decision between images was required rather than a judgment of preference magnitude. The present study utilizes a continuous measure to better assess accuracy in reading relatively calm, unsuspecting faces.

Additionally, this previous study tested whether people accurately read other people's faces only, not assessing accuracy in reading one's own non-social face. Earlier experiments have measured self-accuracy from filmed interactions, finding above-chance matching between self-reports and objective coding of specific behaviors (e.g. laughing, interrupting, and hand gestures; Gosling, John, Craik, & Robins, 1998; Hall, Murphy, &

Schmid Mast, 2007). Other studies have provided evidence of accurate self-attribution of expressiveness after emotional role-playing (Hess, Sénecal, & Thibeault; 2004) and after viewing particularly arousing stimuli (Barr & Kleck, 1995). These experiments have yielded valuable information about self-perception, but each involved either exaggerated emotional situations or information beyond the face.

In contrast, the present experiment explores people's accuracy in reading their own and others' subtle expressions in a non-social context. This investigation helps evaluate the "motor theory" of embodied face perception, according to which people perceive facial expressions by implicitly imitating them (see, e.g., Niedenthal, Barsalou, Winkielman, Krauth-Gruber, & Ric, 2005; Niedenthal, 2007). From this perspective, self-accuracy might outpace other-accuracy, given people's familiarity with controlling their own facial movements. On the other hand, lack of experience with viewing one's face from the outside might inhibit a self-perception advantage.

## Method

**Overview.** The experiment comprised two phases. In the first ("target") phase, participants were clandestinely filmed while viewing and rating images in isolation. In the second ("perceiver") phase, the same participants guessed targets' ratings—including themselves—from the earlier videos (no sound, and just the face showing). Only the videos were shown; the still images viewed by targets in the first phase were never shown to perceivers in the second phase. As in North et al. (2010), no target video clips were excluded from the study.

**Target Phase.** Twenty-eight undergraduates (mean age = 19.96, *SD* = 2.35, 22 female) participated in a "social attitudes from images" study, for course credit.

Participants sequentially viewed a set of 72 images, which composed three different categories (24 each): people, paintings, and animals. Each image appeared for three seconds. Though participants always viewed all 24 people, paintings, and animals respectively (in that order), the order of images within each category was randomized across participants. To reduce the presence of cues such as looking time, participants were instructed to examine each image for the full three seconds before rating it, even if they determined the image's appeal upon first sight. After viewing a given image, participants had five seconds to provide a rating from -10 to 10 indicating their reaction to each person ("How attractive is this person?"), painting ("How much would you like to have this painting on your dorm room wall?") and animal ("How cute is this animal?").

Throughout the task, a built-in computer camera secretly filmed participants' faces. After the experiment, participants were debriefed and asked to sign a film release, or else request that their portion of the recording be deleted. All participants signed the release granting permission to use their footage in subsequent experiments.

***Stimuli.*** Target videos were spliced into individual clips portraying just the 3-second viewing period; only each target's face (and occasionally, shoulders) was in view. A library of 2016 individual clips from the 28 targets was thus available for the next phase.

**Perceiver Phase.** The second phase presented the videos on a computer screen using MediaLab v2008 (Jarvis, 2008). Twenty-six of the initial 28 targets returned to participate as perceivers—at least one month after their initial participation as targets—in exchange for a cash payment.

Because requesting each perceiver to make 2016 individual judgments was not

feasible, the 28 targets were randomly split into groups of four. Each perceiver was asked to judge the four targets within his/her own group (including him/herself): thus perceivers guessed preferences for 288 individual video clips, 72 of which were of him/herself. For each 3-second clip, participants were instructed to guess the targets' ratings of attractiveness, cuteness, or décor, based solely on the facial information contained in the video. The rating employed the same scale (-10 to 10) used by targets.

The order of video clip presentation mirrored the original target phase (people, paintings, animals). The order of video clips for each target was randomized, as was the order of targets within each category. Perceivers viewed 24 consecutive reactions of a given target before moving onto another target; throughout, they were kept informed of the category of the stimuli and preference judgment at issue (e.g., person attractiveness).

## Results

**Target preferences.** For each target and each domain, we computed the difference between the target's maximum and minimum scores in that domain. Across targets ($N = 28$), the median difference for attractiveness, décor, and cuteness was 17, 18, and 19, respectively. Thus, perceivers appeared to make ample use of the 21-point rating scale in indicating their preferences.

**Perceiver guesses.** For each perceiver, we computed four Pearson correlations—one for each of the four targets in his group (including him/herself). Each of the correlations involved 24 pairs: namely, the rating of target versus perceiver for a given image. Thus, zero correlation signifies no ability to read faces, whereas positive correlations reflect successful information transfer. Using a Fisher z-to-r transformation, we subjected these values to additional analyses (see below).

***Other-accuracy.*** For a given perceiver and category, we averaged the perceiver's three other-target (non-self) correlations to calculate an "other-accuracy score" for that category. Mean other-accuracy was significantly greater than zero when guessing painting appeal (Mean $r$ = .14, $SE$ = .04, $t(25) = 3.59$, $p$ = .001, Cohen's d = 0.76) and animal cuteness (Mean $r$ = .21, $SE$ = .02, $t(25) = 8.89$, $p < .001$, Cohen's d = 1.07). Other-accuracy was positive for person attractiveness (Mean $r$ =.06, $SE$ = .04), but not significant ($t(25) = 1.43$, $p$ = .17, Cohen's d = 0.31). When we computed overall other-accuracy from the nine correlations involving others (three categories for each of three targets), perceivers' other-accuracy was reliably positive (Mean $r$ = .14, $SE$ = .03, $t(25) = 5.24$, $p < .001$, Cohen's d = 1.01; see Table 1).

Binomial tests confirmed these findings. Specifically, the number of perceivers (out of 26) with positive other-accuracy was significantly greater than the chance level of 13 for most categories: paintings ($N$ = 20, $p$ = .005), animals ($N$ = 25, $p < .001$), people ($N$ = 15, $p$ = .28, not significant). When we pooled the three categories to create an other-accuracy score based on nine correlations, the number of perceivers with positive scores was again significant ($N$ = 22, $p < .001$; see Table 1).

***Self-accuracy.*** For each of the 26 perceivers, the sole correlation with him/herself as target was used as "self-accuracy score" for a given category. Mean self-accuracy was significantly greater than zero when guessing painting appeal (Mean $r$ = .15, $SE$ = .05, $t(25) = 3.34$, $p$ = .003, Cohen's d = 0.65), animal cuteness ($r$ = .24, $SE$ = .04, $t(25) = 5.43$, $p < .001$, Cohen's d = 0.71), and person attractiveness (Mean $r$ = .21, $SE$ = .07, $t(25) = 3.09$, $p$ = .005, Cohen's d = 0.49). When overall self-accuracy was computed by pooling the three correlations, perceivers' accuracy was reliably positive ($r$ = .19, $SE$ =

.04, $t(25) = 4.81$, $p < .001$, Cohen's d = 0.84; see Table 1).

The number of perceivers with positive self-accuracy was 21, 24, 20 for appeal, cuteness, and attractiveness, respectively; pooling the three correlations, 21 perceivers had positive self-accuracy. These results are all significant via Binomial test ($p$s < .006; see Table 1).[1]

***Comparison of self- versus other-accuracy.*** For each category, and also the pooled category, we compared average other-accuracy with self-accuracy via paired t-tests across the 26 perceivers. None of the differences was reliable; for appeal, cuteness, attractiveness, and the pooled category, $t(25) = 0.24$, 0.58, 1.99, and 1.31, respectively ($p$s > 0.05). The self-other difference for attractiveness, however, approached significance ($p = 0.06$).

We also counted the number of perceivers with greater self-accuracy than other-accuracy. For appeal, cuteness, attractiveness, and the pooled category, 11, 13, 16, and 13 perceivers (out of 26) achieved higher correlations with self than with others. None of these counts is significantly different from chance ($p$s > .10 according to Binomial tests).

***Intercorrelations between accuracy types.*** Finally, we explored intercorrleations between the three different types of accuracy judgments. However, at best only moderate correlations emerged (see Table 2).

## Discussion

These results confirm an earlier demonstration (North et al., 2010) that significant preference information emerges from faces whose owners believe them to be unobserved. The current study goes beyond the prior work (which relied on binary choice) by demonstrating that a continuous assessment of preference yields reliable results. Raters'

sensitivity to faces is all the more remarkable given the experiment's use of low-evocative stimuli, and the inclusion of all recorded facial reactions (no target dropped).

Nevertheless, the correlations obtained were admittedly modest. Such performance echoes prior results (North et al., 2010), and translates to a just-above-chance 57% accuracy score, per Rosenthal & Rubin's (1982) Binomial Effect Size Display conversion (which allows for accuracy score translation between correlational and proportion metrics). Thus, the current paradigm elicited relatively muted (albeit readable) facial affect. Given facial expressions' often-interactive purpose, follow-up studies might explore whether even more readable expressions emerge in a social context (e.g., multiple targets simultaneously viewing the pictures) or with familiar targets (e.g. close friends).

To our knowledge, the present study is the first to compare self- versus other-accuracy in reading mild facial expressions; therefore, it is noteworthy that the data provide little indication that attitudes are more clearly read from the former than the latter. This finding is relevant to embodiment perspectives, including simulation models of face reading (see Goldman, 2006). One such model postulates a generate-and-test loop in which a candidate emotion is simulated on one's own face (perhaps only implicitly, without displacement of facial muscles). The simulated face is then compared to that of the target, with a match leading to emotion attribution. If the simulating agent owns the target face, we might expect that distinguishing matches from non-matches would be easier than for foreign targets. In this case, the predictions of the model differ from our findings; perhaps most people lack extensive knowledge of their facial appearance when programmed to express specific attitudes.

As noted, the present study used all obtained videos, rather than pre-selecting videos based upon a criterion (e.g. apparent expressiveness). Such selection could have artificially inflated the correlations between the ratings of targets and perceivers. Whereas some accuracy paradigms similarly use all recorded sources, others intentionally preselect stimuli known to induce correctness (see Hall et al., 2008, for a review). This practice—as well as variance in exposure durations and video quality—have rendered baseline accuracy levels difficult to measure. As accuracy in social assessments reemerges into the research mainstream (Zaki & Ochsner, 2011), investigators will need to standardize their stimulus-selection practices as much as possible.

References

Ambadar, Z., Schooler, J.W., & Cohn, J.F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science, 16*(5)*,* 403-410.

Baron-Cohen, S., Wheelwright, S., & Jolliffe, T. (1997). Is there a "language of the eyes"? Evidence from normal adults, and adults with autism or Asperger Syndrome. *Visual Cognition, 4*(3), 311-331.

Barr, C. L., & Kleck, R. E. (1995). Self-other perception of the intensity of facial expressions of emotion: Do we know what we show? *Journal of Personality and Social Psychology, 68,* 608-618.

Buck, R. (1979). Measuring individual differences in nonverbal communication of affect: The slide-viewing paradigm. *Human Communication Research, 6,* 47-57.

Ekman, P. (1989). The argument and evidence about universals in facial expressions of emotion. In H. Wagner & A. Manstead (Eds), *Handbook of social psychophysiology.* Chichester: Wiley, pp. 143-164.

Goldman, A.I., & Sripada, C.S. (2005). Simulationist models of face-based emotion recognition. *Cognition, 94,* 193-213.

Goldman, A.I. (2006). The philosophy, psychology, and neuroscience of mindreading. Oxford: Oxford University Press.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology, 72,* 1337-1349.

Gosselin, P., Kirouac, G., & Doré, F.Y. (1995). Components and recognition of facial

expression in the communication of emotion by actors. *Journal of Personality and Social Psychology, 68*(1), 83-96.

Hall, J.A., Murphy, N.A., & Schmid Mast, M. (2007). Nonverbal self-accuracy in interpersonal interaction. *Personality and Social Psychology Bulletin, 33*(12), 1675-1685.

Hall, J.A., Andrzejewski, S.A., Murphy, N.A., Schmid Mast, M., & Feinstein, B.A. (2008). Accuracy of judging others' traits and states: Comparing mean levels across tests. *Journal of Research in Personality, 42,* 1476-1489.

Hess, U., Blairy, S., & Kleck, R.E. (1997). The intensity of emotional facial expressions and decoding accuracy. *Journal of Nonverbal Behavior, 21*(4), 241-257.

Hess, U., Kappas, A., McHugo, G.J., Kleck, R.E., & Lanzetta, J.T. (1989). An analysis of the encoding and decoding of spontaneous and posed smiles: The use of facial electromyopgraphy. *Journal of Nonverbal Behavior, 13*(2), 121-137.

Hess, U., Sénecal, S., & Thibeault, P. (2004). Do we know what we show? Individuals' perceptions of their own emotional reactions. *Current Psychology of Cognition, 22,* 247-265.

Ickes, W. (1997). Introduction. In W. Ickes (Ed.), *Empathic accuracy* (pp. 1-16). New York: Guilford Press.

Jarvis, B. G. (2008). MediaLab (Version 2008.1.33) [Computer Software]. New York, NY: Empirisoft Corporation.

Krumhuber, E., & Kappas, A. (2005). Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. *Journal of Nonverbal Behavior, 29*(1), 3-24.

Niedenthal, P.M., Barsalou, L.W., Winkielman, P., Krauth-Gruber, S., & Ric, F. (2005). Embodiment in attitudes, social perception, and emotion. *Personality and Social Psychology Review, 9*(3)*,* 184-211.

Niedenthal, P.M. (2007). Embodying emotion. *Science, 316,* 1002-1005.

North, M.S., Todorov, A., & Osherson, D.N. (2010). Inferring the preferences of others from spontaneous, low-emotional facial expressions. *Journal of Experimental Social Psychology, 46*(6), 1109-1113.

Rosenthal, R., & Rubin, D.B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166-169.

Wehrle, T., Kaiser, S., Schmidt, S., & Scherer, K.R. (2000). Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology, 78*(1), 105-119.

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathy. *Psychological Science, 19*(4), 399-404.

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion, 9*(4), 478-487.

Zaki, J., & Ochsner, K. (2011). Re-integrating accuracy into social cognition research. *Psychological Inquiry, 22*(3), 159-182.

## Footnote

1. As a non-parametric follow-up, we conducted a Monte Carlo randomization, permuting target and perceiver ratings for each domain ten thousand times. This formed a reference probability distribution of accuracy scores based upon the original data set, to which we could compare the current study's observed accuracy scores. We counted the number of times a given perceiver's other-score was greater than 95% of the other-scores computed from the randomized data. In each case, the success rate exceeded what is expected from chance alone (all $p$s $< .01$), thereby providing corroborating evidence for our findings.

Table 1

*Other- and self-accuracy as a function of stimulus category.*

| | Other | | Self | |
|---|---|---|---|---|
| Category | *r* | No. positive | *r* | No. positive |
| Paintings | 0.14*** | 20** | 0.15** | 21*** |
| Animals | 0.21*** | 25*** | 0.24*** | 24*** |
| People | 0.06 | 15 | 0.21** | 20** |
| Pooled | 0.14*** | 22*** | 0.19*** | 21*** |

***$p \leq .001$; ** $p \leq .01$

Note: *r* is the mean correlation between perceiver and target ratings (based on averages of 1, 3, or 9 coefficients in the self, other, and pooled cases); p-values are determined by t-tests (difference from zero). Number positive is the number of perceivers (out of 26) whose average correlation was positive; p-values are determined by Binomial test (difference from 13, or 50%).

Table 2

*Pearson* r *intercorrelations between accuracy types.*

| | People-Other | Paint-Other | Animal-Other | People-Self | Paint-Self | Animal-Self |
|---|---|---|---|---|---|---|
| People-Other | - | .29 | .11 | .10 | -.23 | -.03 |
| Paint-Other | | - | .12 | .17 | -.13 | .02 |
| Animal-Other | | | - | .14 | .16 | .09 |
| People-Self | | | | - | .22 | .35+ |
| Paint-Self | | | | | - | .08 |
| Animal-Self | | | | | | - |

*Note.* $+p < .10$. Cronbach's $\alpha = .40$ for the 3 types of Other-Accuracy; $\alpha = .45$ for the 3 types of Self-Accuracy. Alphas were fairly low because people who were accurate in one domain weren't necessarily accurate in another, indicating a degree of orthogonality between the three judgments.