

Training Reduces Error in Rating the Intensity of Emotions

Brian T. Leitzke, Rista C. Plate, and Seth D. Pollak
University of Wisconsin–Madison

The ability to recognize how another is feeling is a critical skill, with profound implications for social adaptation. Training programs designed to improve social functioning typically attempt to direct attention toward or away from certain facial configurations, or to improve discrimination between emotions by categorizing faces. However, emotion recognition involves processes in addition to attentional orienting or categorical labeling. The intensity with which someone is experiencing an emotion is also influential; knowing whether someone is annoyed or enraged will guide an observer's response. Here, we systematically examined a novel paradigm designed to improve ratings of facial information communicating emotion intensity in a sample of 492 participants across a series of 8 studies. In Study 1, participants improved precision in recognizing the intensity of facial cues through personalized corrective feedback. These initial findings were replicated in a randomized-control trial comparing training with feedback to viewing and rating faces without feedback. Studies 2 and 3 revealed that these effects generalize to identities and facial configurations not included in the training. Study 4 indicated that the effects were sustained beyond the training session. These findings suggest that individualized, corrective feedback is effective for reducing error in rating the intensity of facial cues.

Keywords: attention training, emotion intensity, facial emotion

Supplemental materials: <http://dx.doi.org/10.1037/emo0000763.supp>

How individuals interpret others' facial configurations has a considerable impact on interpersonal functioning. A perceiver may use facial configurations to make inferences about internal states, intentions, and desires, and apt use of facial cues allows for appropriate responses that facilitate communication (Barrett, Adolphs, Marsella, Martinez, & Pollak, 2019). In contrast, deficits in

this ability can lead to maladaptive, or at least less successful, responses to social situations (Lerner & Arsenio, 2000). For this reason, scientists have sought interventions aimed at improving emotion recognition skills, with the goal of decreasing behavioral problems that often accompany errors in interpreting others' facial cues (Shechner & Bar-Haim, 2016; Stoddard et al., 2016). Extant interventions have focused on using faces for broad, categorical labeling of emotion, which may not map on to how people draw inferences from others' facial movements in their everyday lives (Barrett et al., 2019). For example, recognizing information beyond broad emotion categories—such as if an expresser is mildly annoyed or deeply enraged—provides more nuanced information that is influential in an observer's response. Furthermore, most training paradigms provide participants with only general feedback regarding whether they responded correctly or incorrectly to the task at hand, such as categorizing a face based on a given set of labels; it is possible that more informative feedback that indicates the magnitude, in addition to the direction, of one's response error may improve the effectiveness of training programs (Elfenbein, 2006). Here, we examined a novel approach to training that provided corrective and individualized feedback to improve participants' precision in rating the intensity of emotion-related information as indicated by facial configurations.

Attention Training Interventions

Existing training paradigms target attentional processes that are believed to underlie the development and maintenance of an array of clinical conditions, such as anxiety, depression, and phobias (Cisler & Koster, 2010; MacLeod & Mathews, 2012; Van Bockstaele et al., 2014). Typically, these interventions are designed to

This article was published Online First June 29, 2020.

© Brian T. Leitzke, © Rista C. Plate, and © Seth D. Pollak, Department of Psychology, University of Wisconsin–Madison.

Brian T. Leitzke is now at University of Wisconsin Hospital and Clinics, Madison, Wisconsin. Rista C. Plate is now at the Department of Psychology, University of Pennsylvania.

We thank the individuals who participated in this study, and the research assistants who helped with data collection, particularly Maureen Butler. Funding for this project was provided by the National Institute of Mental Health (MH61285) to Seth D. Pollak and a core grant to the Waisman Center from the National Institute of Child Health and Human Development (U54 HD090256). Brian T. Leitzke, was supported by F31-MH106179 and Rista C. Plate was supported by the National Science Foundation (DGE-1256259) and the Richard L. and Jeanette A. Hoffman Wisconsin Distinguished Graduate Fellowship. The computerized tasks, stimuli, data, and analysis scripts and output are available at: https://osf.io/akgjq/?view_only=8a5a48f954004a3d85c809516bbf742e. Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development.

Correspondence concerning this article should be addressed to Brian T. Leitzke, Department of Psychology, University of Wisconsin–Madison, Waisman Center, 1500 Highland Avenue, Room 399, Madison, WI 53705. E-mail: bleitzke@wisc.edu

reduce or modify maladaptive attention toward disorder-relevant stimuli (e.g., threat). Of the several training programs created to target attention to facial configurations depicting emotion information, the most common has been the visual probe, or “dot probe,” task (MacLeod, Rutherford, Campbell, Ebsworthy, & Holker, 2002).

Visual probe tasks were designed based upon the idea that preferential visual attention toward specific stimuli contributes to the etiology and maintenance of behavioral and mental health difficulties (Shechner & Bar-Haim, 2016). In these types of tasks, participants view two stimuli simultaneously, presented side by side or one above the other. One stimulus is of neutral valence (e.g., a face communicating no specific emotion information), while the other is associated with a targeted valence (e.g., a facial configuration conveying a prototype of anger). Following brief exposure to these stimuli, a probe is presented behind one of the two images and participants are asked to identify a particular characteristic of the probe (e.g., location, orientation). The visual probe tasks allow for the measurement of attentional bias, defined as faster response times to probes located behind the target stimulus (compared with the neutral stimulus), suggesting preferential attention toward the target stimulus. The visual probe paradigm has been modified to serve as a training program. To do so, a contingency is created such that the probe occurs in the location of the neutral stimulus a greater proportion of time than the target stimulus. As participants progress through the task, they learn to preferentially attend away from target cues (such as the anger configuration) in favor of reinforced neutral cues to optimize task performance.

Many reports have indicated that visual probe training paradigms decrease attention biases (Beevers, Clasen, Enock, & Schnyer, 2015; Mogoșe, David, & Koster, 2014; Pergamin-Hight, Naim, Bakermans-Kranenburg, Van Ijzendoorn, & Bar-Haim, 2015; Shechner & Bar-Haim, 2016) as well as reduce psychological symptoms (Beevers et al., 2015; Linetzky, Pergamin-Hight, Pine, & Bar-Haim, 2015; Mogoșe et al., 2014). However, concerns have been raised about the reliability of these effects (Schmukle, 2005). The results of training programs based on visual probe methods yield very modest effect sizes (i.e., small effect sizes for psychological symptom change and medium effect sizes for attention bias change; Mogoșe et al., 2014), reveal large intraindividual variability, and have had a number of failures to replicate (Everaert, Mogoșe, David, & Koster, 2015).

A second type of intervention, interpretation bias training, explicitly targets individuals' ability to discriminate between facial configurations (Penton-Voak, Bate, Lewis, & Munafò, 2012), based upon the assumption that facial cues reliably signal an expresser's internal state. Humans have a propensity to prioritize negative over positive information, such as interpreting faces depicting no emotional information or low levels of happiness as appearing angry; this bias toward negative information is argued to be generally adaptive (Vaish, Grossmann, & Woodward, 2008). In contrast, a tendency to overinterpret ambiguous faces as “angry” is associated with poor outcomes such as chronic irritability (Leibensluft & Stoddard, 2013), aggression (Crick & Dodge, 1996), and long-term social difficulties (Pollak, Messner, Kistler, & Cohn, 2009; Pollak & Sinha, 2002). To address maladaptive emotion discriminability, interpretation bias training aims to shift individuals' identification of faces conveying ambiguous emotional in-

formation—training individuals to assess these faces as less threatening (Penton-Voak et al., 2012; Penton-Voak et al., 2013; Stoddard et al., 2016). In this form of training, participants view exemplars of facial configurations morphed from one display of high intensity (e.g., anger prototypes) to another (e.g., happiness prototypes). Participants then attempt to label the displayed emotion, typically from two forced-choice options. The training component of this program provides feedback (correct vs. incorrect) following stimulus judgment. Interpretation bias training has been implemented in child samples and has yielded encouraging results. Specifically, individuals who decrease their interpretations of ambiguous facial configurations as being “angry” show subsequent reductions in aggression and irritability (Penton-Voak et al., 2012; Stoddard et al., 2016). However, this intervention has not yet provided evidence for generalizability and the categorical nature of the response format is limited and may not map on to the real-world experience of perceiving emotion. Furthermore, relying on discrete emotion categories may be problematic for translation across cultures, and limits the effects to traditional “basic” categories of emotion rather than the full range of emotion experiences (Gendron, Roberson, van der Vyver, & Feldman Barrett, 2014).

Limitations of Intervention Research

The majority of research examining emotion perception has utilized forced-choice paradigms that provide emotion labels, typically from a limited set of prototypes (e.g., angry, happy, sad, etc.). While emotion category labels are ubiquitous in everyday conversation, and can be useful in conveying one's feelings to others, real-time emotion perception likely does not operate categorically (Martinez, 2017). Indeed, a limitation of paradigms focused on facial configurations of common prototypes (such as “anger” or “happiness”), is that people are not restricted to these discrete categories in their everyday interactions. Instead, people use either more general initial distinctions (such as presence/absence of threat, approach vs. avoidance, level of perceived arousal), or more nuanced processes of emotion perception (pleasantly surprised when receiving a desired gift, vs. startled and irritated by an unexpected change) than permitted by discrete emotion categories (Barrett, 2013). Further, using discrete emotion categories assumes that the content in the face conveys a specific internal state of the social partner, a view that is not supported by extant data (Barrett et al., 2019). Thus, it might be useful for the perceiver to attend to nuances that provide clues about intensity.

Additionally, emotion inferences gleaned from facial configurations may be interpreted based on specific features of facial musculature (Martinez, 2017). Differences in such features could indicate subtleties in the subjective state of another person or represent nuances in the emotion information that a social partner wishes to convey. Such details are not well captured by traditional categorical boundaries of emotion. Therefore, an initial assessment of the intensity of a facial display could provide better cues to guide a perceiver's social interaction than a categorical judgment. For this reason, paradigms targeting the perception of facial cues would benefit from allowing participants to improve dimensional, rather than just categorical, ratings. Such paradigms have the potential to unbind training protocols from emotion categories and

labels. However, whether these kinds of judgments are malleable and sensitive to training remains untested.

Current Research

The current research systematically examines the efficacy and generalizability of training aimed at improving individuals' precision in judging the intensity of emotions depicted by facial configurations. To do so, we tested 492 participants across progressive studies. For the first study, we had no a priori evidence on which to base estimated sample or effect sizes. Therefore, we based the sample sizes on previous attention-based intervention programs that are conceptually and methodologically similar to the current research (Penton-Voak et al., 2013; Stoddard et al., 2016). We based subsequent sample sizes on the effect sizes that emerged from our initial study. We first tested whether receiving individualized, corrective feedback could reduce error in rating emotional intensity in faces. Study 1 examined the specificity of the training effects, using a single stimulus for testing and training. Studies 1a and 1b examined these effects with a training group and a randomized control group design, respectively. Studies 2a and 2b examined whether training generalized across stimulus models. Next, Studies 3a, 3b, and 3c tested whether training effects generalized to untrained prototype facial configurations. Finally, Study 4 evaluated the sustainability of training effects over multiple days. Across all eight studies, our hypothesis was that individualized, corrective feedback would improve the precision with which participants perceive dimensional changes in the intensity of emotion depicted by facial configurations.

Analytic Plan

To analyze participants' performance in rating the intensity of emotion depicted in faces, we calculated each participant's error. Error here is defined and calculated as the absolute value of the difference between participants' ratings of intensity and the actual percent morph of each image. We used morphed stimuli to define an objective level of signal intensity in the facial images. Error could range from 0, which indicated a response that matched the actual percent morph exactly, to 100, which indicated a response on the opposite end of the visual analog scale (e.g., a rating of *very happy* when the image was actually 100% angry, or a rating of *very angry* when the image was actually 100% happy). We averaged across all trials at baseline and posttraining. A negative difference score indicates a reduction in error while a positive score indicates an increase in error. We removed all responses that were more than three standard deviations from each participant's mean rating to account for inattentiveness; this accounted for less than 0.05% of the data.

For studies where the outcome of interest involved change from baseline to posttraining within a single group (Studies 1a, 2a, 3a, 3b, and 4), we conducted a linear mixed effects analysis to examine the change in error from baseline to posttraining. We also examined differences in training effects by the emotion prototype conveyed in the face and the stimulus model for which participants were trained. For each statistical model, we included a by-participant random slope for emotion category and percent morph.

For studies where we compared performance between independent groups (Studies 1c, 2b, and 3c), we compared posttraining

performance between groups controlling for baseline performance. We again included a by-participant random slope for emotion category and percent morph. We also examined potential differences in posttraining performance by stimulus model in Studies 1a and 2a. For Study 4, we conducted post hoc comparisons to examine differences between each day of assessment (and controlled for multiple comparisons using the Tukey method). The average age of participants in each study ranged in age from 18 years 5 months to 20 years 6 months, and all participants were students at the same university. Comparing age between studies, participants in Study 4 ($M_{\text{age}} = 20$ years 6 months, $SD = 4$ years 7 months) were older than participants in Studies 1a ($M_{\text{age}} = 19$ years 10 months, $SD = 3$ years 7 months, $p = .05$), 1b ($M_{\text{age}} = 18$ years 5 months, $SD = 9$ months, $p < .001$), 3a ($M_{\text{age}} = 18$ years 9 months, $SD = 12$ months, $p = .04$), and 3c ($M_{\text{age}} = 19$ years 0 months, $SD = 1$ year, 2 months, $p = .03$), ($F(7, 478) = 3.33$, $p = .002$). There were no statistically significant differences in the representation of gender, $\chi^2(7) = 9.27$, $p = .24$, or race/ethnicity, $\chi^2(7) = 56.53$, $p = .066$, between studies.

All analyses were conducted in the R environment (R Core Team, 2018) using the lme4 package, Version 1.1–18.1 (Bates, Maechler, Bolker, & Walker, 2015). Means and standard deviations at baseline and posttraining for each study are presented in Table 1.

Study 1: Effectiveness of Training

Study 1a: Intensity Training

Method. All studies reported in this article were approved by the Institutional Review Board. Participants were undergraduates at a large university in a Midwestern city who participated for course credit. Participants completed the task individually in a dedicated testing room. The task lasted approximately 20–30 min. All participants provided written consent.

Participants. One-hundred and six young adults ($M_{\text{age}} = 19$ years 10 months, $SD = 3$ years 7 months; 70% female, 30% male; 30% Asian, 5% Black or African American, 2% Hispanic, 3% Indian, 1% Middle Eastern, 59% White) participated in Study 1a. We aimed for a large sample size given the possibility that the effects of a computerized intervention may be small.

Stimuli. Stimuli were borrowed from Gao and Maurer (2009). This stimulus set consisted of facial stimuli selected from the MacArthur Network Face Stimuli Set (NimStim; Tottenham et al., 2009). Stimuli depicted prototypes of anger, happiness, and neutrality, herein referred to as angry, happy, and neutral faces, from four different models, consisting of two White males (NimStim: 24M, 25M) and two White females (NimStim: 03F, 10F). For each model, faces were morphed to create 21 images, equally spaced on continua from 100% anger to neutral and 100% happiness to neutral. This created a continuum of 40 equally spaced faces (see Figure 1). We excluded the 100% neutral face to ensure that each facial morph contained a percentage of both the anger or happiness prototype. We randomly counterbalanced participants to receive one of the four stimulus models to view and rate throughout the task, without consideration of the sex of the participant. Participants viewed the stimuli on a 21-in. computer monitor presented with E-Prime software. Each trial consisted of a fixation cross that remained on screen for a random duration between 1,000 and

Table 1
Means and Standard Deviations for Baseline and Posttraining for All Studies

Study	Baseline		Posttraining			
	<i>M (SD)</i>		<i>M (SD)</i>			
Study 1: Effectiveness of training						
1a: Training	8.515 (6.721)		6.156 (6.296)			
	Training	Control	Training	Control		
1b: RCT	8.505 (6.966)	8.954 (7.522)	6.007 (5.841)	7.920 (7.118)		
Study 2: Generalization beyond training model						
2a: Model generalizability	8.559 (6.710)		6.740 (6.155)			
	Trained	Untrained	Trained	Untrained		
2b: Robustness of generalizability	9.565 (7.787)	9.006 (7.219)	6.979 (7.241)	7.644 (7.351)		
Study 3: Generalization beyond training facial configuration						
3a: Other facial configuration	9.234 (7.752)		6.869 (7.184)			
3b: Configuration generalizability	9.361 (7.642)		6.807 (5.990)			
	Trained	Untrained	Trained	Untrained		
3c: Robustness of generalizability	8.490 (7.269)	8.813 (7.439)	6.492 (7.762)	7.154 (7.353)		
	Day 1					
	Baseline	Posttraining	Day 2	Day 3	Day 4	One week after Day 4
Study 4: Sustained effects over time						
4: Sustainability of effects	8.885 (7.445)	6.080 (6.025)	6.910 (6.582)	6.286 (6.553)	6.297 (7.000)	5.777 (6.016)

Note. RCT = randomized-control trial.

2,000 ms, in 100 ms increments. This was followed by stimulus presentation (2,000 ms), a visual noise mask (150 ms), and a response screen. The response screen remained on display until each participant rated the intensity of the face (see Figure 2 for task sequence).

Design. All participants in this study completed the same training task that consisted of three blocks: baseline, training, and posttraining. The only between-subjects difference was the stimulus model on which participants were trained and tested. During the baseline and posttraining blocks, which were identical, participants viewed each of the 40 facial morphs of their assigned model in randomized order. Participants saw the same training model in all three blocks. Following the presentation of each image, participants viewed a response screen that asked "How happy or angry was that person?" Participants made their response by using a computer mouse to move a marker along a visual analog scale that ranged from *completely angry* to *completely happy*. A vertical line indicated the center point of the scale.

Following completion of the baseline block, participants completed the training block, where they viewed each of the 40 facial morphs three times, in random order, for a total of 120 trials. Participants received corrective feedback after each trial. The feedback allowed participants to view their previous response to the actual percent morph of the image by its placement on the visual analog scale. This was the only information presented, and participants did not receive any reward or other indication of being correct/incorrect.

Procedure. Participants completed practice trials with nonfacial stimuli before beginning the task. Practice consisted of two trials where participants viewed two different shades of color and rated them on a scale that ranged from one color to the other. An experimenter explained and supervised the practice trials to ensure understanding of the rating scale. As the purpose of this task was to alter ratings, practice did not require participants to achieve any particular performance threshold. Rather, practice allowed participants to acclimate to the visual analog scale without exposure to

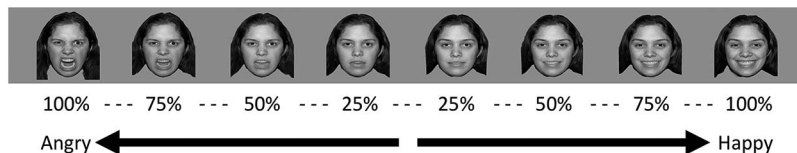


Figure 1. Examples of facial stimuli; stimuli were morphed from a full anger prototype to neutral and neutral to a happiness prototype in 5% increments for each model (25% increments shown here). The neutral image was excluded from each stimulus set. We also included stimuli morphed from fear and sadness prototypes to neutral in Study 3. We obtained permission from the MacArthur Foundation Research Network on Early Experience and Brain Development to reprint select faces from the MacBrain Face Stimulus set.

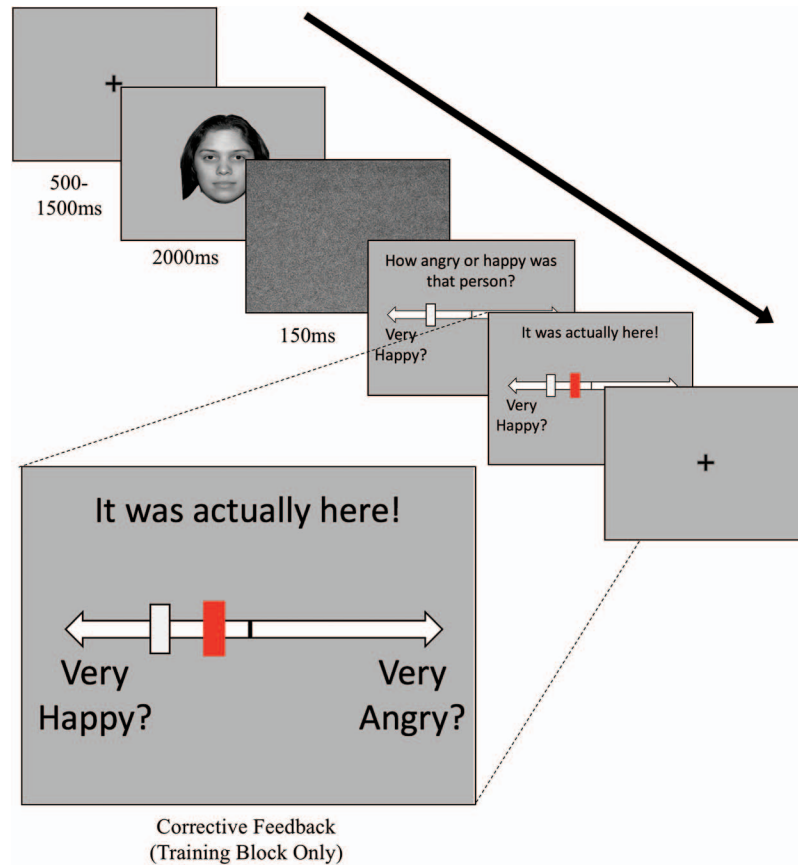


Figure 2. Example of task sequence. Corrective feedback screen presented only during training blocks; participant response (indicated by white rectangle) relative to individualized, corrective feedback (indicated by red rectangle). Studies 1a, 1b, 2a, 2b, and 4 asked participants to infer the intensity of “anger” and “happiness” from facial configurations; Study 3a, 3b, and 3c also included ratings of “fear” and “sadness.” We obtained permission from the MacArthur Foundation Research Network on Early Experience and Brain Development to reprint select faces from the MacBrain Face Stimulus set. See the online article for the color version of this figure.

faces prior to beginning the task. Following the practice trials, participants completed the task.

Results. Participants demonstrated reliable improvements in rating intensity as evidenced by a reduction in error in intensity ratings relative to the actual percent morph from baseline to posttraining ($b = -1.878$, $\chi^2(1) = 47.302$, $p < .001$, 95% CI $[-2.413, -1.343]$). Comparatively, participants’ error in rating intensity was reduced to a greater extent for happy faces relative to angry faces ($b = 1.367$, $\chi^2(1) = 6.267$, $p = .01$, 95% CI $[-2.437, -0.297]$; see Figure 3; baseline and posttraining scores across the morph continuum shown in Figure 4). With regard to possible model-specific effects on posttraining intensity ratings, we found no difference in training effects between any of the four stimulus models, $\chi^2(3) = 0.001$, $p = .999$; all b s < -0.243 , all p s $= 1.0$. Given that there were no differences between training models, we removed stimulus model as a fixed effect in subsequent analyses but continued to include it as a random effect.

Study 1b: Randomized Control Trial

Extending the results from Study 1a, we directly compared the training condition to the control condition. To do so, we conducted

a subsequent experiment where participants were randomly assigned to receive training as in Study 1a or receive no feedback during the training block. Of note, we initially ran a separate study whereby participants did not receive feedback during the training portion (see [online supplemental materials](#)); however, as this was an exploratory study and participants were not randomly assigned to this condition, we ran a separate study with random assignment.

Method. Sixty-five participants who did not participate in Studies 1a (sample size chosen according to post hoc power analysis from Study 1a that indicated that 32 participants per cell was sufficient to detect the effect of interest with .9 power) were randomly assigned to receive training ($N = 33$; $M_{\text{age}} = 18$ years 6 months, $SD = 12$ months; 48% female, 52% male; 24% Asian, 3% African American, 3% Hispanic, 3% Indian, 67% White) or complete the control task ($N = 32$; $M_{\text{age}} = 18$ years 5 months, $SD = 7$ months; 63% female, 37% male; 19% Asian, 3% Hispanic, 78% White). Stimuli and procedures were identical to Study 1a for the training condition. In the control condition participants viewed and rated the faces but did not receive any feedback.

To compare performance between those who received training with feedback versus those who completed the control task, we

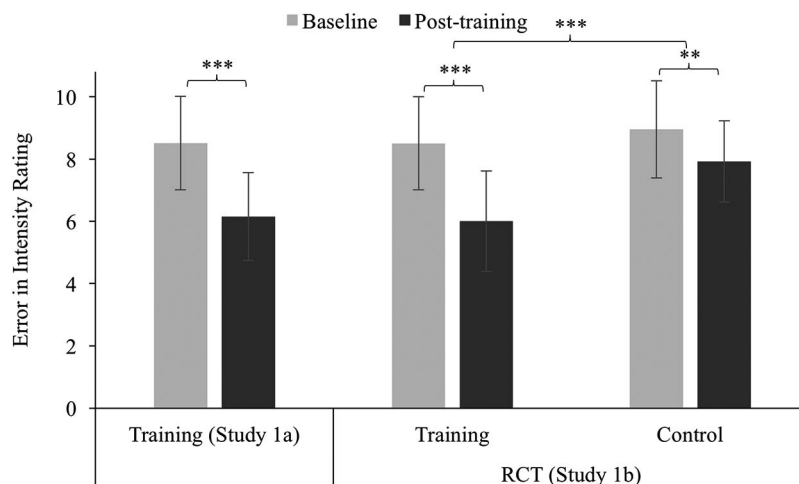


Figure 3. Change in error in intensity ratings for Study 1. Study 1a involved baseline, training, and posttraining on the same model with training consisting of corrective feedback. Study 1b was a randomized-control trial (RCT) whereby participants were randomly assigned to receive training, as in Study 1a or a control that involved baseline, training, and posttraining on the same model with no feedback of any kind during training. Scores collapsed across ratings for “angry” and “happy” facial configurations. Lower values indicate less error; a score of zero indicates no error relative to actual percent morph. Error bars indicate 95% confidence intervals. ** $p < .01$. *** $p < .001$.

compared performance at posttraining between the two conditions controlling for performance at baseline. We again included random effects for model and emotion and percent morph by participant.

Results. Participants who underwent training with feedback showed lower posttraining error in intensity ratings than those who completed the control condition ($b = -1.939$, $\chi^2(1) = 26.128$, $p < .001$, 95% CI $[-2.668, -1.209]$). Similar to Study 1a, precision in intensity ratings following training was better for happy faces than angry faces ($b = -1.759$, $\chi^2(1) = 51.405$, $p < .001$, 95% CI $[-2.23, -1.272]$).

We also conducted a linear mixed effects analysis examining the change in error from baseline to posttraining between the training and control conditions to test whether the magnitude of the training effects differed by condition. Both groups demonstrated reductions in error from baseline to posttraining (feedback: $b = 2.489$, $p < .001$, 95% CI $[1.987, 2.991]$; no feedback: $b = 1.035$, $p = .003$, 95% CI $[0.531, 1.539]$). However, participants who received feedback during training showed a greater change in ratings from baseline to posttraining than did those who did not receive feedback, $b = -1.448$, $\chi^2(1) = 15.560$, $p < .001$, 95% CI

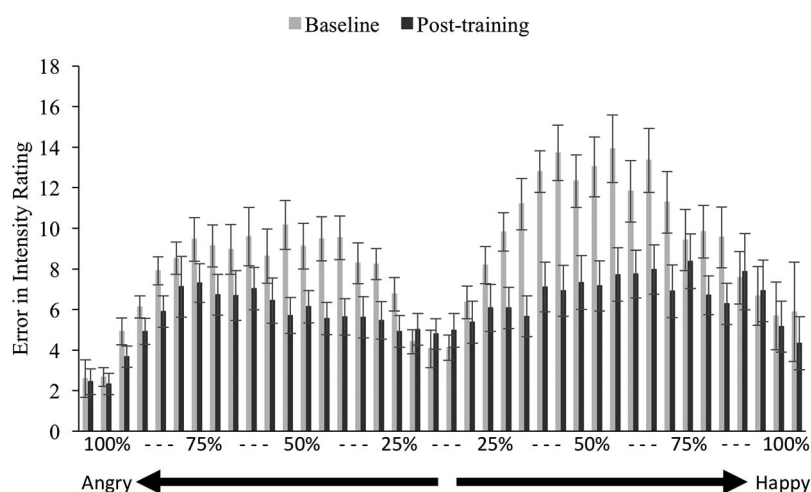


Figure 4. Error in intensity ratings across morph continuum at baseline and posttraining for Study 1a. Lower values indicate less error; a score of zero indicates no error relative to actual percent morph. Error bars indicate 95% confidence intervals.

$[-2.168, -0.730]$. Post hoc tests indicate that while baseline performance between those in the training and control groups was similar, $b = 0.752$, $p = .296$, 95% CI $[-0.089, 1.593]$, posttraining values were lower for those in the training relative to the control group, $b = 2.206$, $p < .001$, 95% CI $[1.363, 3.049]$.

Discussion of Studies 1a and 1b

Taken together, results from Study 1 demonstrate that training with individualized, corrective feedback improved people's precision in rating the intensity of facial configurations. Participants who received training with feedback in Study 1a and 1b showed 27% and 29% reductions in error, respectively. These training effects could not be accounted for by exposure to the facial displays alone, as participants who viewed morphed facial configurations but received no corrective feedback showed less reduction in error (11% reduction in error). It is possible that the individualized feedback provided information to help participants calibrate their perceptual learning. But this individualized feedback may have also helped recruit and engage the participants' attention and motivation throughout the task as they sought to reduce their error gaps in subsequent trials. Notably, some reduction in error occurred whether or not participants received feedback on their responses. While mere exposure may be sufficient to slightly improve the ability to recognize changes in facial musculature, individualized feedback along with exposure resulted in a greater degree of improvement than viewing faces without feedback.

High intensity prototypes of anger, such as those displayed at the 100% endpoints of the morph continua, are rarely seen in the real world (Calvo & Nummenmaa, 2016) and may be of limited utility. Yet, the greatest improvements in intensity ratings were found in the center of the continua (see Figure 4), representing more subtle changes in facial musculature. This facial information is both more prevalent in daily interactions and likely more difficult for social partners to interpret.

Study 2: Generalizability to Untrained Model

The results from Study 1 demonstrated that training improved accuracy in rating intensity. However, because participants saw the same model at baseline, training, and posttest, we do not know whether the effects of the training generalize beyond the stimuli presented. It could be that participants are simply mapping the model's particular facial configuration to the correct response. An effect this specific would be less relevant to actual social behavior because there is variance in the way individuals express emotion. Therefore, we conducted a pair of studies to test the extent to which the effects of training generalize to untrained facial stimuli. Study 2a tested whether training on one model would lead to improved posttraining accuracy when participants rated the intensity of another individual who was not presented during training. Study 2b compared the effects of training on models viewed during training versus models not viewed during training, but included pre- and posttraining accuracy measurements on both trained and untrained faces. Because this was our first test of generalizability, we exceeded the sample size beyond that indicated by the post hoc power analysis from Study 1a in the event that the effect size for generalizability would be smaller.

Study 2a: Model Generalizability

Method. This study included 96 participants ($M_{\text{age}} = 19$ years 3 months, $SD = 2$ years 10 months; 66% female, 34% male; 23% Asian, 1% Black or African American, 2% Hispanic, 1% Indian, 73% White). The stimuli and procedures were identical to that of Study 1a, with the following exceptions. Instead of seeing a single model throughout baseline, training, and posttraining, participants saw one model at baseline and posttraining and a different model during training. Participants were randomly counterbalanced and evenly distributed to all possible model combinations.

Results. Training generalized from one model to another. Participants showed reductions in error from baseline to posttraining when tested on an individual other than the one they were exposed to during training ($b = -1.345$, $\chi^2(1) = 7.836$, $p = .005$, 95% CI $[-2.284, -0.405]$; see Figure 5). There were no differences in reduction in error from baseline to posttraining between happy and angry faces ($b = -0.144$, $\chi^2(11) = 0.026$, $p = .881$, 95% CI $[-2.023, 1.734]$), nor were there differences in posttraining performance by model combination, $\chi^2(11) = 0.074$, $p = 1.0$.

Study 2b: Robustness of Model Generalizability

Study 2a demonstrated that the training effects generalized to a novel model. However, Study 2a cannot speak to the robustness of the effect. Study 2b addressed this question by including both a trained and untrained model at baseline and testing, allowing us to directly compare improvements on trained and untrained models.

Method. Ninety-four participants ($M_{\text{age}} = 19$ years 5 months, $SD = 3$ years 5 months; 67% female, 33% male; 18% Asian, 5% Black or African American, 5% Hispanic, 1% Middle Eastern, 71% White) completed Study 2b. Procedures in this study were the same as in Study 1a with the following exceptions. Participants were trained on one model, as in Study 1a; however, baseline and posttraining included both the model presented during training and a model not presented during training. Trained and untrained models each comprised half of the pre- and posttest trials. Because participants viewed two models at baseline and posttraining, participants viewed facial emotions morphed in 5% increments for each model (as opposed to 2.5% as in the previous studies) to ensure that the task length was consistent with Study 1a. Participants were again randomly counterbalanced and evenly distributed across all model combinations (for both trained and untrained models).

To compare differences between trained versus untrained models, we compared performance at posttraining between trained and untrained models, controlling for baseline ratings. We also examined differences by emotion with random effects for model and emotion and percent morph by participant.

Results. Results from this study revealed lower posttraining error for models viewed during training relative to those not viewed during training ($b = -0.743$, $\chi^2(1) = 12.053$, $p < .001$, 95% CI $[-1.163, -0.324]$). While there were reductions in error for both trained ($b = -2.589$, $\chi^2(1) = 134.374$, $p < .001$, 95% CI $[-3.026, -2.151]$) and untrained models ($b = -1.376$, $\chi^2(1) = 39.087$, $p < .001$, 95% CI $[-1.807, -0.945]$), the degree of this change was greater for models for which participants received corrective feedback during training ($b = -1.21$, $\chi^2(1) = 14.723$, $p < .001$, 95% CI $[-1.828, -0.592]$). There was no difference in

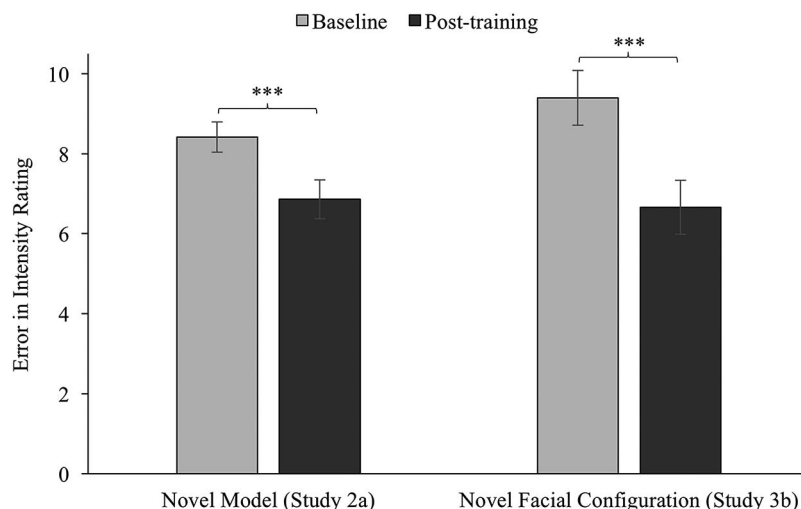


Figure 5. Change in error in intensity ratings for those trained and tested on different models or facial configurations. In Study 2a, participants completed baseline and posttraining on one model with training on another; Study 3b involved training on anger/happiness facial configurations and baseline and posttraining assessment on fear/happiness or sadness/happiness facial configurations (same model). Scores for Study 3b are collapsed across ratings for fearful and happy, and sad and happy facial prototypes. Lower values indicate less error; a score of zero indicates no error relative to actual percent morph. Error bars indicate 95% confidence intervals. *** $p < .001$.

performance based on the combination of trained and untrained models, $\chi^2(1) = 0.733$, $p = .392$, 95% CI $[-0.440, 1.124]$.

Discussion of Studies 2a and 2b

These two studies provide evidence that the effects of training generalize beyond the training stimuli to new individuals. While improvements were greater for models viewed during training, participants also demonstrated improvement at rating intensity of untrained models as well. This suggests that the training effects are not specific to presented stimuli, but have the possibility of generalizing to faces or individuals encountered during training.

Study 3: Generalizability to Untrained Facial Configurations

Given the generalizability of the training effects observed in Studies 1 and 2, we next tested whether the training would be effective for emotion prototypes other than anger and happiness. Facial configurations associated with “anger” may signal an environmental threat and especially attract attention. In addition, we sought to determine if the training would generalize to completely different sets of muscular configurations than seen during training. To address these questions, Study 3a tested whether the training program would be effective for improving intensity ratings of facial configurations associated with prototypes of fear and happiness. Studies 3b and 3c examined whether training on one emotion category prototype would lead to improved accuracy in rating intensity of a facial configuration not encountered during training.

Study 3a: Training on a Fearful Facial Prototype

Method. Thirty-three participants ($M_{\text{age}} = 18$ years 9 months, $SD = 12$ months; 52% female, 48% male; 15% Asian, 3% Black

or African American, 3% Hispanic, 3% Middle Eastern, 76% White; sample size reduced due to medium-large effect sizes obtained in Study 2) were trained on a continuum of fear-to-happiness. The procedures were identical to that of Study 1a. We used the same four models (NimStim White females 03F, 10F; White males 24M, 25M) as in Study 1a, however, in this study the facial images of each model were morphed from “fear” to “happiness.”

Results. Similar to the findings in Study 1a, training reduced error in intensity ratings ($b = -2.693$, $\chi^2(1) = 22.906$, $p < .001$, 95% CI $[-3.794, -1.593]$). Participants demonstrated greater reduction in error for facial configurations representing happy relative to fear ($b = -2.021$, $\chi^2(1) = 3.224$, $p = 0.073$, 95% CI $[-4.222, 0.181]$), as observed in Study 1a.

Study 3b: Facial Configuration Generalizability

We next tested whether training effects would generalize to a different set of facial configurations. Participants in this study were trained on facial configurations associated with anger and happiness, as in Study 1a, but viewed facial configurations of fearful and happy prototypes during baseline and posttraining. In other words, participants were trained on facial configurations associated with anger and happiness, but not fear.

Method. Thirty-two participants ($M_{\text{age}} = 18$ years 11 months, $SD = 1$ year, 4 months; 66% female, 34% male; 38% Asian American, 59% White, 3% did not disclose; sample size consistent with Study 3a) were trained on “angry” and “happy” faces (morphed on a continuum) but were tested at baseline and post-training on “fearful” and “happy” faces (morphed on a continuum). The procedure was identical to Study 1a and stimuli were the same facial morphs used in Studies 1a and 3a, respectively.

Results. Training on prototypical angry and happy facial configurations yielded reductions in error in rating the intensity of

faces on the fear-happy continuum ($b = -2.971$, $\chi^2(1) = 27.795$, $p < .001$, 95% CI $[-4.073, -1.869]$; see Figure 5). Participants again demonstrated greater improvement in error when rating happy faces than fearful faces ($b = -2.252$, $\chi^2(1) = 3.994$, $p = .046$, 95% CI $[-4.456, -0.048]$).

Study 3c: Robustness of Facial Configuration Generalizability

In this study, we directly compared changes in intensity ratings for facial images viewed during training (anger and happiness prototypes) and those not seen during training (fear and happiness or sadness and happiness prototypes).

Method. Sixty-four participants ($M_{\text{age}} = 19$ years 0 months, $SD = 1$ year, 2 months; 75% female, 25% male; 25% Asian, 2% Hispanic, 73% White; sample size for each emotion prototype trained consistent with Study 3a) received training with one model whose facial images were morphed from depicting anger and happiness as in Study 1a. However, during the baseline and posttraining blocks, participants were tested on images of the same model morphed from depicting either fear to happiness ($N = 32$) or sadness to happiness ($N = 32$). Thus, for half of the trials, participants viewed stimuli drawn from the same morphed continuum as they viewed during training, and for the other half, they viewed stimuli morphed from “happy” facial configurations to an emotion category they did not encounter during training. Fear and sadness prototypes were chosen to test generalizability to other emotion categories of negative valence. Stimulus models were the same four as in all previous studies. In the baseline and posttraining blocks, facial morphs ranged from angry faces, fearful prototype configurations, or sad to happy prototype configurations in 5% increments to maintain consistent task length across studies. Participants were evenly distributed and counterbalanced to all model combinations.

Similar to Study 2b, we examine performance at posttraining between trained and untrained facial configurations controlling for baseline ratings. We also examined the difference between positively valenced and negatively valenced emotion categories, and again included random effects for model, emotion category, and percent morph by participant.

Results. While there were differences between trained and untrained stimuli in posttraining error, participants showed reduction in error from baseline to posttraining for both trained (anger to happiness; $b = -1.484$, $\chi^2(1) = 7.085$, $p = .008$, 95% CI $[-2.575, -0.394]$) and untrained emotions (fear to happiness and sadness to happiness; $b = -0.964$, $\chi^2(1) = 3.246$, $p = .072$, 95% CI $[-2.011, 0.083]$). Moreover, there was no difference in the degree of improvement from baseline to posttraining between trained and untrained emotions ($b = -0.352$, $\chi^2(1) = 0.786$, $p = 0.375$, 95% CI $[-1.128, 0.425]$). Similar to previous studies, participants showed greater improvements for “happy” facial configurations relative to negatively valenced facial configurations (angry/fearful/sad; $b = -1.249$, $\chi^2(1) = 9.914$, $p = 0.002$, 95% CI $[-2.025, -0.472]$). Posttraining error did not differ between fearful or sad faces ($b = -0.015$, $\chi^2(1) = 0.001$, $p = 0.981$, 95% CI $[-1.288, 1.275]$).

Discussion of Studies 3a, 3b, and 3c

Study 3 extended assessment of the training paradigm in three ways. First, training was effective for improving accuracy for fear prototypes similarly to the effects observed for happy and angry prototypes. Second, training on one type of facial configuration (i.e., angry, happy) led to improved accuracy on an untrained configuration (i.e., fear). Third, training reduced participant’s errors in rating untrained emotion categories to the same extent as it reduced errors on trained categories. Taken together, the evidence suggests that the effects of the training are robust and transfer to new stimuli.

Study 4: Sustainability of Effects

Prior research on interpretation bias training found sustained improvements after 4 days of training (Penton-Voak et al., 2012; Stoddard et al., 2016). Therefore, Study 4 examined whether improvements in intensity ratings were sustained for 1 week following four consecutive days of training.

Method

Fifty-eight young adults ($M_{\text{age}} = 20$ years 6 months, $SD = 4$ years 7 months; 67% female, 33% male; 38% Asian, 5% African American, 2% Hispanic, 55% White) who completed Study 1a agreed to return to complete a total of four consecutive days of training followed by a follow-up assessment 7 days after the fourth day of training. Of the initial 58 participants who agreed to participate in this multiday training, two did not return after the first day, one discontinued after 2 days, one discontinued after 3 days, and 12 did not return for the follow-up assessment 1 week later. Therefore, 42 young adults ($M_{\text{age}} = 20$ years 11 months, $SD = 5$ years 3 months; 69% female; 5% African American, 38% Asian American, 2% Hispanic, 55% White) completed all five sessions and are included here. There were no differences in baseline performance between participants who did and did not complete all 5 days of training, $t(56) = -1.09$, $p = .282$, $d = .14$, 95% CI $[-1.47, 0.44]$. However, those who went on to complete all 5 days showed greater gains in accuracy from baseline to posttraining on Day 1 than those who did not, $t(56) = 2.66$, $p = .010$, $d = .35$, 95% CI $[0.43, 3.03]$. Participant recruitment and compensation, as well as task stimuli and procedures for the first session of this multiday training were identical to that of Study 1a. During the second, third, and fourth sessions, participants completed only the baseline and training blocks of Study 1a, which lasted approximately 15 min. In the fifth and final session, participants completed only the posttraining block, which lasted less than 10 min. We randomly counterbalanced participants to one of the four stimulus models on which they were trained and tested during all sessions.

Results

The subset of participants involved in the multiday training replicated results from Study 1a, showing improvements in intensity ratings from baseline to posttraining on Day 1 ($b = -2.798$, $p < .001$, 95% CI $[-3.161, -2.436]$). Participants showed continued improvements in accuracy on Days 2, 3, and 4 relative to

baseline on Day 1 (all $ps < .001$). These improvements were maintained 1 week after the conclusion of training ($b = -2.929$, $p < .001$, 95% CI $[-3.309, -2.559]$).

Examining performance beyond Day 1, error in rating intensity on Day 2 was reduced from baseline on Day 1 ($b = -1.988$, $p < .001$, 95% CI $[-2.359, -1.617]$) though not compared with *post-training* on Day 1 ($b = -0.811$, $p = .003$, 95% CI $[0.440, 1.181]$). Error in ratings on Day 3 of training was improved from baseline on Day 1 ($b = -2.631$, $p < .001$, 95% CI $[-3.008, -2.253]$) and comparable with posttraining on Day 1 ($b = 0.168$, $p = 0.869$, 95% CI $[-0.210, 0.545]$). A similar pattern was found for performance on Day 4, with improved error relative to baseline on Day 1 ($b = -2.648$, $p < .001$, 95% CI $[-3.023, -2.272]$), and similar performance compared to posttraining on Day 1 ($b = 0.151$, $p = 0.787$, 95% CI $[-0.224, 0.526]$).

To measure the impact of multiple days of training on performance, we examined change in error from both baseline and posttraining on Day 1 to error measured 1 week following completion of all 4 days of training. Relative to baseline on Day 1, multiple days of training resulted in improved error in rating intensity 1 week following completion of training ($b = -2.929$, $p = 0$, 95% CI $[-3.309, -2.549]$). These patterns are summarized in Figure 6. Error 1 week after training was no different than that at posttraining on Day 1 ($b = 0.130$, $p = 0.985$, 95% CI $[-0.249, 0.510]$), Day 3 ($b = 0.298$, $p = 0.67$, 95% CI $[-0.093, 0.689]$), or Day 4 ($b = 0.282$, $p = 0.716$, 95% CI $[-0.107, 0.671]$). However, error 1 week after training was improved from error on Day 2 ($b = 0.941$, $p < .001$, 95% CI $[0.556, 1.328]$).

Discussion of Study 4

Study 4 indicates that improvements in error achieved following four consecutive days of training were sustained 1 week later. Error improved after training on Day 1, replicating Study 1a, but then dropped off on Day 2. This pattern of results suggests that a single day of training may not be sufficient to achieve lasting effects. Consistent with this view, accuracy levels returned to Day

1 posttraining levels after subsequent training on Day 2. Participants' improvements were sustained from Day 3 through the 1-week follow-up. Participants who did not return to complete all 5 days of training did not differ from those who did complete training on baseline (pretraining) performance. Even though baseline performance did not differ, participants who did not complete all 5 days of training had higher error rates at posttraining on Day 1. While both groups did improve during Day 1, it is possible that those who benefited less from the training self-selected out of the study.

Comparison of Effects Across Studies

We conducted an internal meta-analysis to better understand the overall effect of the training. Such analyses provide information about the consistency of effects and confidence for interpreting cumulative results (see Goh, Hall, & Rosenthal, 2016 for further discussion on this topic). We excluded Study 4 from this analysis because the participants from Study 4 were also included in Study 1a. We also excluded participants assigned to the control condition in Study 1b because we did not hypothesize that they would show a reduction in error. Analyses were conducted using the "meta" package in R with the means and standard deviations of error at baseline and posttraining across the remaining studies reported here. These results indicate the training paradigm was effective in reducing error when rating facial configurations across all studies where participants received corrective feedback (fixed effects model: $SMD = -.256$, $z = -3.87$, $p < .001$, 95% CI $[-0.386, -0.126]$; random effects model: $SMD = -.255$, $t = -6.15$, $p < .001$, 95% CI $[-0.356, -0.153]$). There was no evidence of heterogeneity for either model ($ps > .8$).

General Discussion

The studies presented here provide initial evidence of the efficacy, generalizability, and short-term sustainability of a novel paradigm to improve people's precision in perceiving changes in

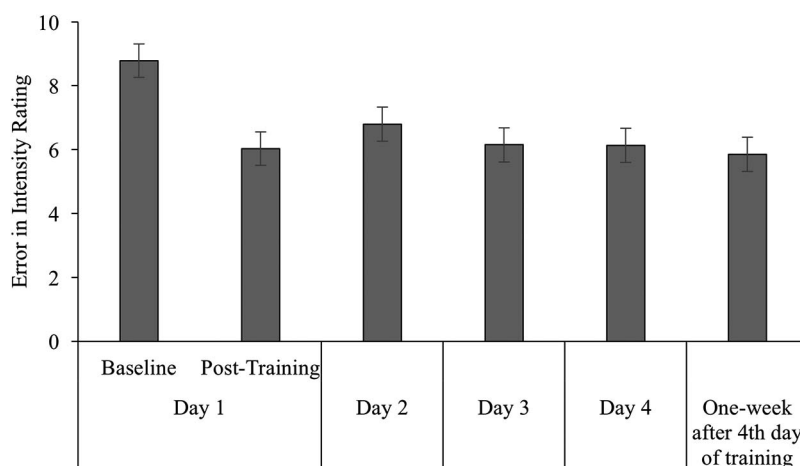


Figure 6. Change in error in intensity ratings for Study 4. Baseline and posttraining were completed on Day 1; only baseline and training completed on Days 2 through 4; follow-up session completed 1 week following four consecutive days of training consisted of posttraining only. Lower values indicate less error while a score of zero indicates no error relative to actual percent morph. Error bars indicate 95% confidence intervals.

the intensity of facial configurations used to infer emotions. Individualized, corrective feedback increased participants' attention to the intensity of signal strength conveyed in the face. This yielded reductions in error that generalized beyond the trained stimuli and that were sustained 1 week following training (see Figure 7 for summaries).

The current research was designed to test whether individuals could improve in their ability to rate emotion intensity through facial configurations. Participants generally performed relatively well at baseline, with error rates less than 10 points from the calculated percent morph along each emotion continuum. Improvements were generally on the order of 2–3 points. While this change may appear modest, it represented a 25–40% improvement from baseline performance and yielded consistent and reliable reductions in error.

Mechanisms of Change

This training procedure did not include any extrinsic rewards or incentive to enhance motivation, or punishment to indicate whether participants were incorrect. Participants saw only their response and the gap in relation to the actual percentage of the morph to which they had just responded. This feedback provided the magnitude and directionality of one's error. Our working hypothesis is that as participants became more engaged in trying to reduce the magnitude of their errors, they heighten and sustain their attention to the faces. Thus, intrinsic motivation, in addition to attention, may be a factor in this effect. On this view, partici-

pants were developing a meta-awareness and strategy that generalized beyond any emotion prototype or specific set of facial musculature. The feedback itself may have provided enough motivation to recruit and enhance attentional processes (see Engelmann & Pessoa, 2007). Therefore, the combination of repeated exposure paired with personalized and specific feedback may engage attention and/or motivation to approach the task in a way that reduced errors. An alternative possibility is that the effects are driven primarily (if not solely) by bottom-up perceptual processes and do not bear on the perceiver's conscious or top-down emotion concepts. On this view, the pattern of results could reflect changes in how participants perceived changes of facial muscles across images. That there was some improvement in the control condition of Study 1b provides evidence that such mechanisms may be involved. It is likely that changes in one's bottom-up perception would in turn influence emotion concepts, however, this set of studies cannot directly disentangle the precise level of processing affected.

Relevant to the consideration of mechanisms, the classic visual-probe paradigm does not provide the participant with feedback, which may in part explain why effects are inconsistent. Feedback may influence participants' motivation to attend more closely to features of the face, therefore reducing error. Perhaps closer attention to the facial features via bottom-up perceptual processing, paired with feedback integration, improved perceptual precision, an effect that would also result in error reduction. However, it is not clear whether simply being reminded or encouraged to pay









Summary of Task Sequences and Results Summaries				
Study	N	Design (Pretest - Training - Posttest)	Example (Pretest - Training - Posttest)	Summary
Study 1: Basic Effectiveness of Training				
1a	106	A - A - A		Training on one model displaying one emotion prototype improves rating on the same model and facial configuration
1b	65	A - A - A (RCT: Training vs. Control)		Training with feedback leads to greater improvements than exposure to stimuli without feedback via randomized-control trial
Study 2: Generalizability to Untrained Model				
2a	96	A - B - A		Training on one model generalizes to improvements on a novel model
2b	97	AB - A - AB		Training on one model generalizes to improvements on more than one model
Study 3: Generalizability to Untrained Facial Configuration				
3a	33	A _{fear} - A _{fear} - A _{fear}		Training generalizes beyond facial configurations tested in Study 1
3b	32	A _{fear} - A _{anger} - A _{fear}		Training on one facial configuration generalizes to improvements on a different facial configuration
3c	64	A _{anger} B _{fear} - A _{anger} - A _{anger} B _{fear}		Training on one facial configuration generalizes to improvements on more than one facial configuration
Study 4: Sustained Effects over Time				
4	42	A - A - (one week) - A		Training effects are sustained at least one week after training

Figure 7. Task sequences for each study with pictorial representations. Facial images are example stimuli. "A" represents one model; "B" represents a model different from "A"; RCT = randomized control trial. Studies 3a-c included the same model through all phases of the design. We obtained permission from the MacArthur Foundation Research Network on Early Experience and Brain Development to reprint select faces from the MacBrain Face Stimulus set. See the online article for the color version of this figure.

close attention to facial features might result in similar improvements. This experiment compared performance when participants received feedback versus no feedback, but did not examine other strategies to enhance attention. Thus, we cannot be certain that feedback alone is the critical component to the efficacy of this training paradigm. Future research can go on to test these specific mechanisms.

Applications

Because this training focused on improving continuous intensity ratings, it has the potential for broad application beyond extant studies which tend to focus on labeling traditional categories of emotion. Studies that rely heavily on emotion labels and categories fall prey to cultural constraints, using concepts and words that are bound to a specific culture or language. There is evidence that there is some variance in how societies recognize and label emotions (Barrett et al., 2019; Crivelli, Jarillo, Russell, & Fernández-Dols, 2016; Crivelli, Russell, Jarillo, & Fernández-Dols, 2016, 2017; Tracy & Robins, 2008). However, despite differences in emotion categorization, ratings of emotion intensity may be more consistent (Crivelli et al., 2017). Further, myriad factors can influence emotion expressivity including individual (Friedman, DiMatteo, & Taranta, 1980; Kring & Gordon, 1998) and cultural differences (Niedenthal, Rychlowska, & Wood, 2017). Successful social perceivers must attend to and update responses according to these differences (Girard & McDuff, 2017; Rychlowska et al., 2015; Wood, Rychlowska, & Niedenthal, 2016). The present training paradigm, therefore, may help social agents adapt to newly encountered norms of emotion expressivity.

Limitations and Future Directions

The present research assessed a novel approach to a training intervention; however, in this early stage of development, there are limitations to the current studies that can be addressed in future research. First, with the exception of Study 1b, the research presented here was preexperimental in design, conducted without comparison or control groups. While training effects were observed in each study, the control group in Study 1b also showed reductions in error, though not to the same degree as the treatment group. The observed reductions in error across these eight studies show promise in this novel training paradigm and warrant further research. Second, the facial stimuli used in this study consisted of images morphed from one prototype of an emotion to another, resulting in an artificial approximation of ecologically valid emotional displays. While there is adequate extant research to support the utility of morphed facial stimuli as a window into understanding emotion perception (Gao & Maurer, 2009; Herba, Landau, Russell, Ecker, & Phillips, 2006; Penton-Voak et al., 2013; Stoddard et al., 2016), subsequent research can extend these hypothesized mechanisms using real-world variations in emotion intensity. This issue is cogently discussed in Gao and Maurer (2009). Additionally, by using static, established images of facial configurations, we had an objective measure of emotion intensity with which to calibrate participant responses. Nevertheless, the stimuli used present limitations for generalizability of the results to real-world emotion perception and interpretation. Taking this research to live, dynamic social interactions could increase the ecological

validity of the effects of training. Additionally, increasing the number of models in which participants must discern the intensity of cues might clarify how the effects of training translate to social settings in which social partners vary in expressivity. As individuals are able to maintain and update emotional information for multiple social agents simultaneously (Plate, Wood, Woodard, & Pollak, 2019), we would expect that the training could translate to more complex social situations.

Another limitation is that the stimulus set presented here was relatively narrow, consisting of four White models. While we found generalizability to other models in this study, we did not test generalizability to identities outside of this stimulus set, such as individuals of different ages or racial and ethnic background. The decision to eliminate ethnicity as a factor in stimulus selection was because we would not have had the sample size and power to fully cross ethnicity in stimuli with participant ethnicity in this initial phase of research. Given the strength of the present results, such an extension is now both motivated and important for exploring individual differences in the effects of this paradigm. Relatedly, the participants included in this study were predominantly White university students. While this sample was representative of their local community, it is relatively homogenous and limits our ability to interpret our findings across demographic groups. Given the unique nature and problems of making generalizations from American undergraduates (Henrich, Heine, & Norenzayan, 2010; Rad, Martingano, & Ginges, 2018), caution is necessary when interpreting the results from the current young adult population. However, the changes in accuracy seen in this study set the stage for testing whether this paradigm may be effective for different populations. An important avenue for future research is to ask whether the effects equivalently generalize across model gender, race, age, and other demographic factors. It is also worth noting that the average age of participants in Study 4 was older than three of the other studies. While the patterns across studies were similar and there is no reason to believe this age difference would result in any meaningful difference in performance in this paradigm, caution is still warranted when interpreting these results.

In addition to expanding characteristics of the sample for exploring generalizability, the primary aim of developing attention training paradigms is to have intervention options to serve clinical groups with attention biases. There is significant variability in inferring others' emotions, particularly when viewing morphed facial configurations (Barrett & Niedenthal, 2004), with some individuals more skilled than others. Clinical groups in particular may have aberrations in attention to, and identification of, emotion, as has been identified in a host of developmental and emotional problems including depression, (Dalili, Penton-Voak, Harmer, & Munafò, 2015; Harrison & Gibb, 2015), autism, (Uljarevic & Hamilton, 2013), schizophrenia (Green & Horan, 2010), anxiety (Lichtenstein-Vidne et al., 2017; Pergamin-Hight, Naim, Bakermans-Kranenburg, van Ijzendoorn, & Bar-Haim, 2015), addiction (Field & Cox, 2008), eating disorders (Hendrikse et al., 2015), aggression (Penton-Voak et al., 2013), and irritability (Hommer et al., 2014); all of which may benefit from novel approaches to treatment targeting attentional mechanisms. Knowledge would be furthered by the examination of variability in abilities in perceiving or inferring emotional intensity, particularly among clinical groups, and whether this training paradigm might

be effective at altering such attentional processes in high-risk populations.

One critical avenue for future research is whether recognizing changes in these stimuli signifies an ability to infer changes in a social agent's internal emotion state versus the ability to solely track perceptual changes in facial musculature. Research examining how individuals use information gleaned from facial cues to predict internal states and behaviors of social partners would further our understanding of the role of this paradigm in facilitating socioemotional communication.

Conclusion

The current research found that training individuals to attend to distinctions in the way different intensities of emotion are conveyed in the face was effective at improving peoples' perceptions or inferences. This is the first emotion training paradigm that we are aware of that focuses on improving ratings of emotion intensity as opposed to categorical recognition of specific emotions. Training effects from this paradigm have the capacity to improve emotion recognition in the service of adaptive interpersonal behavior.

References

- Barrett, L. F. (2013). Psychological construction: The Darwinian approach to the science of emotion. *Emotion Review*, 5, 379–389. <http://dx.doi.org/10.1177/1754073913489753>
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20, 1–68. <http://dx.doi.org/10.1177/1529100619832930>
- Barrett, L. F., & Niedenthal, P. M. (2004). Valence focus and the perception of facial affect. *Emotion*, 4, 266–274. <http://dx.doi.org/10.1037/1528-3542.4.3.266>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48. <http://dx.doi.org/10.18637/jss.v067.i01>
- Beevers, C. G., Clasen, P. C., Enock, P. M., & Schnyer, D. M. (2015). Attention bias modification for major depressive disorder: Effects on attention bias, resting state connectivity, and symptom change. *Journal of Abnormal Psychology*, 124, 463–475. <http://dx.doi.org/10.1037/abn0000049>
- Calvo, M. G., & Nummenmaa, L. (2016). Perceptual and affective mechanisms in facial expression recognition: An integrative review. *Cognition and Emotion*, 30, 1081–1106. <http://dx.doi.org/10.1080/02699931.2015.1049124>
- Cisler, J. M., & Koster, E. H. W. (2010). Mechanisms of attentional biases towards threat in anxiety disorders: An integrative review. *Clinical Psychology Review*, 30, 203–216. <http://dx.doi.org/10.1016/j.cpr.2009.11.003>
- Crick, N. R., & Dodge, K. A. (1996). Social information-processing mechanisms in reactive and proactive aggression. *Child Development*, 67, 993–1002. <http://dx.doi.org/10.2307/1131875>
- Crivelli, C., Jarillo, S., Russell, J. A., & Fernández-Dols, J. M. (2016). Reading emotions from faces in two indigenous societies. *Journal of Experimental Psychology: General*, 145, 830–843. <http://dx.doi.org/10.1037/xge0000172>
- Crivelli, C., Russell, J. A., Jarillo, S., & Fernández-Dols, J. M. (2016). The fear gasping face as a threat display in a Melanesian society. *Proceedings of the National Academy of Sciences of the United States of America*, 113, 12403–12407. <http://dx.doi.org/10.1073/pnas.1611622113>
- Crivelli, C., Russell, J. A., Jarillo, S., & Fernández-Dols, J. M. (2017). Recognizing spontaneous facial expressions of emotion in a small-scale society of Papua New Guinea. *Emotion*, 17, 337–347. <http://dx.doi.org/10.1037/emo0000236>
- Dalili, M. N., Penton-Voak, I. S., Harmer, C. J., & Munafò, M. R. (2015). Meta-analysis of emotion recognition deficits in major depressive disorder. *Psychological Medicine*, 45, 1135–1144. <http://dx.doi.org/10.1017/S0033291714002591>
- Elfenbein, H. A. (2006). Learning in emotion judgments: Training and the cross-cultural understanding of facial expressions. *Journal of Nonverbal Behavior*, 30, 21–36. <http://dx.doi.org/10.1007/s10919-005-0002-y>
- Engelmann, J. B., & Pessoa, L. (2007). Motivation sharpens exogenous spatial attention. *Emotion*, 7, 668–674. <http://dx.doi.org/10.1037/1528-3542.7.3.668>
- Everaert, J., Mogoşe, C., David, D., & Koster, E. H. W. (2015). Attention bias modification via single-session dot-probe training: Failures to replicate. *Journal of Behavior Therapy and Experimental Psychiatry*, 49, 5–12. <http://dx.doi.org/10.1016/j.jbtep.2014.10.011>
- Field, M., & Cox, W. M. (2008). Attentional bias in addictive behaviors: A review of its development, causes, and consequences. *Drug and Alcohol Dependence*, 97, 1–20. <http://dx.doi.org/10.1016/j.drugalcdep.2008.03.030>
- Friedman, H. S., DiMatteo, M. R., & Taranta, A. (1980). A study of the relationship between individual differences in nonverbal expressiveness and factors of personality and social interaction. *Journal of Research in Personality*, 14, 351–364. [http://dx.doi.org/10.1016/0092-6566\(80\)90018-5](http://dx.doi.org/10.1016/0092-6566(80)90018-5)
- Gao, X., & Maurer, D. (2009). Influence of intensity on children's sensitivity to happy, sad, and fearful facial expressions. *Journal of Experimental Child Psychology*, 102, 503–521. <http://dx.doi.org/10.1016/j.jecp.2008.11.002>
- Gendron, M., Roberson, D., van der Vyver, J. M., & Feldman Barrett, L. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, 14, 251–262. <http://dx.doi.org/10.1037/a0036052>
- Girard, J. M., & McDuff, D. (2017). Historical heterogeneity predicts smiling: Evidence from large-scale observational analyses. *Automatic face & gesture recognition (FG 2017)*, 2017 12th IEEE International Conference on automatic face and gesture recognition (pp. 719–726). <http://dx.doi.org/10.1109/FG.2017.135>
- Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, 10, 535–549. <http://dx.doi.org/10.1111/spc3.12267>
- Green, M. F., & Horan, W. P. (2010). Social cognition in schizophrenia. *Current Directions in Psychological Science*, 19, 243–248. <http://dx.doi.org/10.1177/0963721410377600>
- Harrison, A. J., & Gibb, B. E. (2015). Attentional biases in currently depressed children: An eye-tracking study of biases in sustained attention to emotional stimuli. *Journal of Clinical Child and Adolescent Psychology*, 44, 1008–1014. <http://dx.doi.org/10.1080/15374416.2014.930688>
- Hendrikse, J. J., Cachia, R. L., Kothe, E. J., McPhie, S., Skouteris, H., & Hayden, M. J. (2015). Attentional biases for food cues in overweight and individuals with obesity: A systematic review of the literature. *Obesity Reviews*, 16, 424–432. <http://dx.doi.org/10.1111/obr.12265>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466, 29. <http://dx.doi.org/10.1038/466029a>
- Herba, C. M., Landau, S., Russell, T., Ecker, C., & Phillips, M. L. (2006). The development of emotion-processing in children: Effects of age, emotion, and intensity. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 47, 1098–1106. <http://dx.doi.org/10.1111/j.1469-7610.2006.01652.x>

- Hommer, R. E., Meyer, A., Stoddard, J., Connolly, M. E., Mogg, K., Bradley, B. P., . . . Brotman, M. A. (2014). Attention bias to threat faces in severe mood dysregulation. *Depression and Anxiety, 31*, 559–565. <http://dx.doi.org/10.1002/da.22145>
- Kring, A. M., & Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology, 74*, 686–703. <http://dx.doi.org/10.1037/0022-3514.74.3.686>
- Leibenluft, E., & Stoddard, J. (2013). The developmental psychopathology of irritability. *Development and Psychopathology, 25*, 1473–1487. <http://dx.doi.org/10.1017/S0954579413000722>
- Lemerise, E. A., & Arsenio, W. F. (2000). An integrated model of emotion processes and cognition in social information processing. *Child Development, 71*, 107–118. <http://dx.doi.org/10.1111/1467-8624.00124>
- Lichtenstein-Vidne, L., Okon-Singer, H., Cohen, N., Todder, D., Aue, T., Nemets, B., & Henik, A. (2017). Attentional bias in clinical depression and anxiety: The impact of emotional and non-emotional distracting information. *Biological Psychology, 122*, 4–12. <http://dx.doi.org/10.1016/j.biopsycho.2016.07.012>
- Linetzky, M., Pergamin-Hight, L., Pine, D. S., & Bar-Haim, Y. (2015). Quantitative evaluation of the clinical efficacy of attention bias modification treatment for anxiety disorders. *Depression and Anxiety, 32*, 383–391. <http://dx.doi.org/10.1002/da.22344>
- MacLeod, C., & Mathews, A. (2012). Cognitive bias modification approaches to anxiety. *Annual Review of Clinical Psychology, 8*, 189–217. <http://dx.doi.org/10.1146/annurev-clinpsy-032511-143052>
- MacLeod, C., Rutherford, E., Campbell, L., Ebsworthy, G., & Holker, L. (2002). Selective attention and emotional vulnerability: Assessing the causal basis of their association through the experimental manipulation of attentional bias. *Journal of Abnormal Psychology, 111*, 107–123. <http://dx.doi.org/10.1037/0021-843X.111.1.107>
- Martinez, A. M. (2017). Computational models of face perception. *Current Directions in Psychological Science, 26*, 263–269. <http://dx.doi.org/10.1177/0963721417698535>
- Mogoșe, C., David, D., & Koster, E. H. W. (2014). Clinical efficacy of attentional bias modification procedures: An updated meta-analysis. *Journal of Clinical Psychology, 70*, 1133–1157. <http://dx.doi.org/10.1002/jclp.22081>
- Niedenthal, P. M., Rychlowska, M., & Wood, A. (2017). Feelings and contexts: Socioecological influences on the nonverbal expression of emotion. *Current Opinion in Psychology, 17*, 170–175. <http://dx.doi.org/10.1016/j.copsyc.2017.07.025>
- Penton-Voak, I. S., Bate, H., Lewis, G., & Munafò, M. R. (2012). Effects of emotion perception training on mood in undergraduate students: Randomised controlled trial. *The British Journal of Psychiatry, 201*, 71–72. <http://dx.doi.org/10.1192/bjp.bp.111.107086>
- Penton-Voak, I. S., Thomas, J., Gage, S. H., McMullan, M., McDonald, S., & Munafò, M. R. (2013). Increasing recognition of happiness in ambiguous facial expressions reduces anger and aggressive behavior. *Psychological Science, 24*, 688–697. <http://dx.doi.org/10.1177/0956797612459657>
- Pergamin-Hight, L., Naim, R., Bakermans-Kranenburg, M. J., van Ijzendoorn, M. H., & Bar-Haim, Y. (2015). Content specificity of attention bias to threat in anxiety disorders: A meta-analysis. *Clinical Psychology Review, 35*, 10–18. <http://dx.doi.org/10.1016/j.cpr.2014.10.005>
- Plate, R. C., Wood, A., Woodard, K., & Pollak, S. D. (2019). Probabilistic learning of emotion categories. *Journal of Experimental Psychology: General, 148*, 1814–1827.
- Pollak, S. D., Messner, M., Kistler, D. J., & Cohn, J. F. (2009). Development of perceptual expertise in emotion recognition. *Cognition, 110*, 242–247. <http://dx.doi.org/10.1016/j.cognition.2008.10.010>
- Pollak, S. D., & Sinha, P. (2002). Effects of early experience on children's recognition of facial displays of emotion. *Developmental Psychology, 38*, 784–791. <http://dx.doi.org/10.1037/0012-1649.38.5.784>
- R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of *Homo sapiens*: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 11401–11405. <http://dx.doi.org/10.1073/pnas.1721165115>
- Rychlowska, M., Miyamoto, Y., Matsumoto, D., Hess, U., Gilboa-Schechtman, E., Kamble, S., . . . Niedenthal, P. M. (2015). Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles. *Proceedings of the National Academy of Sciences of the United States of America, 112*, E2429–E2436. <http://dx.doi.org/10.1073/pnas.1413661112>
- Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality, 19*, 595–605. <http://dx.doi.org/10.1002/per.554>
- Shechner, T., & Bar-Haim, Y. (2016). Threat monitoring and attention-bias modification in anxiety and stress-related disorders. *Current Directions in Psychological Science, 25*, 431–437. <http://dx.doi.org/10.1177/0963721416664341>
- Stoddard, J., Sharif-Askary, B., Harkins, E. A., Frank, H. R., Brotman, M. A., Penton-Voak, I. S., . . . Leibenluft, E. (2016). An open pilot study of training hostile interpretation bias to treat disruptive mood dysregulation disorder. *Journal of Child and Adolescent Psychopharmacology, 26*, 49–57. <http://dx.doi.org/10.1089/cap.2015.0100>
- Tottenham, N., Tanaka, J. W., Leon, A. C., McCarry, T., Nurse, M., Hare, T. A., . . . Nelson, C. (2009). The NimStim set of facial expressions: Judgments from untrained research participants. *Psychiatry Research, 168*, 242–249. <http://dx.doi.org/10.1016/j.psychres.2008.05.006>
- Tracy, J. L., & Robins, R. W. (2008). The nonverbal expression of pride: Evidence for cross-cultural recognition. *Journal of Personality and Social Psychology, 94*, 516–530. <http://dx.doi.org/10.1037/0022-3514.94.3.516>
- Uljarevic, M., & Hamilton, A. (2013). Recognition of emotions in autism: A formal meta-analysis. *Journal of Autism and Developmental Disorders, 43*, 1517–1526. <http://dx.doi.org/10.1007/s10803-012-1695-5>
- Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin, 134*, 383–403. <http://dx.doi.org/10.1037/0033-2909.134.3.383>
- Van Bockstaele, B., Verschuere, B., Tibboel, H., De Houwer, J., Crombez, G., & Koster, E. H. W. (2014). A review of current evidence for the causal impact of attentional bias on fear and anxiety. *Psychological Bulletin, 140*, 682–721. <http://dx.doi.org/10.1037/a0034834>
- Wood, A., Rychlowska, M., & Niedenthal, P. M. (2016). Heterogeneity of long-history migration predicts emotion recognition accuracy. *Emotion, 16*, 413–420. <http://dx.doi.org/10.1037/emo0000137>

Received April 12, 2018

Revision received January 29, 2020

Accepted February 22, 2020 ■