Emotion
2011, Vol. 11, No. 4, 896–906

1528-3542/11/$12.00 DOI: 10.1037/a0023746

Metacognition of Emotional Face Recognition

Karen J. Kelly and Janet Metcalfe
Columbia University

While humans are adept at recognizing emotional states conveyed by facial expressions, the current literature suggests that they lack accurate metacognitions about their performance in this domain. This finding comes from global trait-based questionnaires that assess the extent to which an individual perceives him or herself as empathic, as compared to other people. Those who rate themselves as empathically accurate are no better than others at recognizing emotions. Metacognition of emotion recognition can also be assessed using relative measures that evaluate how well a person thinks s/he has understood the emotion in a particular facial display as compared to other displays. While this is the most common method of metacognitive assessment of people's judgments of learning or their feelings of knowing, this kind of metacognition—"relative meta-accuracy"—has not been studied within the domain of emotion. As well as asking for global metacognitive judgments, we asked people to provide relative, trial-by-trial prospective and retrospective judgments concerning whether they would be right or wrong in recognizing the expressions conveyed in particular facial displays. Our question was: Do people know when they will be correct in knowing what expression is conveyed, and do they know when they do not know? Although we, like others, found that global meta-accuracy was unpredictive of performance, relative meta-accuracy, given by the correlation between participants' trial-by-trial metacognitive judgments and performance on each item, were highly accurate both on the Mind in the Eyes task (Experiment 1) and on the Ekman Emotional Expression Multimorph task (in Experiment 2).

Keywords: metacognition, emotion, relative, global, meta-accuracy

This article addresses the question of whether people know when they understand others' emotional expressions. Metacognition—knowing when one knows and when one does not know—is a skill that is both separate from the more basic level emotion recognition itself, and that has the potential to confer a number of benefits. If one is accurate in knowing that one knows for certain what another person is feeling, then it is not necessary to devote time and effort to further determining what the other person's emotion is. One can act, or choose not to act, on one's knowledge of the other person's emotional state directly. In contrast, there is a distinct value to being able to correctly assess that one does *not* know what the other person's emotional expression is conveying. Rather than acting precipitously, an individual—with good metacognitions—can, when they know that they don't know, take the time to further explore the person's expression or directly ask the person what he or she is feeling, rather than jumping to a wrong conclusion that could have negative consequences for the relationship. Appropriate uncertainty about one's own lack of understanding can be valuable. Furthermore, one of the well-known benefits of accurate metacognition, in any domain, is that it can be used to promote learning. The individual can choose to allocate attention and time appropriately (see, Finn, 2008; Metcalfe & Finn, 2008; Thiede, Anderson, & Therriault, 2003), and learn to hone their emotion recognition skills effectively. If an individual does not know what they don't know (i.e., they have poor metacognitions), they are not in a position to remedy the situation (Benjamin, Bjork, & Schwartz, 1998). As in other domains, then, meta-accuracy in the domain of emotional recognition should be helpful in promoting task accuracy, minimizing error, ensuring flexibility in a dynamic social context, and in promoting learning over time (Nelson & Narens, 1990).

Metacognition is broadly defined as the capacity to actively monitor and reflect upon one's own performance and abilities (Flavell, 1979; Dunlosky & Metcalfe, 2009). It can be measured in several ways, however, and these methods of assessing metacognition address different questions and allow different inferences. The *global meta-accuracy* and *relative meta-accuracy* of one's judgments about one's knowledge are the two most important and prominent measures used in the field of metacognition. Global measures usually use a Pearson correlation to calculate the relationship between people's responses on questionnaires about their abilities, preferences, predispositions, or, in this case, emotional sensitivity, and their ultimate overall performance on a relevant task, as compared to that of other people. They require the participant to assess his or her performance in a domain, in comparison with other members of the sample, and also to provide an objective measure—usually by taking a test—of performance on the task. Then the participants' overall metacognitive judgments about their

Karen J. Kelly and Janet Metcalfe, Department of Psychology, Columbia University.

This research was supported, in part, by National Institute of Mental Health Grant RO1-MH60637 and by Grant 220020166 from the James S. McDonnell Foundation. The authors, however, are wholly responsible for the content of this article. We thank Natalie Sebanz, Bridgid Finn, Jamil Zaki, Patrick Kennedy, and David Miele.

Correspondence concerning this article should be addressed to Karen J. Kelly, Department of Psychology, Columbia University, 406 Schermerhorn Hall, 1190 Amsterdam Ave MC 5501, New York, NY 10027. E-mail: kk2557@columbia.edu

skills or propensities as compared to others are compared to the objective assessments of how they performed on the criterion task as compared to others. For example, a questionnaire might ask how good a student a person thinks he or she is, and then relate that rating to GPA, plotting this correspondence across a sample of people. If people who rate themselves as good students, are, in fact, the students with high GPAs, then this global meta-accuracy correlation will be high: people who think they are good students are good students.

In contrast to these global measures, relative measures look at the metacognition within the individual and provide information on whether the person knows what he or she knows and does not know. Such measures require the individual to discriminate among items on a task by providing judgments about whether they will get particular items right or not on an upcoming test. Then the participant is given the test, and the predictions that were made about the particular items are correlated with whether those items were right or wrong on the test, usually using a gamma correlation (see Nelson & Narens, 1990, for discussion of this statistic; cf., Masson & Rotello, 2008). A gamma correlation is simply a nonparametric correlation coefficient that relates the individual's trial-by-trial judgments to whether they were right or wrong on each item on the criterion test. A correlation of 1.0 means that the person thought they would be correct on those items that he or she got right, and thought they would be wrong on those items they got wrong. A correlation of zero means that the person's judgments were unrelated to what they got right or wrong on the test. A correlation of −1.0 means that the person thought he or she would get right the answers he or she got wrong, and the wrong answers he or she got right. (One rarely sees this pattern in real data). If the sample mean over people—where gamma is the dependent measure—is significantly above zero, then one can conclude that people have good metacognition, meaning that, on average, they know what they know and don't know. While relative meta-accuracy is most commonly used in the metacognitive literature on learning and memory (Dunlosky & Metcalfe, 2009) both of these measures—those measuring global meta-accuracy and those measuring relative meta-accuracy—are commonly, and correctly, called metacognition. Even so, their meaning is different. The former assesses whether people know how good they are with respect to other people. The latter measures whether or not people know what they know.

Within the context of emotion recognition and the ultimate goal of social and interpersonal connectivity and understanding, there is little doubt that metacognition would seem to be a valuable skill. The application of metaknowledge to future social-emotional encounters should result in the opportunity to learn to correct one's emotional errors and promote the future likelihood of an accurate situational response while preserving and nurturing the relevant relationship. Accurate metacognitions, at a relative meta-accuracy level, should promote correct and confident interpretations: an angry expression, confidently understood as such, should elicit an immediate and appropriate response. In contrast, uncertainty in the interpretation of an ambiguous facial expression may prompt the perceiver to seek additional information until a confident assessment can be made, maximizing the likelihood of an adaptive response. Emotional displays signal important information about an individual's internal states (Darwin, 1872), essential for smooth social functioning and ultimately for survival. Insofar as most individuals can accurately recognize nonverbal emotional expressions (Adolphs, 2006; Ekman, 1982; Tracy & Robins, 2008), and insofar as metacognition is thought to have a central role in acquiring and maintaining this skill, one would expect the literature to show that people have good metacognition of emotion.

A number of studies have examined people's metacognitive judgments concerning emotion recognition. Surprisingly, the consensus, as exhaustively reviewed below[1] and as noted by Ickes (l993), is that people appear to almost entirely lack metaknowledge in the domain of emotion recognition. For example, Ickes, Stinson, Bissonnette, and Garcia (1990) designed a study to assess empathic accuracy in naturalistic dyads. The researchers recorded two participants interacting and later asked the participants to review the tapes and note what they and their partner had been feeling. Following this, the participants were asked to complete a self-report questionnaire including personality and self-monitoring measures. While the researchers demonstrated that the participants varied in their ability to recognize the emotions exchanged during the unstructured dyadic interaction, the global metacognitive measures failed to predict performance on this emotion recognition task. Individuals who reported high levels of empathy, skill at deciphering others' intentions, and emotional intelligence were no more accurate at identifying emotions than those who reported less competence.

Similarly, Levenson and Ruef (1992) adopted a related paradigm to measure empathic accuracy and the effect of physiological synchrony between the target and the perceiver. Participants viewed a dyadic marital interaction and rated the emotional intensity of the observed target person. These ratings were compared with the target's self-ratings. The accuracy of detecting negative emotions improved when the target and perceiver were synchronized in their autonomic and somatic responses. However, the traditional global empathy scales and the participants' ratings of self-perceived task accuracy and task difficulty did not predict task performance.

In keeping with this line of research Marangoni, Garcia, Ickes, and Teng (1995) used the dyadic interaction paradigm to assess empathic accuracy in simulated psychotherapy sessions. They were interested in individual differences in empathy and whether empathic accuracy improved with increased exposure to the target and with feedback concerning the target's feelings. Each participant watched three videotaped interactions between a target and a clinician. The videotapes were paused 30 times each while the participant evaluated the target's emotional state. At posttest, the participants were asked to provide a global judgment of their overall accuracy for each target. The authors found individual differences and general improvement in empathic accuracy following increased exposure to the target and feedback. However,

[1] There is one additional study that might be relevant that is not included in our literature review—an unpublished 1996 master's thesis by Mortimer (cited in Ickes & Aronson, 2003), in which people were shown video segments and asked if they would be able to answer questions about these segments later. They were then asked inferential questions such as "Who is the child of the two adults?" or "What is the relationship between the man and the woman?" Participants' level of metacognition regarding their trial-by-trial empathic accuracy was measured using a Pearson correlation. Mortimer found that "the correlations were large enough to be statistically meaningful for only 8 of the 72 perceivers—about 11% of them."

they found no relation between self-assessment and task accuracy. They concluded that while task accuracy was improved, the meta-accuracy was not improved.

Realo et al. (2003) created the Mind-Reading Belief Scale to more specifically assess metacognition about one's own emotion recognition ability. The scale included items designed to assess one's ability to judge another's traits, mind states, intentions, and emotions. The questionnaire requires the participant to use a 5-point Likert scale to respond to questions such as, "I can read people's intentions in their faces." While the scale was psychometrically sound, self-reported empathy on the Mind-Reading Belief Scale failed to predict performance on a face or voice emotional expression recognition task.

Ames and Kammrath (2004) asked participants to complete a number of questionnaires to assess individual differences in extraversion, narcissism, self-esteem, and social skills. Participants completed the Interpersonal Perception Task (IPT) (Archer & Costanzo, 1988), in which one must answer questions about video clips that depict various interpersonal interactions. In keeping with previous research, there was no relation between questionnaire responses and performance on the IPT. Those who reported higher levels of social fluency were no more accurate at interpreting the emotional content of the videos. After completing the IPT, participants were paired together and asked to perform a simulated negotiation task in which they adopted either the role of an entrepreneur selling a family business or that of an executive seeking to purchase the business. The negotiation task was followed by a report of one's own and one's partner's intentions and feelings during the negotiation task. Again, those who reported higher levels of social fluency were no better at deciphering their partner's intentions and feelings during the negotiation. The authors did, however, find a strong positive relationship between an individual's predicted task performance and self-reported levels of narcissism. The finding that the more narcissistic the individual the more likely s/he was to predict but not exhibit good performance on emotion recognition tasks renders the utility of these self-reported global metacognitive measures questionable.

Recent work has attempted to discern in more detail the conditions, if any, under which global metacognitive judgments might be predictive. Zaki, Bolger, and Ochsner (2008), using video clips of emotional expressions concerning personal episodes, did find a situation in which self-reported empathic accuracy predicted people's emotional accuracy. This positive relation, showing that people who thought they were highly sensitive actually performed better, occurred only when the most highly emotionally expressive clips were used. This is the only case in the published literature in which people have shown any above-chance metacognitive accuracy on emotion recognition, but even this occurred with only a subset of the materials used. Despite this single provocative finding, the consensus remains that people lack metacognition about emotional expression understanding.

According to this review, Ickes' (1993) early summary of the evidence that individuals lack metacognitive insight into emotion recognition—that self-assessment measures show no predictive value with respect to emotion recognition—seems to still hold today. The self-assessment questionnaires, while unable to predict performance on various emotion recognition tasks, do sometimes relate to levels of narcissism, however, rendering their interpretation particularly dubious. Finally, some questionnaires have even demonstrated a negative correlation between perceived and actual ability (Ickes et al., 1990).

Ickes (1993) suggested a number of reasons why individuals may lack metaknowledge in this realm. He argued that while individuals differ in their ability to identify emotions, people generally have very little insight into this ability because of privacy norms. As a result of these norms, individuals rarely solicit or receive accurate feedback concerning mistakes in emotional interpretation. Further, individuals may intentionally mask their emotions or intentionally provide misleading feedback that makes it difficult to assess one's accuracy.

While the conclusion suggested by these studies seems straightforward, a complete absence of metaknowledge, nevertheless, seems incompatible with the general notion that humans are emotionally intelligent and socially adapted. Humphrey (1984) compared the complexity of social interaction to a chess game in which individuals must maintain and monitor their own goals while continually monitoring others' behaviors, building alliances, deceiving, cooperating, strategizing, and manipulating. Successfully navigating such a socially complex world seems unlikely, absent any metacognitive insight. Adolphs (2006) argued that one's perception of emotion is active and inferential, in that one actively seeks out relevant cues and infers meaning based on observations and past experiences. This, too, suggests that individuals monitor their knowledge state, implementing a very basic metacognitive strategy in social interactions. People's obvious skill in negotiating the social world and their apparent lack of metacognition, reflected in the results reviewed above, seem irreconcilable.

Before accepting the conclusion that people have such extremely poor metacognition in the domain of emotional recognition, it is important to note that none of the above-cited studies on emotional metacognition used the relative measures of metacognition discussed earlier. All of the studies described above used questionnaires to assess interpersonal sensitivity based on the individual's tendency to agree with statements such as, "I am generally sensitive to others' feelings"—a global measure. However, as noted earlier, global meta-accuracy is only one sort of metacognitive accuracy. It measures whether people know where they stand with respect to one another; it does not measure whether or not they know what they know. Without evidence concerning whether people's relative meta-accuracy is above chance or not, it could be premature to conclude that people actually have poor emotional metacognitive capabilities.

The present study investigated whether individuals can accurately predict and assess their performance on two distinct emotion recognition tasks. We explored both relative and global measures of meta-accuracy and the relation between them. We further investigated the sensitivity of relative measures across differing levels of emotional content by including stimuli that ranged from relatively neutral to the more dramatically obvious expressions. In Experiment 1 we investigated people's metacognition of emotion recognition using the Mind in the Eyes task—a task that assesses one's ability to determine another's mental state by viewing just the eyes (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). In Experiment 2 we used the Emotional Expression Multimorph Task (Blair, Colledge, Murray, & Mitchell, 2001) that presents an individual's full face in differing stages of transition from neutral to fully expressed emotions.

Experiment 1

In Experiment 1, participants completed a global metacognitive questionnaire and also performed a relative metacognitive task. The global questionnaire assessed self-perceived levels of empathy, perspective taking, and theory of mind. To evaluate relative meta-accuracy, we then presented stimuli from the Mind in the Eyes test in which individuals viewed sets of eyes and selected the expressed mental state from various response options (Baron-Cohen et al., 2001). During the presentation of each item, participants provided a prospective judgment of whether they would be able to recognize the correct emotion. After making judgments on all items, they then answered each question. After answering the question each participant then made a retrospective judgment of whether they had answered correctly.

We expected to replicate the finding that global metacognitions were unpredictive of performance. We also predicted, however, that relative meta-accuracy would be good and that individuals would be able to make highly accurate item-by-item estimates of their performance. We thought that both the prospective and retrospective relative metacognition judgments would show above-zero meta-accuracy. Usually, though, retrospective metacognitive judgments show higher relative accuracy than do prospective judgments because people, when making prospective judgments, do not know what the alternatives will be (see Glenberg, Sanocki, Epstein, & Morris, 1987). Thus, we expected that while both prospective and retrospective relative meta-accuracy judgments would be above zero, we hypothesized that the retrospective relative meta-accuracy would be higher than the prospective relative meta-accuracy.

Method

Participants

Ninety-nine Columbia University students (44 males, 51 females, 4 unknown, M_{age} = 21.94 years, SD_{age} = 4.69) participated for course credit or cash. All participants in both experiments were treated in accordance with APA ethical guidelines.

Materials

The questionnaire that assessed global meta-accuracy contained 24 items selected from the Mind-Reading Belief Scale (Realo et al., 2003), and the Interpersonal Reactivity Index–Perspective-Taking (IRI-P) and Empathic-Concern subscales (IRI-E) (Davis, 1980). The items represented abilities related to understanding another's mental states. Sample items from each included: "I can read people's intentions in their faces" (Mind-Reading Belief Scale), "I sometimes try to understand my friends better by imagining how things look from their perspective" (IRI-P), and "I am often quite touched by things that I see happen" (IRI-E). We also included 2 task-specific questions, "Relative to other Columbia students I am able to determine what a person is feeling by looking into their eyes" and "Relative to other Columbia students I am *not* able to determine what a person is feeling by looking into their eyes." We termed this the Columbia Empathy Measure (CEM). For our combined samples, the internal consistencies were measured using Cronbach's alpha and were: Mind-Reading Belief Scale α = .70, IRI-P α = .78, IRI-E α = .80, CEM α = .76. All responses were made on a 5-point Likert scale with higher scores indicating that the statement was extremely characteristic of them.

The stimuli for Experiment 1 were derived from the Mind in the Eyes task (see Figure 1 for an example) in which individuals view 36 sets of eyes and select the expressed mental state from four response options (Baron-Cohen et al., 2001) that, for reasons outlined below, we increased to six options. The stimuli consisted of various individuals posing complex mental states (e.g., contemplative, desirous, and aghast). Each image was presented in black and white and cropped to include only the actor's eye region. To decrease the probability of chance guessing that spuriously decreases gamma scores—making it difficult to detect a real correlation (see Schwartz & Metcalfe, 1994)—we added two distractors to the standard response options. Additional distractors were selected from the complete list of response options. Trials in which the added distractors were chosen more often than chance were eliminated, yielding 25 stimuli.

Procedure

The testing session began with the questionnaire, which was immediately followed by the Mind in the Eyes task. The stimuli were divided into four blocks (nine eyes per block). The displayed emotions and the response options were quite varied and the participants were told they would be given a six-alternative multiple-choice test. Each block contained a prospective judgment phase and a retrospective judgment phase. During the prospective phase, participants viewed each successive set of eyes and after seeing each were asked, "How confident are you that you will be able to correctly identify the emotion?" This confidence level was the prospective judgment. During the retrospective phase, participants viewed the same stimuli from the previous phase now accompanied by the test that included the response options and the instruction to select the expressed mental state. After choosing a response, they were then asked to provide a retrospective confidence judgment expressing the likelihood that the given response was correct. Both prospective and retrospective judgments were made using a slider that was anchored at 0% and 100% at each end. We then converted these percentages to proportions.

Figure 1. Stimuli for Experiment 1 (Mind in the Eyes). These particular images represent the emotions: playful, desire, and worried. Reprinted with permission from Simon Baron-Cohen, 2011.

Results

Performance

Mean accuracy on the Mind in the Eyes task was .73 (SD = .13). Chance performance on the task was .17. A one-sample t test against the value of chance showed that our sample performed well above chance [$t(98) = 41.68$, $p < .00$]. In keeping with the literature, this indicated that participants were good at identifying emotional expressions (Tracy & Robins, 2008).

To determine whether or not prospective and retrospective confidence judgments differed overall, we performed a paired samples t test. Mean prospective confidence judgments (M = .57, SD = .11) were significantly lower than mean retrospective confidence judgments (M = .65, SD = .12) [$t(98) = 11.06$, $p < .001$, d = 1.58]. This is in keeping with our expectations that confidence judgments would change once the response choices were presented, and with the fact that people had seen the stimuli twice by the time they made the retrospective judgments.

Relative Meta-Accuracy

Two gamma correlations were computed for each participant. The first was between prospective confidence ratings on each item and whether each item was correct or incorrect; the second was between the retrospective confidence ratings on each item, and whether the items were correct or incorrect. These gamma correlations were taken as dependent measures for each participant from which the sample means of gammas were computed and compared to the null hypothesis of zero, which would indicate no relative meta-accuracy. Both the prospective and retrospective gamma correlations (G) between confidence judgments and accuracy were significantly greater than zero [prospective G = .07, SD = .32, $t(98) = 2.21$, $p = .03$; retrospective G = .40, SD = .33, $t(98)$ = 12.20, $p < .001$] (see Figure 2). As predicted, retrospective relative meta-accuracy was greater than the prospective relative meta-accuracy [$t(98) = 8.72$, $p < .001$, $d = 1.25$].

Relation of Relative Meta-Accuracy to Performance

Relative meta-accuracy was good, indicating that people do have knowledge about what they know and do not know concerning emotional expressions. We also used the gammas themselves for each participant, to see whether the magnitude of this correlation was, itself, predictive of performance on the Mind in the Eyes task. As noted by Nelson (l984), the gamma correlation (unlike the Pearson correlation) is robust under different levels of task performance. So if a correlation between gammas and performance on the Mind in the Eyes task were found, it would not be spuriously attributed to the mere level of performance, a priori, but rather would indicate that good metacognizers about their own emotion recognition were also good emotion recognizers. However, no relation between the two was found in this experiment. Participants

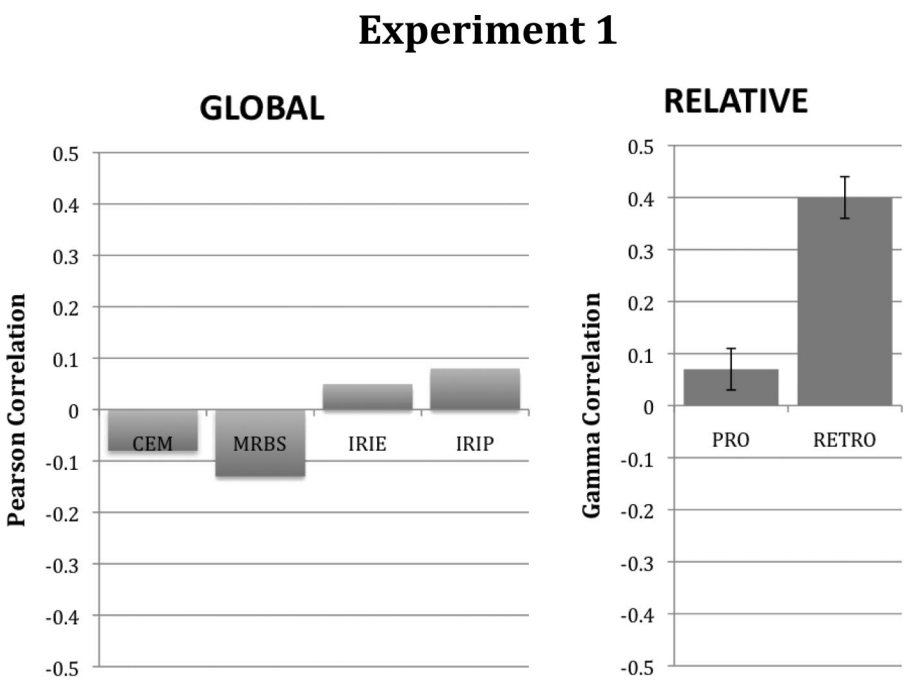

Figure 2. Global (Pearson correlation between questionnaire response and proportion correct on the emotion recognition task) and relative (gamma correlation, means of within participant correlations between confidence ratings and correct vs. incorrect responses on the emotion recognition task) meta-accuracy in Experiment 1 (Mind in the Eyes). The Global questionnaires were the Columbia Emotion Measure (CEM), Mind-Reading Belief Scale (MRBS), the Interpersonal Reactivity Index Empathy (IRI-E) and the Perspective-Taking (IRI-P) subscales. Gamma correlations are given for prospective (pro) and retrospective (retro) confidence judgments as related to item accuracy.

who had particularly high relative meta-accuracy scores did not show particularly high performance on the Mind in the Eyes task, either when prospective gammas were related to proportion correct [$r_{\textit{between individuals' gamma scores and their proportions correct}} = -.01, p = .92$] or when retrospective gammas were related to proportion correct [$r = .08$, $p = .39$].

Global Meta-Accuracy

None of the methods of assessing global meta-accuracy—whether people knew that they were good or bad at doing the emotion recognition task—were related to their actual performance on the Mind in the Eyes task. We computed a Pearson correlation between responses on the questionnaires and accuracy on the task. Higher scores on the questionnaires indicate high self-rated levels of mind-reading ability, empathy, and interpersonal sensitivity. Mind-Reading Belief Scale was not correlated with task performance, $r = -.13$, $p = .21$, *ns*; Interpersonal Reactivity Index (Perspective-Taking subscale) was not correlated with task performance, $r = .08$, $p = .45$, n.s.; Interpersonal Reactivity Index (Empathic Concern subscale) was not correlated with task performance $r = .05$, $p = .63$, n.s.; and Columbia Empathy Measure was not correlated with task performance $r = -.08$, $p = .44$, n.s. These findings replicate the previous literature.

We also investigated whether global self-assessment on any scale was predictive of any measures other than task accuracy. Perhaps not surprisingly, responses on the Mind-Reading Belief Scale (a global measure) were marginally predictive of the overall magnitude of prospective confidence judgments ($r = .19$, $p = .07$), such that there was a trend toward individuals who thought they were proficient at mind-reading being more confident, overall, in their prospective confidence judgments than individuals who reported lower mind-reading competence. This finding harkens back to the earlier reported finding that narcissism is correlated with questionnaire responses and overall confidence.

The means of the prospective confidence judgments for each participant were uncorrelated with any other global metacognitive measure; however, the Interpersonal Reactivity Index (Perspective-Taking subscale) was not correlated with mean confidence judgments, $r = .00$, $p = .98$, n.s.; the Interpersonal Reactivity Index (Empathic Concern subscale) was not correlated with confidence judgments, $r = .11$, $p = .28$, n.s.; the Columbia Empathy Measure was not correlated with confidence judgments, $r = .14$, $p = .16$, n.s. Mean retrospective confidence judgments were also not correlated with any global metacognitive measures: the Mind-Reading Belief Scale was not correlated with confidence judgments, $r = .16$, $p = .11$, n.s.; scores on the Interpersonal Reactivity Index (Perspective-Taking subscale) were not correlated with confidence judgments, $r = -.07$, $p = .50$, n.s.; scores on the Interpersonal Reactivity Index (Empathic Concern subscale) were not correlated with confidence judgments, $r = .02$, $p = .85$, n.s.; the Columbia Empathy Measure was not correlated with confidence judgments, $r = .12$, $p = .25$, n.s. And, none of the above global measures of self-assessed metacognition as measured by the personality scales were correlated with relative meta-accuracy gamma scores (all p's $> .05$).

Discussion

The most important finding of this experiment was that people were able to make accurate relative metacognitive assessments concerning their emotion recognition performance. They correctly discriminated between items that they answered correctly and incorrectly, particularly when the response options had been presented, and they had just made their emotion recognition judgment. This indicates that people do know whether they know or do not know—they have accurate metacognition—in the domain of emotion recognition, in contrast to the conclusion that would have been drawn from previous studies on this topic.

As expected, we also replicated the finding that global questionnaire responses did not predict performance on the task. People who reported interpersonal sensitivity, empathy, and the ability to read and interpret mental states and intentions were no better at identifying the facial expressions as depicted in the Mind in the Eyes than were people who reported that they were less competent.

While these results are intriguing, there are a few limitations of the Mind in the Eyes test that might warrant further exploration with a different task. The stimuli are ambiguous and the response options represent complex mental states that are based on subtle distinctions. In addition, using just the eyes to identify emotions may not be readily comparable to the previous studies of emotion recognition, or to situations that people encounter in their daily lives. Furthermore, in the Mind in the Eyes task, there is no a priori way to distinguish between easy and difficult items, and Zaki et al.'s (2008) data indicate that this could be an interesting variable to investigate. Finally, Glenberg et al. (1987) have noted that accurate metacognitive judgments depend upon the participant knowing the nature of the criterion test. But the test response options were unknown at time of judgment in the Mind in the Eyes task, a factor that, as Glenberg et al. (l987) noted, should have negatively impacted on prospective relative meta-accuracy (which gamma scores, while significantly greater than zero, were small). This lack of knowledge of the test may also have accounted for the observed differences in the prospective and retrospective gamma scores.

In spite of these limitations, individuals did reliably assess their meta-accuracy on this task, suggesting a level of metacognitive insight not previously recognized. Even so, it seemed prudent to replicate with a task that circumvented the difficulties of the Mind in the Eyes test.

Experiment 2

The results of Experiment 1 indicated, for the first time, that people do know what they know emotionally; they have good relative meta-accuracy. In Experiment 2 we were interested in replicating this finding using different and better stimuli. Accordingly, we used the Ekman Emotional Expression Multimorph Task (Blair et al., 2001). This stimulus set uses the Pictures of Facial Affect Series (Ekman & Freisen, 1976) to create a set of stimuli that depicts faces that gradually transition from neutral to fully expressed emotion in 10 stages. The original Ekman picture set contains both males and females expressing the basic emotions that Ekman argued are universally recognizable: happiness, fear, anger, sadness, and disgust. (While surprise is often included, our set did not contain pictures expressing this emotion). The set also

contains each actor posing a neutral expression. Blair et al. (2001) used these pictures to create a series of morphed images that systematically vary the amount of emotional content available in each image. As the images gradually morph from a neutral expression to a fully expressed emotion, more emotional content becomes available and the posed emotion is slowly revealed. The graded emotional expressions allow objective control of emotional content and categorization of stimuli according to difficulty. This is an important feature of these stimuli in so far as measures of relative meta-accuracy are quite sensitive to restricted range. If the images do not represent a broad enough range of difficulty, the ability to detect metacognitive accuracy is hampered, as the participant is not able to reliably differentiate known items from unknown items.

In contrast to the Mind in the Eyes test, which only displayed a fragment of the face, the morphs display the entire face. Presenting the entire face is also more in keeping with the typical way in which facial expressions are processed in more naturalistic circumstances that allow the viewer to use all available facial cues and information (Tracy & Robins, 2008; Adolphs, 2006). This full-face presentation should provide more ecological validity than seen in the previous experiment.

Further, the Ekman morphs limit the response options such that the same basic emotional choices are presented on each trial, controlling for the problems of subtlety and ambiguity associated with the Mind in the Eyes task. This feature also addresses the problem of the unknown nature of the test in the previous experiment. We expected that presenting the same basic emotion response options on each trial would decrease or possibly entirely eliminate the discrepancy between prospective and retrospective confidence judgments found in Experiment 1. We also hypothesized that the relative meta-accuracy gammas for the prospective judgments would be higher, given that the test was more transparent.

We expected to replicate the overall findings from Experiment 1—that global metacognitive judgments would not predict task performance but that relative metacognitive judgments would show that people did have the metaknowledge of what they knew and did not know in the emotional face recognition domain. Further, we expected individuals to be sensitive to the amount of emotional content presented in the faces, adjusting their judgments according to stimulus difficulty such that they would be more likely to say they knew the emotion in the easy, expressive faces, and that they did not know the emotion in the more neutral faces, in which the emotions were more difficult to discern and ambiguous.

Method

Participants

One hundred Columbia University students (38 males; 60 females; 2 unknown, $M_{age} = 21.4$ years, $SD_{age} = 5.36$) participated for course credit or cash.

Materials

We used the previously described questionnaires and scoring—that is the Mind-Reading Belief Scale, the Interpersonal Reactivity Index–Perspective-Taking, the Interpersonal Reactivity Index–Empathic Concern, and the Columbia Empathy Measure—with one small exception. The task-specific question read "Relative to other Columbia students, I am able to determine what a person is feeling by looking at *their face*."

We selected 90 faces from the Emotional Expression Multimorph Task: six actors (3 male; 3 female) posing five emotions (sad, happy, disgust, anger, and fear) (Blair et al., 2001). The morphs were taken from a previously created set consisting of 10 gradations from neutral to fully expressed emotion. We selected the second, fourth, and sixth levels of each emotion to serve as difficult, medium, and easy items, respectively. Each face was presented in black and white on a black backdrop (see Figure 3).

Procedure

Global metacognitive judgments were collected using the previously described questionnaires. Relative item-by-item metacognitive judgments were also obtained according to the procedure outlined in Experiment 1. Participants provided prospective and retrospective confidence judgments of their predicted accuracy on each item. There were six blocks of 15 trials each. Block arrangement and presentation followed the format described in Experiment 1. Each block was divided into two sections: prospective confidence judgments and test accompanied by retrospective confidence judgments.

Participants were told that each model was expressing sadness, happiness, disgust, anger, or fear. Therefore, during prospective confidence judgments, participants knew that the emotion response options would remain constant and that only these options would be presented during the test.

Figure 3. Sample stimuli for Experiment 2 (Emotional Expression Multimorph Task). a) These images represent one actor posing fear at morphed levels 2, 4, and 6. b) These images represent another actor posing happiness at morphed levels 2, 4, and 6. Reprinted from Pictures of Facial Affect, Consulting Psychologists Press, Palo Alto, CA, P. Ekman and W. Friesen, 1976 with permission from Paul Ekman, Ph.D./Paul Ekman Group, LLC, 2011.

Results

Performance

Mean proportion correct on the Ekman Emotional Expression Multimorph Task was .62 (SD = .06). Participants performed significantly better than chance (.20) on this task as indicated by a one-sample t test against the chance value [$t(99) = 67.95$, $p < .001$]. This once again confirmed that individuals are generally quite adept at identifying emotional expressions (Tracy & Robins, 2008).

In spite of the fact that emotion response alternatives remained constant throughout the experiment, a paired samples t test revealed that mean retrospective judgments were significantly higher than mean prospective judgments (prospective $M = 71.45$, $SD = 11.90$, retrospective $M = 74.28$, $SD = 11.74$) [$t(99) = 5.86$, $p < .001$, $d = .85$], though the difference was not as great as in Experiment 1. Although the difference was very small, people's mean confidence about what they would get right a few minutes hence, was slightly lower than their posttest confidence upon having seen the stimulus a second time and choosing the response alternative. This slight increase in confidence might be due to mere repetition of the question, and was also seen in the first experiment.

Relative Meta-Accuracy

The main finding of this study was that both the prospective and retrospective gamma correlations (G) between judgments and test performance were significantly greater than zero [prospective $G = .38$, $SD = .22$, $t(99) = 17.14$, $p < .001$; retrospective $G = .45$, $SD = .45$, $t(99) = 21.63$, $p < .001$] (see Figure 4). This finding demonstrated that once again, relative meta-accuracy was good, indicating that people have metacognitive insight into what they know and do not know on an emotion recognition task. Retrospective gamma correlations were slightly greater than prospective gamma correlations [$t(99) = 2.97$, $p < .004$, $d = .31$].

Relation of Relative Meta-Accuracy to Performance

Interestingly, and unlike in Experiment 1, task accuracy—the proportion correct on the emotion recognition test—was related to both prospective ($r = .32$, $p = .001$) and retrospective ($r = .43$, $p < .001$) gammas. People who were better at the emotion recognition task also tended to be more accurate in their relative metacognitive assessments. This is an intriguing finding, suggesting that those with good metacognitive insight into what they know and do not know in the emotional domain are also better able to read others' emotional expressions.

Global Meta-Accuracy

As in Experiment 1, global questionnaire responses were not predictive of task performance: The Mind-Reading Belief Scale was not correlated with task performance, $r = .05$ $p = .65$, n.s.; Interpersonal Reactivity Index (Perspective-Taking subscale) was not correlated with task performance, $r = .12$, $p = .25$, n.s.; Interpersonal Reactivity Index (Empathic Concern subscale) was

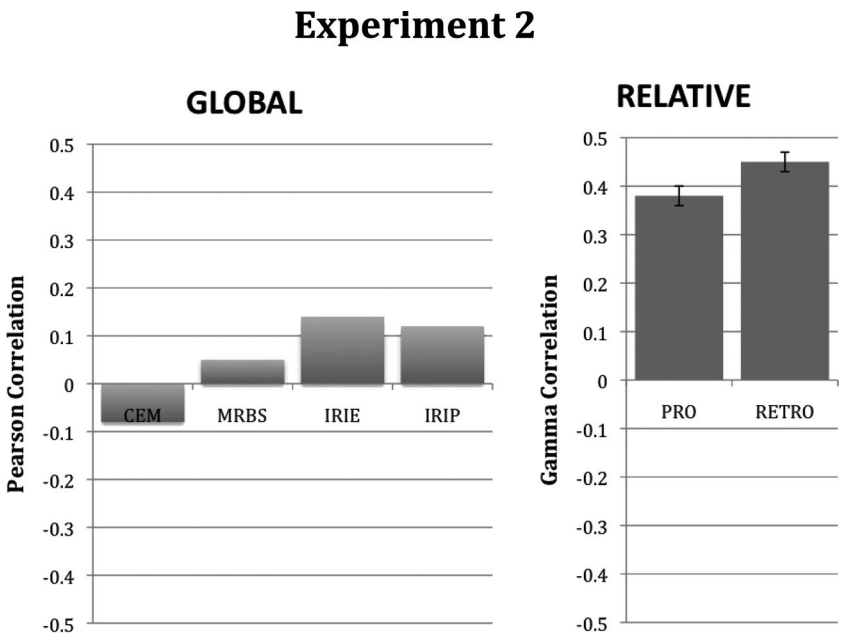

Figure 4. Global (Pearson correlation between questionnaire response and proportion correct on the emotion recognition task) and relative (gamma correlation, means of within participant correlations between confidence ratings and correct vs. incorrect responses on the emotion recognition task) meta-accuracy in Experiment 2 (The Ekman Multi-Morph Task). The Global questionnaires were the Columbia Emotion Measure (CEM), the Mind-Reading Belief Scale (MRBS), the Interpersonal Reactivity Index Empathy (IRI-E) and the Perspective-Taking (IRI-P) subscales. Gamma correlations are given for Prospective (pro) and retrospective (retro) confidence judgments as related to item accuracy.

not correlated with task performance, $r = .14$, $p = .18$, n.s.; the Columbia Empathy Measure was not correlated with task performance, $r = -.12$, $p = .25$, n.s.

The questionnaire responses (specifically the Mind-Reading Belief Scale) predicted magnitude of mean overall confidence judgments (prospective $r = .23$, $p = .02$ and retrospective $r = .30$, $p < .001$); that is, if people were overconfident on their global judgments they tended to be overconfident overall on the individual judgments as well.

Item Difficulty

As expected, participants were sensitive to the range of item difficulty reflected in the morphs. This sensitivity is captured in their mean accuracy, confidence judgments, and response time. Each pairwise comparison across-item difficulty was significant at the $p < .001$ level (see Table 1).

We attempted to further investigate and corroborate the finding by Zaki et al. (2008) that global metacognitive measures were accurate when used to predict performance on highly expressive items. However, when we used performance on the easiest morphs, questionnaire responses did not predict task accuracy (Mind-Reading Belief Scale $r = .07$, $p = .48$, n.s.; Interpersonal Reactivity Index (Perspective-Taking subscale) $r = .03$, $p = .76$, n.s.; Interpersonal Reactivity Index (Empathic Concern subscale) $r = .07$, $p = .49$, n.s.; Columbia Empathy Measure $r = -.21$, $p = .45$, n.s.). Interestingly, we did find that the Columbia Empathy Measure was negatively correlated with performance on the most difficult items. People who said they were most proficient at identifying emotions had the worst performance on the task (when only the most difficult items were used) $r = -.21$, $p = .04$. We also found that the Empathic Concern subscale of the Interpersonal Reactivity Index predicted performance on the medium difficulty items $r = .20$, $p = .05$.

In order to further investigate item difficulty, we computed separate gamma correlations for items at each level of difficulty using both prospective and retrospective confidence judgments correlated with accuracy. All prospective and retrospective gammas for easy, medium, and difficult items were significantly greater than zero at $p < .001$, Easy $G_{pro} = .28$ ($SD = .39$); $G_{retro} = .38$ ($SD = .34$); Medium $G_{pro} = .31$ ($SD = .27$); $G_{retro} = .39$ ($SD = .27$); Difficult $G_{pro} = .16$ ($SD = .26$); $G_{retro} = .23$ ($SD = .29$). We performed a 2×3 ANOVA to further investigate the effects of judgment phase (prospective or retrospective) and item difficulty (easy, medium, difficult). No significant interactions were found. However, as expected, there was a significant effect of judgment phase such that gamma correlations between prospective judgments and trial accuracy were significantly lower than gamma correlations between retrospective judgments and trial accuracy $F(1, 95) = 12.53$, $p = .001$, $d = .12$. There was also a main effect of item difficulty when the gammas were computed individually on only those subsets of items within participant, such that the most difficult items had significantly lower gamma correlations $F(2, 190) = 23.54$, $p = .00$, $d = .39$.

Discussion

Experiment 2 replicated the results of Experiment 1 with respect to both relative and global meta-accuracy. Individuals were able to accurately predict their performance on an item-by-item basis, prospectively, and evaluate it, retrospectively. This relative meta-accuracy indicates that individuals have the ability to both predict and evaluate their performance on this emotion recognition task.

In Experiment 1, test response choices had differed across trials and had not been available during prospective judgments, a fact that probably rendered the prospective judgments less accurate than when participants knew what the test alternatives would be, as was the case in Experiment 2. The similarity of prospective and retrospective judgment accuracy in Experiment 2 is likely a function of the equivalent information being available during the prospective and retrospective phases.

Global meta-accuracy, as captured by the questionnaires, was not predictive of task performance. Participants' estimations of their emotional and interpersonal sensitivity were unrelated to their actual ability to identify the expressed emotions. Unlike Zaki et al. (2008), we did not find that global metacognitive self-assessments predicted performance for highly expressive target stimuli. However, we used static faces rather than dynamic video clips. The static faces may have been processed differently than emotionally charged exchanges captured on video.

Participants were sensitive to the difficulty of items as reflected in their prospective and retrospective item-by-item judgments. Easy items were given higher mean judgments than items of medium difficulty, which in turn were assigned higher mean judgments than very difficult items. Interestingly, people who were good at discriminating what they knew and did not know in the relative accuracy task were also the people who performed best on the task—at least in Experiment 2 (though not in Experiment 1). Thus, this particular metacognitive index, in which people made specific judgments about whether they could correctly assess the emotion in particular faces, related to overall performance on the emotion recognition task itself, while none of the global questionnaire measures did.

Table 1
Experiment 2: Ekman Morphs—Performance According to Item Difficulty

Morph Difficulty	Mean Accuracy	Mean Prospective judgment (0–1)	Mean Retrospective judgment (0–1)	Mean test RT (in ms)
Easy	M = .80 (.09)	M = .82 (.11)	M = .84 (.10)	M = 4782.61 (1324.77)
Medium	M = .67 (.09)	M = .73 (.12)	M = .75 (.12)	M = 6024.02 (2013.28)
Difficult	M = .39 (.08)	M = .60 (.16)	M = .64 (.16)	M = 7432.98 (2483.73)

Note. Mean performance (standard deviations) according to item difficulty.

Summary and Concluding Discussion

These studies found, for the first time, that while individuals have great difficulty making accurate global trait judgments of their own empathy, they are able to make relative metacognitive assessments, correctly identifying the particular emotional exemplars that are more difficult for them to recognize. The failure of global metacognitive self-report measures to predict performance on emotion recognition tests had led to the assumption that individuals lack metacognitive awareness of their interpersonal sensitivity, but the high relative accuracy that we found here belies that conclusion. Our findings are consistent with the notion that human beings possess a highly developed social mind and are, therefore, socially adept and skilled at recognizing emotions (Tracy & Robins, 2008; Adolphs, 2006; Ekman, 1982). It is likely that reaching this advanced level of social functioning is accompanied by some metacognitive awareness, as our results indicate. Interpersonal emotional interactions, in the real world, are highly variable and success often requires that the individuals engaged recognize both their accuracy and inaccuracy.

While these data demonstrate that individuals have good relative meta-accuracy when viewing static faces, it would be interesting to know whether these results generalize to more dynamic displays. Most of the previous work on metacognition of emotion recognition has used videotaped interactions between live actors engaged in emotional exchanges. Assessing relative meta-accuracy while viewing such naturalistic interactions would extend and add ecological validity to the current findings.

Also, while the current findings demonstrate that people can reliably make accurate relative metacognitive judgments, we do not know how they use this metacognitive knowledge. However, accurately assessing one's performance on a given test of emotion recognition might result in improved performance on other emotion recognition tasks if the person is able to use this metacognitive knowledge to appropriately seek new information in the face of uncertainty. Such a feedback loop might, in the real world, give rise to a positive relation between metacognitive ability and improved emotional recognition. We offer this possibility with caution, since no relation between relative meta-accuracy and task performance was observed in Experiment 1. However, in Experiment 2—an experiment with stronger metacognitive results for reasons mentioned above—those people with more accurate metacognitive judgments were also better at the emotion recognition task itself. A link between metacognitive insight and task performance may be of particular importance within the realm of emotion recognition. Both abilities, emotion recognition and metacognition, require that the participant reflect upon and understand mental states.

Establishing whether or not people's metacognition could be leveraged to improve emotion recognition may be particularly important when working with people who lack proficiency at emotion recognition, such as people with autism or with Asperger's syndrome. Interventions are often designed to teach people how to correctly identify emotions based on facial and other nonverbal cues. If the individual lacks metacognitive insight into his or her ability to recognize emotions, he or she may not know when to use these newly learned strategies. In order to effectively implement strategies designed to teach emotion recognition, a reasonable strategy may be to first determine if a metacognitive deficit exists, or not. If it does, then that might be the starting point for therapy. If not, then this metaknowledge might be implemented in appropriately directing the emotional learning to where it is needed most.

In summary, global judgments are unlikely to capture the specific intricacies that categorize individual emotional encounters. However, when measured on an item-by-item basis, people are good at making metacognitive assessments of their emotional knowledge, and some people are *very* good at it. And this kind of metacognition could be very useful in the real world—allowing people to seek more information when they need it, and to learn from the consequences of such seeking. While both perspectives on metacognition—the global and the relative—are valid, the relative accuracy perspective, in which the person evaluates whether s/he knows or does not know in each instance according to the dynamics of each specific encounter, may be more representative of people's nuanced ability to successfully navigate real-world interpersonal encounters.

References

Adolphs, R. (2006). Perception and emotion: How we recognize facial expressions. *Current Directions in Psychological Science, 15,* 222–226. doi:10.1111/j.1467–8721.2006.00440.x

Ames, D. R., & Kammrath, L. K. (2004). Mind-reading and metacognition: Narcissism, not actual competence, predicts self-estimated ability. *Journal of Nonverbal Behavior, 28,* 187–209. doi:10.1023/B:JONB.0000039649.20015.0e

Archer, D., & Costanzo, M. (1988). *The Interpersonal Perception Task.* Berkeley, CA: University of California Extension Media Center.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry, 42,* 241–251. doi:10.1017/S0021963001006643

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68. doi:10.1037/0096–3445.127.1.55

Blair, R. J. R., Colledge, E., Murray, L., & Mitchell, D. G. V. (2001). A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies. *Journal of Abnormal Child Psychology, 29,* 491–498. doi:10.1016/S0278-2626(03)00276–8

Darwin, C. (1872). *The expression of the emotions in man and animals* (3rd ed.) New York: Oxford University Press.

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10,* 85–103.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Washington D. C.: Sage Publications.

Ekman, P. (1982). *Emotion in the human face.* New York, New York: Cambridge University Press.

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

Finn, B. (2008). Framing effects on metacognitive monitoring and control. *Memory & Cognition, 36,* 813–821. doi:10.3758/MC.36.4.813

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34,* 906–911. doi:10.1037/0003-066X. 34.10.906

Glenberg, A. M., Sanocki, T., Epstein, W., & Morris, C. (1987). Enhancing calibration of comprehension. *Journal of Experimental Psychology: General, 116,* 119–136. doi:10.1037/0096–3445.116.2.119

Humphrey, N. (1984). *Consciousness regained.* New York: Oxford University Press.

Ickes, W. (1993). Empathic accuracy. *Journal of Personality, 61,* 587–610.

Ickes, W., & Aronson, E., (2003). *Everyday mind reading* (pp. 169–17). Amherst, New York: Prometheus Books.

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology, 59,* 730–742. doi:10.1037/0022–3514.59.4.730

Levenson, R. W., & Ruef, A. M. (1992). Empathy: A physiological substrate. *Journal of Personality and Social Psychology, 63,* 234–246. doi:10.1037/0022–3514.63.2.234

Marangoni, C., Garcia, S., Ickes, W., & Teng, G. (1995). Empathic accuracy in a clinically relevant setting. *Journal of Personality and Social Psychology, 68,* 854–869. doi:10.1037/0022–3514.68.5.854

Masson, M. E. J., & Rotello, C. M. (2009). Sources of bias in the Goodman-Kruskal/gamma correlation coefficient measure of association: Implications for studies of metacognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 509–527.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review, 15,* 174–179. doi:10.3758/PBR.15.1.174

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133. doi:10.1037/0033–2909.95.1.109

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation, Vol. 26* (pp. 125–173). New York: Academic Press.

Realo, A., Allik, J., Nolvak, A., Valk, R., Ruus, T., Schmidt, M., & Eilola, T. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality, 37,* 420–445. doi:10.1016/S0092-6566(03)00021–7

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In J. Metcalfe and A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 93–113). Cambridge, MA: MIT Press.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95,* 66–75. doi 10.1037/0022–0663.95.1.66

Tracy, J. L., & Robins, R. W. (2008). The automaticity of emotion recognition. *Emotion, 8,* 81–95. doi:10.1037/1528–3542.8.1.81

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science, 19,* 399–404. doi 10.1111/j. 1467–9280.2008.02099.x

Received November 4, 2009
Revision received July 8, 2010
Accepted July 30, 2010 ■