2020, American Psychological Association https://doi.org/10.1037/pas0000818

Valence in the Reading the Mind in the Eyes Task

Chloe C. Hudson[a]

Amanda L. Shamblaw[a]

Kate L. Harkness[a]

Mark A. Sabbagh[a]

[a]Queen's University, Canada

Author note: Correspondence should be addressed to Chloe Hudson. Address: Department of Psychology, 62 Arch Street, Queen's University, Kingston, Ontario, Canada, K7L3N6. E-mail address: c.hudson@queensu.ca. This work was supported by a grant awarded to Dr. Harkness and Dr. Sabbagh from the Social Sciences and Humanities Research Council [435-2012-1536, 2012].

Abstract

The Reading the Mind in the Eyes task (RMET; Baron-Cohen et al., 2001) is commonly used to assess theory of mind abilities in adults. In the task, participants pair one of four mental state descriptors with a picture of the eye region of a face. The items have varying emotional valence, and nearly 100 studies have examined whether performance on this task varies with item valence. However, efforts to address this question have been hampered by cross-study inconsistencies in how item valence is assessed. Thus, the goal of this study was to establish reference ratings for the valence of RMET items. In Study 1, we recorded valence ratings for each RMET item with a large sample of raters (n = 164). We illustrated how valence categories are essentially arbitrary and largely influenced by sample size. In addition, valence ratings were continuously distributed, further questioning the validity of imposing categorical distinctions. In Study 2, we used an archival dataset to demonstrate how the different categorization schemes resulted in conflicting conclusions about the association between item valence and RMET performance. However, when we examined the association between item valence and performance in a continuous manner, a clear U-shaped pattern emerged: items that had more extreme valence ratings (negative or positive) were associated with better performance than items with more neutral ratings. We conclude that using the item valence ratings we report, and treating item valence as a continuous rather than categorical predictor, will help bring consistency to the study of the association between item valence and performance in the RMET.

Keywords: Reading the Mind in the Eyes task, valence, theory of mind

Public Significance Statement

We assessed the valence of items in the Reading the Mind in the Eyes task (RMET; Baron-Cohen, Wheelright, Hill, Raste, & Plumb, 2001) using the largest sample of raters to date. While past researchers have categorized the items into valence categories, we found that the valence of items was continuously distributed. We provide reference valence ratings that can be used in future studies, which we hope will help to bring consistency and validity to the study of valence in the RMET.

Valence in the Reading the Mind in the Eyes Task

Theory of mind is the understanding that people's observable actions are motivated by their mental states, including their intentions, beliefs, and desires (Sabbagh & Bowman, 2018). Subtle differences in using theory of mind understandings to make appropriate and accurate judgments about the contents of others' mental states are observed in the general population and are associated with several important social qualities, including cooperation (Paal & Bereczkei, 2007) and empathy (Bruneau & Saxe, 2012; Gonzalez-Liencres, Shamay-Tsoory, & Brüne, 2013). Less accurate theory of mind understanding relative to healthy populations has been found in several clinical conditions, including schizophrenia (Bora, Yucel, & Pantelis, 2009), autism spectrum disorder (Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998), bipolar disorder (Bora, Bartholomeusz, & Pantelis, 2016), and major depressive disorder (Bora & Berk, 2016). In the clinical literature, poorer performance on theory of mind tasks is associated with poorer interpersonal functioning (Couture, Penn, & Roberts, 2006; Fett et al., 2011; Tager-Flusberg, 2003).

The most commonly used task to assess individual differences and clinical group differences in theory of mind in adults is the Reading the Mind in the Eyes task (RMET; Baron-Cohen, Wheelright, Hill, Raste, & Plumb, 2001). In this task, participants are presented with a series of 36 black-and-white photographs of the eye-region of faces with four adjectives (the correct response and three distractors) placed at each corner of the photograph (see Figure 1 for a sample item). Participants are asked to choose which of the four adjectives best represents the mental state of the individual in the photograph. Accuracy is operationalized as either the total number, or percent, of correct responses. Correct response options were originally chosen by the task creator and piloted on groups of eight judges. In order for an item to be validated, a

minimum of five of the eight judges had to agree on the target word, and no more than two judges were to pick any single distractor item.

The RMET was developed to examine overall accuracy in theory of mind decoding. However, since its development, many researchers have capitalized on the fact that the mental states being judged vary in emotional valence; some are negative (e.g., "hostile"), some neutral (e.g., "pensive"), and others positive (e.g., "friendly"). There are a number of reasons to consider item valence when investigating theory of mind accuracy. For example, healthy individuals show attentional biases away from negative stimuli and towards positive stimuli (Joorman & Gotlib, 2007). Therefore, one could hypothesize that individuals may be more accurate and faster at decoding positive versus negative or neutral mental states. Further, individual differences in healthy adults' attentional biases are predictive of important beliefs about the world, such as optimism (Segerstrom, 2001) and personality traits, such as extroversion and neuroticism (Amin, Constable, & Canli, 2004). Thus, individual differences in performance across item valence may help to explain individual differences in everyday traits or behaviours. In contrast, individuals with depression show attentional biases towards negative stimuli, as well as biases in the recognition of sad versus happy emotion in faces (Bourke, Douglas, & Porter, 2010). It is possible that biases in the foundational skill of mental state decoding may provide the basis for negative biases in these higher-order cognitive and behavioral domains (see Harkness et al., 2005).

However, since the RMET was not developed with valence in mind, researchers interested in this question have had to develop their own valence categories. The first valence categorization scheme was provided by Harkness et al. (2005). They gave a small sample of women (n = 12) all 36 pictures from the RMET with the correct adjective centred below the

photograph. Participants ranked each picture-adjective pair on a 7-point scale (1 = *very negative*, 4 = *neutral*, and 7 = *very positive*). Items that had mean ratings significantly above neutral were categorized as positive (n = 8), items with mean ratings significantly below neutral were categorized as negative (n = 12), and items that did not differ significantly from neutral were categorized as neutral (n = 16). Other studies examining item valence have created their own valence categories using their own, even smaller, samples of raters (n = 3 to 5; Ali & Chamorro-Premuzic, 2010; Purcell, Phillips, & Gruber, 2013; Sandvik, Hansen, Johnsen, & Laberg, 2014). Still other categorizations have been developed from ratings of different aspects of the stimuli: Only the RMET picture (Hünefeldt, Laghi, & Ortu, 2013; Kometer, Schmidt, Bachmann, Studerus, Seifritz, & Vollenweider, 2012; Kynast & Schroeter, 2018; Meijer-van Abbema & Koole, 2017; Meyer, 2009; Scott, Levy, Adams, & Stevenson, 2011); or only the correct adjective (Hezel & McNally, 2014; Oldershaw et al., 2011); or based on the correct response options' score on the Dictionary of Affect in Language (Whissell, 2009), which contains over 4000 words that are scored along affective dimensions (Konrath, Corneille, Bushman, & Luminet, 2014; Luminet, Grynberg, Ruzette, & Mikolajczak, 2011).

Studies seeking to adopt a valence categorization scheme, therefore, have had many such schemes to choose from. The resulting valence categories from the Harkness et al. (2005) sample are by far the most popular, and have now been used in over 50 studies, including 11 in the past two years (Balter et al., 2018; Belardinelli, Huenefeldt, Maffi, Squitieri, & Migliore, 2018; Anupama, Bhola, Thirthalli, & Mehta, 2018; Caldú et al., 2019; da Costa, Vrabel, Zeigler-Hill, & Vonk, 2018; Kattan, 2018; Kopera et al., 2018; Lischke et al., 2019; Pahnke et al., 2019; Rnic et al., 2018; Yu, Chen, Tu, Tsao, & Tan, 2018). Other researchers have adopted the valence categorizations that were created using only the RMET pictures (Berenson et al., 2018; Frick et

al., 2012; Meyer & Morey, 2015; Tay, Hulbert, Jackson, & Chanen, 2017; Weinstein et al., 2016) or only the correct response option (Kenyon et al., 2012; Medina-Pradas Navarro, Álvarez-Moya, Grau, & Obiols, 2012; Pedreño, Pousa, Navarro, Pàmias, & Obiols, 2017; Radke & de Bruijn, 2015; Tapajóz Pereira de Sampaio, Soneira, Aulicino, & Allegri, 2013; Zainal & Newman, 2017). Further, two studies used a combination of multiple categorization schemes that were created using different methods (Baribeau et al., 2015; Fossati et al., 2014). Finally, a number of studies reported on valence effects but did not report whether they used preexisting valence categorization schemes or whether they created their own (Chan, 2008; Espinós, Fernández-Abascal, & Ovejero, 2018; Hanna, 2015; Jermakow & Brzezicka, 2016; Kenyon, Alvares, Hickie, & Guastella, 2013; Lenton-Brym, Moscovitch, Vidovic, Nilsen, & Friedman, 2017; Mannava, 2012; Neal & Chartrand, 2011; Nejati, 2018; Oldershaw, Hambrook, Tchanturia, Treasure, & Schmidt, 2010; Overbeck & Droutman, 2013; Reid, 2017; Shaw, Bramham, Lawrence, Morris, Baron-Cohen, & David, 2005; Uzefovsky, Shalev, Israel, Knafo, & Ebstein, 2012).

Not surprisingly, the different valence categorization schemes differ from one another in terms of the items included in each valence category. To illustrate, Table 1 summarizes the results of six studies that used their own sample to create valence categories *and* reported the items in each category (Harkness et al., 2005; Hezel & McNally, 2014; Kometer et al., 2012; Konrath et al., 2014; Meijer-van Abbema & Koole, 2017; Scott et al., 2011). Three of these studies had the authors or research assistants from their lab categorize the RMET items (Hezel & McNally, 2014; Konrath et al., 2014; Meijer-van Abbema & Koole, 2017), two studies recruited undergraduate participants (Harkness et al., 2005; Scott et al., 2011), and one study did not indicate how participants were recruited (Kometer et al., 2012). See Table 1 for sample size and

gender distributions when provided. Alarmingly, when items are compared across these categorization schemes, there are four items that were categorized as negative, neutral, and positive, depending on which scheme was used. This discordance has important implications as it is likely to result in different conclusions being drawn regarding the fundamental question of how individuals and groups differ in their decoding of negative versus neutral versus positive mental states.

In summary, nearly 100 studies have examined the association between item valence and performance in the RMET. The RMET is the only task to our knowledge that assesses participants' ability to decode complex and subtle mental states (i.e., theory of mind decoding). Further, the stimuli depict a range of valences, which many researchers have categorized as positive, neutral, and negative to examine how accuracy in decoding mental states varies with item valence. However, as noted above, the field lacks a consistent operationalization of item valence in this task. In the current study, we sought to establish reference ratings for the valence of the RMET items that can serve as the standard for examining questions relating to valence in theory of mind decoding. In Study 1 we obtained valence ratings from largest sample of raters to date to gain confidence in determining which stimuli are likely to belong to a given valence category (i.e., negative, neutral, positive), based on relatively strict criteria. To preview our results, we find that although criteria can be applied to create categories, the item ratings themselves follow a continuous distribution without clear, non-arbitrary category boundaries. In Study 2, we illustrate the shortfalls of the categorical approach to defining valence in the RMET by showing that the relation between valence and performance changes drastically depending on which of the valence categorization schemes is used, including that provided in Study 1. Finally,

we propose and empirically examine a novel continuous method of defining item valence in the RMET.

Study 1

The goal of Study 1 was to define the valence of RMET items using the largest sample of raters to date. We used a rating method similar to Harkness et al. (2005), such that the valence of both the picture and the correct response option were rated together. Our reason for choosing this approach is that pictures and correct response options are not always perfectly matched in their valence. For example, while "distrustful" appears to be a negatively-valenced word, the picture associated with distrustful (item 34) is not overtly negative. Indeed, when items were classified using only the picture, item 34 was classified as a *positive* item (Scott et al., 2011). Further, while the distractor words also carry valence, we did not want participants' potential experience of being unsure which of the four adjectives was best to influence their valence ratings. That is, if items are rated as more negative simply because they are harder (and not because of the conceptual content of the items) then this would induce a spurious correlation between item valence and performance. Therefore, as in Harkness et al. (2005), we reasoned that the picture and the correct response option together capture the relevant stimulus information that participants are exposed to while performing the original RMET.

Method

Participants. This study was approved by the General Research Ethics Board at Queen's University according to the recommended principles of Canadian ethics guidelines and University policies. Participants were 225 undergraduate students (M_{age} = 18.41 years, range = 17-23, 82% female) enrolled in an introductory psychology course. The majority of participants identified as White (69.6%), followed by East Asian (17.3%), South Asian (5.6%), bi-

racial/other ethnicity (5.0%), Black (1.4%), and Hispanic (0.9%). Participants were recruited through a departmental website through which they signed up for the study, and they were compensated with course credit. All participants were of age to consent to research, capable of providing informed consent, and spoke English. To increase the likelihood that the sample was representative of the population of first-year undergraduates at this institution there were no other inclusion or exclusion criteria.

Measures and procedure. Participants completed the study through an online survey website (surveymonkey.com). Participants first provided informed consent. They then read instructions and completed a modified version of the RMET task using a computerized presentation. In this task, participants were presented with the original 36 black-and-white photographs taken of the eye-region, from just above the eyebrow to midway down the bridge of the nose, in random order (Baron-Cohen et al., 2001). Each photograph measured 15cm x 6cm. For the purposes of the current study, the correct mental state term used to describe the photograph was centered under each photograph. Participants were given the following instructions: "*In the following task, you will see 36 photographs of eyes, one at a time, with an adjective centered below the photograph. You will be asked to judge whether each eye-adjective pair conveys a positive, neutral, or negative emotion on a 7-point scale from Very Negative to Very Positive*." Participants were responded on a 7-point Likert-type scale with the following anchors: 1 = *Very Negative*, 4 = *Neutral*, 7 = *Very Positive*. Upon completion, participants were informed of the goal of the study and were granted course credit.

Results

Preliminary analyses. Responses were inspected for evidence of non-systematic responding. A mean valence score for each item was calculated using the data from all 225

participants. We conducted a Pearson correlation for each participant to explore the relation between each participant's individual responses on the 36 items and the mean valence ratings on the 36 items. Correlations close to zero suggest random responding, while negative correlations suggest scale reversal. Participants whose overall ratings correlated with mean ratings at less than $r = .20$ were excluded ($n = 57$). Further, four additional participants were excluded for restricted variability in responding (i.e., greater than 50% of their responses were restricted to one value). This approach is a novel method to detect non-systematic responding and requires further research.

The final dataset included 164 participants (82% female) with a mean age of 18.41 years ($SD = 1.06$ years). The ethnicity distribution of this final sample was as follows: White (76.6%), East Asian (13.3%), bi-racial/other ethnicity (6.0%), South Asian (3.2%), and Black (1.3%). There were no significant differences in age, identified gender, or ethnicity between participants included versus excluded from the final dataset.

Valence analyses. The distribution of responses, mean score, and median score for each item are presented in Figure 2. We first conducted a series of one-sample *t*-tests comparing the mean valence rating of each stimulus with the neutral rating (i.e., 4). Only 7 items failed to differ significantly from neutral. The remaining items were classified as negative (13 items) or positive (16 items). The results of this categorization scheme are presented in Table 1 (Study 1a).

The large sample size in this study resulted in relatively few items being categorized as neutral. Even small deviations from neutral (e.g., less than .2 scale degrees) were significant. In contrast, prior studies that created valence categories were based on much smaller sample sizes, which require larger deviations to be captured as statistically different from neutral. To illustrate the impact of sampling errors of estimates on valence, we conducted an additional analysis using

bootstrapped confidence intervals (1000 iterations) for the *t*-statistic of each item based on a sample size of 60. This sample size was selected because it reflects the average sample size for studies that employ the RMET.

The average *t*-scores and 95% confidence intervals are presented in Figure 2. Thirteen items were categorized as positive because they had a 95% confidence interval that fell entirely above +2.00 (i.e., corresponding with a *p*-value less than .05 or statistically significant), five items were categorized as negative because they had a 95% confidence interval that fell entirely below -2.00, and 18 items were categorized as neutral because they had a 95% confidence interval that fell within the -2.00 to +2.00 range. The results from this categorization procedure are presented in Table 1 (Study 1b) to enable comparison with the categories from the first analysis and prior studies.

Discussion

There are two important implications of the results of Study 1 that we subsequently followed up on in Study 2. First, the one-sample *t*-test approach to defining categories is unduly influenced by sample size. With our large sample of raters, the *t*-tests resulted in a very small number of items categorized as neutral (i.e., most items differed significantly from neutral). To illustrate the impact of sample size on the resulting valence categories, we used a bootstrap approach in which we randomly selected sub-samples of 60 from our larger sample and created confidence intervals around the means resulted in a categorization scheme with substantially more items being categorized as neutral. That is, changes in sample size (and corresponding sampling error in estimates) had a substantial impact on whether items were categorized as positive, neutral, or negative.

Nevertheless, our second conclusion is that we did not observe any natural discontinuities at the boundaries between the negative, neutral, and positive categories (see Figure 2). Instead, valence ratings appeared to be distributed in a continuous fashion across items. For example, the difference between the mean valence rating for "regretful" (categorized as neutral in the bootstrapped analyses) and "uneasy" (categorized as negative) was .05 on our 7-point rating scale. Similarly, the difference between the mean valence rating for "reflective" (categorized as neutral) and "playful" (categorized as positive) was .04. The categorical boundaries that we chose to impose (i.e., confidence intervals within or outside the -2.00 to +2.00 range), while reasoned, were essentially arbitrary, and movement in one direction or the other would simply have increased or decreased the number of neutral items relative to the number of negative or positive items. As is the case with any continuous construct, imposing categorical boundaries will force distinctions that may not be meaningful. Therefore, an important question that arises from our findings above is that if item valence is distributed in a continuous manner, should it also be operationalized and analyzed as a continuous variable in measurement?

The goals of Study 2 address the above question in two inter-related ways. First, an important question is whether the different valence categorization schemes lead to different results. We apply seven different valence categorization schemes to illustrate how the differences across schemes influence the relation of RMET performance by valence. Given that the authors of each valence categorization scheme made justifiable but essentially arbitrary decisions about where to differentiate neutral items from negative or positive items, we expect that the differences in the valence categorization scheme will lead to different results on this fundamental research question. If the use of different categorization schemes leads to different conclusions *in the same dataset*, we have evidence for the importance of a universal reference valence rating

and empirically illustrate the problems with conceptualizing item valence as a categorical construct. Conversely, it is possible that in the aggregate, the different categorization schemes might lead to broadly similar conclusions about ways in which item valence affects performance on the task. Second, we illustrate an approach that seeks to capture the relation of performance across item valence without drawing arbitrary distinctions among categories.

Study 2

The first goal of Study 2 was to illustrate the problem of arbitrary category boundaries by showing that different valence categorization schemes result in different patterns of performance within the same dataset. In the current study we used an archival dataset to provide a simple illustration of this phenomenon. Specifically, we examined differences in decoding accuracy and response time across items of a negative versus neutral versus positive valence for each of the six prior categorization schemes described in the literature, and the bootstrapped analyses from Study 1. We acknowledge that most studies investigating valence in the RMET are interested in interaction effects related to individual and group differences in performance across valence. However, if the main effect of valence is different across different valence categorization schemes, then researchers have little basis upon which to base a priori hypotheses regarding interaction effects.

The second goal of Study 2 was to illustrate the relation of accuracy and valence using a continuous metric of valence in the same archival sample as above. As proof-of-concept, we test the linear and quadratic trends in the data, although we acknowledge that more complex trends are possible and could be modelled based on the particular research question.

Method

Participants. The current sample included 122 undergraduate students (M_{age} = 20.34, 76.23% female, 55.74% White) who were drawn from a larger study examining the relation of theory of mind and mood (see Hudson et al., 2018). Participants were recruited through an introductory psychology participant pool and advertisements.

Measures. ***The Reading the Mind in the Eyes Task*** (RMET; Baron-Cohen et al., 2001). Participants completed the original RMET to assess theory of mind decoding skill. The task consists of 36 black-and-white photographs of the eye region of faces with each standardized to the same size (15cm x 6 cm). Participants were asked to make a forced choice between four adjectives (the target response option and three distractor response options) that describe what the person is thinking or feeling. The RMET was administered by computer using E-Prime software (Psychology Software Tools, 2016) and participants responded by pressing the S, X, K, or M key on the keyboard. These keys are spatially analogous to the adjectives presented on the screen. Accuracy was defined as percent correct and response time was recorded in milliseconds.

Using the raw data from the RMET we calculated accuracy and response time scores on negative, neutral, and positive eyes based on the six prior categorization schemes described in the literature, and the bootstrapped analyses from Study 1. Konrath et al. (2012) only provided negative and positive categories. Therefore, each participant received seven separate negative and positive accuracy and response time scores and six separate neutral accuracy and response time scores.

Procedure. The General Research Ethics Board at Queen's University approved the study protocol. All participants provided written, informed consent. The larger study from which these participants were drawn involved two in-person sessions separated by one week. The RMET was completed in session 1. Participants were first provided with a practice trial followed

by the full 36-item RMET. Participants were instructed to respond as quickly and accurately as possible. Responses and response times (in milliseconds) were recorded digitally with measurement error not exceeding 16 ms. Participants received course credit or $40 for their participation.

Results

Valence categorization comparison. One-way repeated-measures analyses of variance (ANOVAs) were conducted to examine the pattern of results across the seven valence categorization schemes in overall accuracy and response time, respectively, by valence (negative, neutral, positive). Significant main effects were followed-up with pairwise comparisons using the least significant difference test.

Accuracy. Performance by valence across the different valence categorization schemes is presented visually in Figure 3. The main effect of valence was significant for three of the seven valence categorization schemes. Using Hezel and McNally's (2014) scheme, accuracy on the negative and positive items was significantly higher than on the neutral items, $t(121) = 2.10$, $p = .04$ and $t(121) = 2.86$, $p = .005$, respectively. In direct contrast, using Meijer-van Abbema and Koole's (2017) scheme, accuracy on the neutral items was significantly higher than on the negative or positive items, $t(121) = 3.09$, $p = .002$ and $t(121) = 3.23$, $p = .002$, respectively. Finally, using the Scott et al. (2011) scheme, accuracy on the negative items was significantly higher than on the neutral or positive items, $t(121) = 4.42$, $p < .001$ and $t(121) = 5.07$, $p < .001$, respectively.

Response time. Performance by valence across the different valence categorization schemes is presented visually in Figure 4. RMET response times differed significantly by valence for three of the categorization schemes. Using Hezel and McNally's (2014) scheme,

participants were significantly faster on positive compared to neutral items, $t(121) = 2.26$, $p = .03$. Using the Konrath et al. (2014) scheme, participants were significantly faster on negative compared to positive items, $t(121) = 2.17$, $p = .03$. Finally, using the Scott et al. (2011) scheme, participants were significantly faster on the negative items than the neutral or positive items, $t(121) = 2.95$, $p = .02$ and $t(121) = 3.18$, $p = .003$, respectively.

Continuous analyses. Two mixed-effects regression analyses using the lme4 package for R (Bates, Mächler, Bolker, & Walker, 2015) were performed to examine accuracy and response time, respectively, across valence in a continuous fashion. For those interested, we have posted the rudimentary R code and the data for these analyses on the OSF site for this project (along with all the data from Study 1) at the following link: https://osf.io/7f92k/

Accuracy. In the first regression, the linear and quadratic terms for the valence ratings from Study 1 were entered as within-subject predictor variables, and accuracy on each item was entered as the dichotomous outcome variable (0 = inaccurate, 1 = accurate). In this analysis, the linear valence term was not a significant predictor of accuracy, $b = 4.80$, $SE = 2.86$, $z = 1.68$, $p = .09$. The quadratic valence rating term was a significant predictor of accuracy, $b = 14.93$, $SE = 2.68$, $z = 5.56$, $p < .001$ (see Figure 5a). Performance was more likely to be accurate the more extremely negative or positive the items became.

Response time. In the second regression, the linear and quadratic terms for the valence ratings from Study 1 were entered as within-subject predictor variables, and response time on each item was entered as the continuous outcome variable. Similarly to the accuracy analyses, the linear valence term was not a significant predictor of response time, $b = -2239.7$, $SE = 2343.8$, $t = -0.96$, $p = .34$, whereas the quadratic valence term was a significant predictor of

response time, b = -9661.1, SE = 2206.3, t = -4.38, $p < .001$ (see Figure 5b). Response times were shorter the more extremely negative or positive the items became.

General Discussion

As expected, in Study 2, the pattern of accuracy and response time across mental states of negative, neutral, and positive valence differed dramatically based on the particular valence categorization scheme that was employed. Indeed, all patterns were represented, with negative eyes, neutral eyes, and positive eyes each taking a turn showing the highest level of accuracy. As such, conclusions regarding which types of mental states are easiest or most difficult for adults to decode differ dramatically depending on which valence categories are used.

This lack of a clear picture for the relation of accuracy by valence not only makes comparison of results across studies difficult, but it also impedes the ability of researchers to formulate other hypotheses for between-group individual differences comparisons across item valence. For example, if the field cannot agree on how theory of mind performance is affected by the valence of mental states in a sample of undergraduate students, then it has little basis for developing hypotheses regarding accuracy differences by valence between healthy individuals and those with clinical conditions associated with general theory of mind impairments. One particularly worrying implication of the availability of multiple valence categorization schemes is that authors could engage in "*p*-hacking." That is, researchers may be tempted to analyze their data using each of the different categorization schemes and then choose the method that supports their hypotheses. Furthermore, categorization schemes with few items in a category may result in scores that are less reliable because large changes in valence category scores can be caused by relatively small changes in performance. For example, Meijer-van Abbema and Koole's (2017)

neutral category contains six items. As such, participants' neutral performance will drop by 17% each time they answer a single item incorrectly.

In Study 1 we argued that categorizing the items obscured the fundamentally continuous nature of the item valences. When we used the same ratings derived from our large sample of raters in Study 1, but erased the category boundaries, a much clearer picture emerged. Specifically, there was a U-shaped association between item valence and performance. Accuracy was poorer and response times were longer for items that were closer to a neutral rating. Conversely, accuracy was higher and response times were shorter the closer an item was to being extremely negative or extremely positive. Very concretely, the results of the analyses for Study 2 demonstrate clearly that theory of mind judgments were easier for the more strongly valenced (negative *or* positive) items, and more difficult for the items as they approached neutral ratings.

There are a number of potential explanations for the above pattern of findings, a full discussion of which is beyond the scope of the current methodological paper. For example, mental states that carry strong emotions, such as "despondent", "hostile," or "friendly," may have corresponding social cues that more closely resemble those of facial expressions of the basic emotions that are universally recognized in humans (e.g., "sadness", "anger", "joy", etc.; Ekman & Friesen, 1986). In contrast, mental states that are less emotionally laden, such as "preoccupied" or "serious," may not involve social cues that obviously correspond with these basic emotions. The implications for the current discussion, however, are that this general pattern of performance in a sample of undergraduate students can now be used to develop hypotheses for samples that may be expected to diverge from this general pattern due to individual difference characteristics (e.g., depression or schizophrenia, pre-existing biological and/or temperamental characteristics, etc.) or for performance following an experimental and/or treatment

manipulation. For example, researchers could explore whether the association between item valence and RMET performance differs between clinical and non-clinical groups by introducing a between-subject diagnostic status variable (e.g., depressed vs. non-depressed) and exploring the interaction between diagnostic status and valence ratings. That is, researchers could use a mixed-effect linear or logistic regression to explore the interaction between continuous valence ratings and diagnostic status (presence vs. absence of diagnosis) on RMET performance (accuracy or response time).

There are important limitations that should be acknowledged with the current study. The samples from both Study 1 and Study 2 comprised undergraduate volunteers from a suburban Canadian university and, thus, they are relatively homogenous in terms of age and ethnicity. Therefore, generalization of the current valence ratings and accuracy results to more representative large samples of adults is required. In addition, althought the goal of this paper was to establish a set of reference valence ratings that are representative of a broad population (i.e., all undergraduate students, including those who may have clinical diagnoses), it does not address a number of interesting questions about the kinds of population-level variables that affect valence ratings. For example, because we did not collect information on psychiatric history in Study 1, we could not explore whether psychiatric diagnoses influenced valence ratings. Whether particular populations (e.g., those a diagnosis of major depressive disorder) would rate the valence of RMET items differently than other populations (e.g., those without any clinical diagnoses) remains an open question. Similarly, more research is needed to understand whether one's *experience* of valence influences accuracy on the RMET. Nevertheless, our study provides the novel contribution of creating valence reference ratings that are representative of a broad population that can be used with any subset of this population.

There are also a number of limitations to using the RMET to explore how valence influence theory of mind decoding accuracy. First, while the RMET items depict a range of valences, there are more positive items than negative ones. Second, while Baron-Cohen specified that distractor items were meant to have the same emotional valence as the target response option "as far as possible" (Baron-Cohen et al., 2001), a review of all four response options for each item demonstrates that the response options are not always matched on emotional valence. For example, the correct answer for the item presented in Figure 1 is "playful". However, if one thought that the correct answer was "irritated" or "bored", then the item might be considered more negative. Consequently, perception of valence in the original RMET may be biased by accuracy. Finally, negative items are more likely to be taken from male and older faces than they are to be taken from female or younger faces (Kynast & Schroeter, 2018). As such, there is good reason to suggest that researchers whose primary research focus is on the association between item valence and mental state decoding accuracy may benefit from a new task that addresses these issues.

In summary, we find that when RMET item valence is assessed with a large number of participants, the item valences vary continuously and do not show clear evidence of having category boundaries. Researchers' tendencies to impose such category boundaries is likely, at least in part, responsible for inconsistencies in the literature with respect to how item valence affects RMET performance. We propose an alternative approach that involves treating the item valence ratings that we provide as a continuous predictor, and present evidence suggesting that this approach most clearly captures the relation between item valence and performance accuracy in the RMET in a sample of university students. Ultimately, we hope that by adopting this approach, researchers will be able to better apply the RMET to study the possible interplay

between theory of mind and emotional valence, and the various conditions that can affect this important aspect of everyday social cognitive functioning.

References

Ali, F., & Chamorro-Premuzic, T. (2010). Investigating theory of mind deficits in nonclinical psychopathy and Machiavellianism. *Personality and Individual Differences, 49*(3), 169-174. doi: 10.1016/j.paid.2010.03.027

Amin, Z., Constable, R. T., & Canli, T. (2004). Attentional bias for valenced stimuli as a function of personality in the dot-probe task. *Journal of Research in Personality, 38*(1), 15-23.

Anupama, V., Bhola, P., Thirthalli, J., & Mehta, U. M. (2018). Pattern of social cognition deficits in individuals with borderline personality disorder. *Asian Journal of Psychiatry, 33*, 105-112. doi: 10.1016/j.ajp.2018.03.010

Balter, L. J., Hulsken, S., Aldred, S., Drayson, M. T., Higgs, S., van Zanten, J. J. V., ... & Bosch, J. A. (2018). Low-grade inflammation decreases emotion recognition–Evidence from the vaccination model of inflammation. *Brain, Behavior, and Immunity, 73*, 216-221. doi: 10.1016/j.bbi.2018.05.006

Baribeau, D. A., Doyle-Thomas, K. A., Dupuis, A., Iaboni, A., Crosbie, J., McGinn, H., ... & Schachar, R. J. (2015). Examining and comparing social perception abilities across childhood-onset neurodevelopmental disorders. *Journal of the American Academy of Child & Adolescent Psychiatry, 54*(6), 479-486. doi: 10.1016/j.jaac.2015.03.016

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*(2), 241-251. doi: 10.1111/1469-7610.00715

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48. doi: http://dx.doi.org/10.18637/jss.v067.i01

Belardinelli, M. O., Huenefeldt, T., Maffi, S., Squitieri, F., & Migliore, S. (2019). Effects of stimulus-related variables on mental states recognition in Huntington's disease. *International Journal of Neuroscience, 129*(6), 563-572. doi: 10.1080/00207454.2018.1552691

Berenson, K. R., Dochat, C., Martin, C. G., Yang, X., Rafaeli, E., & Downey, G. (2018). Identification of mental states and interpersonal functioning in borderline personality disorder. *Personality Disorders: Theory, Research, and Treatment, 9*(2), 172. doi: 10.1037/per0000228

Bora, E., Bartholomeusz, C., & Pantelis, C. (2016). Meta-analysis of Theory of Mind (ToM) impairment in bipolar disorder. *Psychological Medicine, 46*(2), 253-264. doi: 10.1017/S0033291715001993

Bora, E., & Berk, M. (2016). Theory of mind in major depressive disorder: A meta-analysis. *Journal of Affective Disorders, 191*, 49-55. doi: 10.1016/j.jad.2015.11.023

Bora, E., Yucel, M., & Pantelis, C. (2009). Theory of mind impairment in schizophrenia: meta-analysis. *Schizophrenia Research, 109*(1-3), 1-9. doi: 10.1016/j.schres.2008.12.020

Bourke, C., Douglas, K., & Porter, R. (2010). Processing of facial emotion expression in major depression: a review. *Australian and New Zealand Journal of Psychiatry, 44*(8), 681-696. doi: 10.3109/00048674.2010.496359

Bruneau, E. G., & Saxe, R. (2012). The power of being heard: The benefits of 'perspective-giving' in the context of intergroup conflict. *Journal of Experimental Social Psychology, 48*(4), 855-866. doi: 10.1016/j.jesp.2012.02.017

Caldú, X., Ottino‐González, J., Sánchez-Garre, C., Hernan, I., Tor, E., Sender‐Palacios, M. J., ... & Jurado, M. Á. (2019). Effect of the catechol-O-methyltransferase Val 158Met polymorphism on theory of mind in obesity. *European Eating Disorders Review*. Advance online publication. doi: 10.1002/erv.2665

Chan, E. (2008). The roles of theory of mind and empathy in the relationship between dysphoria and poor social functioning. (Unpublished doctoral dissertation). Queen's University, Ontario, Canada.

Couture, S. M., Penn, D. L., & Roberts, D. L. (2006). The functional significance of social cognition in schizophrenia: a review. *Schizophrenia Bulletin, 32*(suppl_1), S44-S63. doi: 10.1093/schbul/sbl029

da Costa, H. P., Vrabel, J. K., Zeigler-Hill, V., & Vonk, J. (2018). DSM-5 pathological personality traits are associated with the ability to understand the emotional states of others. *Journal of Research in Personality, 75*, 1-11. doi: 10.1016/j.jrp.2018.05.001

Ekman, P. & Friesen, W. V. (1986). A New Pan-Cultural Facial Expression of Emotion. *Motivation and Emotion, 10(2)*, 159-168.

Espinós, U., Fernández-Abascal, E. G., & Ovejero, M. (2018). What your eyes tell me: Theory of mind in bipolar disorder. *Psychiatry Research, 262*, 536-541. doi: 10.1016/j.psychres.2017.09.039

Fett, A. K. J., Viechtbauer, W., Penn, D. L., van Os, J., & Krabbendam, L. (2011). The relationship between neurocognition and social cognition with functional outcomes in

schizophrenia: A meta-analysis. *Neuroscience & Biobehavioral Reviews, 35*(3), 573-588. doi: :10.1016/j.neubiorev.2010.07.001

Fossati, A., Feeney, J., Maffei, C., & Borroni, S. (2014). Thinking about feelings: Affective state mentalization, attachment styles, and borderline personality disorder features among Italian nonclinical adolescents. *Psychoanalytic Psychology, 31*(1), 41. doi: 10.1037/a0033960

Frick, C., Lang, S., Kotchoubey, B., Sieswerda, S., Dinu-Biringer, R., Berger, M., ... & Barnow, S. (2012). Hypersensitivity in borderline personality disorder during mindreading. *PloS one, 7*(8), e41650. doi: 10.1371/journal.pone.0041650

Gonzalez-Liencres, C., Shamay-Tsoory, S. G., & Brüne, M. (2013). Towards a neuroscience of empathy: Ontogeny, phylogeny, brain mechanisms, context and psychopathology. *Neuroscience & Biobehavioral Reviews, 37*(8), 1537-1548. doi: 10.1016/j.neubiorev.2013.05.001

Hanna, J. T. (2015). Ostracism Increases Positive Valence Theory of Mind Decoding Accuracy. (Unpublished undergraduate honours thesis). King's University College at Western University Canada, Ontario, Canada.

Harkness, K., Sabbagh, M., Jacobson, J., Chowdrey, N., & Chen, T. (2005). Enhanced accuracy of mental state decoding in dysphoric college students. *Cognition & Emotion, 19*(7), 999-1025. doi: 10.1080/02699930541000110

Hezel, D. M., & McNally, R. J. (2014). Theory of mind impairments in social anxiety disorder. Behavior therapy, 45(4), 530-540. doi: 10.1016/j.beth.2014.02.010

Hudson, C. C., Shamblaw, A. L., Wilson, G. A., Roes, M. M., Sabbagh, M. A., & Harkness, K. L. (2018). Theory of mind, excessive reassurance-seeking, and stress generation in

depression: A social-cognitive-interpersonal integration. *Journal of Social and Clinical Psychology, 37*(9), 725-750. doi: 10.1521/jscp.2018.37.9.725

Hünefeldt, T., Laghi, F., & Ortu, F. (2013). Are anxiously attached women better mindreaders? *Cognitive Processing, 14*(3), 317-321. doi: 10.1007/s10339-013-0556-2

Jermakow, N., & Brzezicka, A. (2016). How autistic are anorectic females? Similarities and differences between anorexia nervosa and autism spectrum disorders. *Clinical Neuropsychiatry, 13*(4/5), 53-58.

Joormann, J., & Gotlib, I. H. (2007). Selective attention to emotional faces following recovery from depression. *Journal of Abnormal Psychology, 116*(1), 80. doi: 0.1037/0021-843X.116.1.80

Kattan, D. (2018). *Aging: Implications for theory of mind* (Unpublished doctoral dissertation). State University of New York at Stony Brook, Stony Brook, NY.

Kenyon, M., Samarawickrema, N., DeJong, H., Van den Eynde, F., Startup, H., Lavender, A., ... & Schmidt, U. (2012). Theory of mind in bulimia nervosa. *International Journal of Eating Disorders, 45*(3), 377-384. doi: 10.1002/eat.20967

Kenyon, A. R., Alvares, G. A., Hickie, I. B., & Guastella, A. J. (2013). The effects of acute arginine vasopressin administration on social cognition in healthy males. *Journal of Hormones*. doi: 10.1155/2013/386306

Kometer, M., Schmidt, A., Bachmann, R., Studerus, E., Seifritz, E., & Vollenweider, F. X. (2012). Psilocybin biases facial recognition, goal-directed behavior, and mood state toward positive relative to negative emotions through different serotonergic subreceptors. *Biological Psychiatry, 72*(11), 898-906. doi: 10.1016/j.biopsych.2012.04.005

Konrath, S., Corneille, O., Bushman, B. J., & Luminet, O. (2014). The relationship between narcissistic exploitativeness, dispositional empathy, and emotion recognition abilities. *Journal of Nonverbal Behavior, 38*(1), 129-143. doi: 10.1007/s10919-013-0164-y

Kopera, M., Trucco, E. M., Jakubczyk, A., Suszek, H., Michalska, A., Majewska, A., ... & Brower, K. J. (2018). Interpersonal and intrapersonal emotional processes in individuals treated for alcohol use disorder and non-addicted healthy individuals. *Addictive Behaviors, 79*, 8-13. doi: 10.1016/j.addbeh.2017.12.006

Kynast, J., & Schroeter, M. L. (2018). Sex, age and emotional valence: Revealing possible biases in the 'Reading the Mind in the Eyes' task. *Frontiers in Psychology, 9*, 570. doi: 10.3389/fpsyg.2018.00570

Lenton-Brym, A. P., Moscovitch, D. A., Vidovic, V., Nilsen, E., & Friedman, O. (2018). Theory of mind ability in high socially anxious individuals. *Anxiety, Stress, & Coping, 31*(5), 487-499. doi: 10.1080/10615806.2018.1483021

Lischke, A., Pahnke, R., Koenig, J., Homuth, G., Hamm, A. O., & Wendt, J. (2018). COMTVal158Met genotype affects complex emotion recognition in healthy men and women. *Frontiers in Neuroscience, 12*(1007), 1-6. doi: 10.3389/fnins.2018.01007

Luminet, O., Grynberg, D., Ruzette, N., & Mikolajczak, M. (2011). Personality-dependent effects of oxytocin: greater social benefits for high alexithymia scorers. *Biological Psychology, 87*(3), 401-406. doi: 10.1016/j.biopsycho.2011.05.005

Mannava, S. (2012). Age-related differences in emotion recognition ability: Visual and auditory modalities. *Vanderbilt Undergraduate Research Journal, 8*, 1-5.

Medina-Pradas, C., Blas Navarro, J., Álvarez-Moya, E. M., Grau, A., & Obiols, J. E. (2012). Emotional theory of mind in eating disorders. *International Journal of Clinical and Health Psychology, 12*(2), 189-202.

Meijer-van Abbema, M., & Koole, S. L. (2017). After God's image: prayer leads people with positive God beliefs to read less hostility in others' eyes. *Religion, Brain & Behavior, 7*(3), 206-222. doi: 10.1080/2153599X.2016.1236033

Meyer (2009). In the eye of the beholder: Attachment style differences in emotion perception. *The Penn State McNair Journal, 16*, 74-87.

Meyer, J. K., & Morey, L. C. (2015). Borderline personality features and associated difficulty in emotion perception: An examination of accuracy and bias. *Personality and Mental Health, 9*(3), 227-240. doi: 10.1002/pmh.1299

Neal, D. T., & Chartrand, T. L. (2011). Embodied emotion perception: amplifying and dampening facial feedback modulates emotion perception accuracy. *Social Psychological and Personality Science, 2*(6), 673-678. doi: 10.1177/1948550611406138

Nejati, V. (2018). Negative interpretation of social cue in depression: Evidence from reading mind from eyes test. *Neurology, Psychiatry and Brain Research, 27*, 12-16. doi: 10.1016/j.npbr.2017.11.001

Oldershaw, A., Hambrook, D., Rimes, K. A., Tchanturia, K., Treasure, J., Richards, S., ... & Chalder, T. (2011). Emotion recognition and emotional theory of mind in chronic fatigue syndrome. *Psychology & Health, 26*(8), 989-1005. doi: 10.1080/08870446.2010.519769

Oldershaw, A., Hambrook, D., Tchanturia, K., Treasure, J., & Schmidt, U. (2010). Emotional theory of mind and emotional awareness in recovered anorexia nervosa patients. *Psychosomatic Medicine, 72*(1), 73-79. doi: 10.1097/PSY.0b013e3181c6c7ca

Overbeck, J. R., & Droutman, V. (2013). One for all: Social power increases self-anchoring of traits, attitudes, and emotions. *Psychological Science, 24*(8), 1466-1476. doi: 10.1177/0956797612474671

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual differences, 43*(3), 541-551.

Pahnke, R., Mau-Moeller, A., Junge, M., Wendt, J., Weymar, M., Hamm, A. O., & Lischke, A. (2019). Oral contraceptives impair complex emotion recognition in healthy women. *Frontiers in Neuroscience, 12*(1041), 1-9. doi: 10.3389/fnins.2018.01041

Pedreño, C., Pousa, E., Navarro, J. B., Pàmias, M., & Obiols, J. E. (2017). Exploring the components of advanced theory of mind in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 47*(8), 2401-2409. doi: 10.1007/s10803-017-3156-7

Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Retrieved from https://www.pstnet.com/

Purcell, A. L., Phillips, M., & Gruber, J. (2013). In your eyes: does theory of mind predict impaired life functioning in bipolar disorder? *Journal of Affective Disorders, 151*(3), 1113-1119. doi: 10.1016/j.jad.2013.06.051

Radke, S., & de Bruijn, E. R. (2015). Does oxytocin affect mind-reading? A replication study. *Psychoneuroendocrinology, 60*, 75-81. doi: 10.1016/j.psyneuen.2015.06.006

Reid, S. (2017). What Am I Thinking Right Now?: Social Anxiety Symptomology and Its Impact on Theory of Mind Ability. (Unpublished undergraduate honours thesis). Butler University, Indianapolis, Indiana.

Rnic, K., Sabbagh, M. A., Washburn, D., Bagby, R. M., Ravindran, A., Kennedy, J. L., ... & Harkness, K. L. (2018). Childhood emotional abuse, physical abuse, and neglect are

associated with theory of mind decoding accuracy in young adults with depression. *Psychiatry Research, 268*, 501-507. doi: 10.1016/j.psychres.2018.07.045

Sabbagh, M. A. & Bowman, L. C. (2018). Theory of mind. In S. Ghetti (Ed.) *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (4th ed.). New York: Wiley.

Sandvik, A. M., Hansen, A. L., Johnsen, B. H., & Laberg, J. C. (2014). Psychopathy and the ability to read the "language of the eyes": Divergence in the psychopathy construct. *Scandinavian Journal of Psychology, 55*(6), 585-592. doi: 10.1111/sjop.12138

Scott, L. N., Levy, K. N., Adams Jr, R. B., & Stevenson, M. T. (2011). Mental state decoding abilities in young adults with borderline personality disorder traits. *Personality Disorders: Theory, Research, and Treatment, 2*(2), 98-112. doi: 10.1037/a0020011

Segerstrom, S. C. (2001). Optimism and attentional bias for negative and positive stimuli. *Personality and Social Psychology Bulletin, 27*(10), 1334-1343.

Shaw, P., Bramham, J., Lawrence, E. J., Morris, R., Baron-Cohen, S., & David, A. S. (2005). Differential effects of lesions of the amygdala and prefrontal cortex on recognizing facial expressions of complex emotions. Journal of cognitive neuroscience, 17(9), 1410-1419. doi: 10.1162/0898929054985491

Tager-Flusberg, H. (2003). Exploring the relationship between theory of mind and social-communicative functioning in children with autism. In B. Repacholi & V. Slaughter (Eds.), *Individual Differences in Theory of Mind: Implications for Typical and Atypical Development* (pp. 197-212). New York, NY, US: Psychology Press.

Tapajoz Pereira de Sampaio, F., Soneira, S., Aulicino, A., & Allegri, R. F. (2013). Theory of mind in eating disorders and their relationship to clinical profile. *European Eating Disorders Review, 21*(6), 479-487. doi: 10.1002/erv.2247

Tay, S. A., Hulbert, C. A., Jackson, H. J., & Chanen, A. M. (2017). Affective and cognitive theory of mind abilities in youth with borderline personality disorder or major depressive disorder. *Psychiatry Research, 255*, 405-411. doi: 10.1016/j.psychres.2017.06.016

Uzefovsky, F., Shalev, I., Israel, S., Knafo, A., & Ebstein, R. P. (2012). Vasopressin selectively impairs emotion recognition in men. *Psychoneuroendocrinology, 37*(4), 576-580. doi: 10.1016/j.psyneuen.2011.07.018

Weinstein, S. R., Meehan, K. B., Cain, N. M., Ripoll, L. H., Boussi, A. R., Papouchis, N., ... & New, A. S. (2016). Mental state identification, borderline pathology, and the neglected role of childhood trauma. *Personality Disorders: Theory, Research, and Treatment, 7*(1), 61-71. doi: 10.1037/per0000139

Whissell, C. (2009). Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological Reports, 105*(2), 509-521. doi: 10.2466/PR0.105.2.509-521

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin, 124*(3), 283. doi: 10.1037/0033-2909.124.3.283

Yu, R. L., Chen, P. S., Tu, S. C., Tsao, W. C., & Tan, C. H. (2018). Emotion-Specific Affective Theory of Mind Impairment in Parkinson's Disease. *Scientific Reports, 8*(1), 16043. doi: 10.1038/s41598-018-33988-6

Zainal, N. H., & Newman, M. G. (2018). Worry amplifies theory-of-mind reasoning for negatively valenced social stimuli in generalized anxiety disorder. *Journal of Affective Disorders, 227*, 824-833. doi: 10.1016/j.jad.2017.11.084

Table 1.

RMET items organized by valence.

Note. Items are organized based on the continuous analysis conducted in the current study from the most positively valenced item to the most negatively valenced item according to the bootstrapped *t*-statistic. The coloured cells depict the categorization of items as positive (blue), neutral (green), negative (red). Cells that are black were not classified.

P = Participants rated the valence of the picture without the corresponding target response option.

R = Participants rated the valence of the target response option without the corresponding picture.

P + R = Participants rated both the picture and the target response option.

n = number of participants used to create the valence categories. When *n* = 0, the authors determined the valence categories rather than a sample of participants.

k = number of studies that have used this valence categorization scheme.

playful comforting

irritated bored

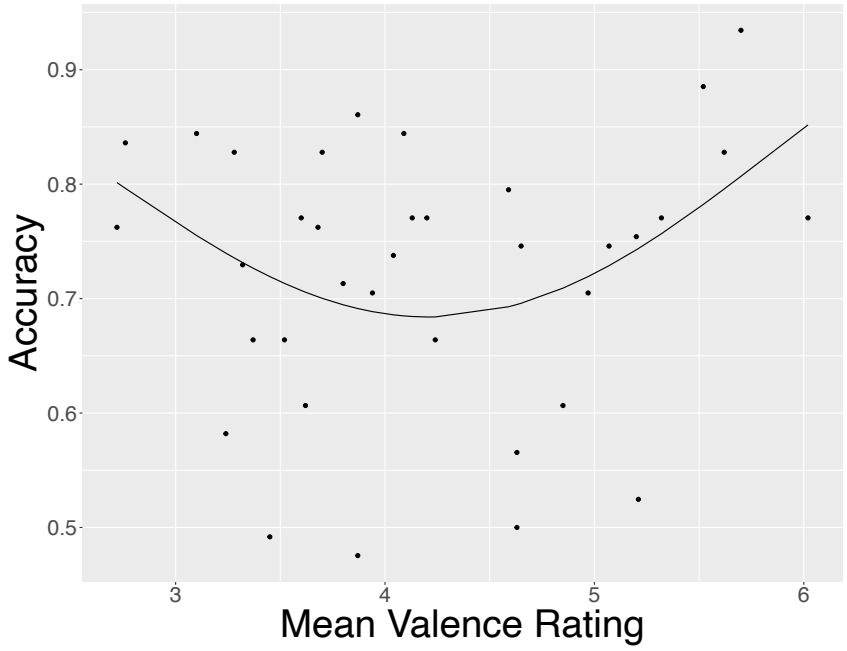
Figure 1. Sample eye from the Reading the Mind in the Eyes task (Baron-Cohen et al., 2001). Obtained from the Autism Research Centre Downloadable Tests website (https://www.autismresearchcentre.com/arc_tests)

Figure 2. Valence ratings from Study 1, sorted by mean valence rating. From left to right, the figure depicts the item number and target response adjective, the distribution of participant responses, the mean valence rating, the median valence rating, the average bootstrapped *t*-value (mean response for each item compared to the neutral score), and the bootstrapped 95% confidence intervals (n = 60) for the mean valence ratings.

Figure 3. Mean accuracy on the Reading the Mind in the Eyes Task stratified by valence categories (negative, neutral, positive). Errors bars represent 95% confidence intervals. *Repeated-measures ANOVA significant at $F > 5.29$, $p < .004$. Connecting lines illustrate significant pairwise comparisons.

Figure 4. Mean response time on the Reading the Mind in the Eyes Task stratified by valence categories (negative, neutral, positive). Errors bars represent 95% confidence intervals. *Repeated-measures ANOVA significant at $F > 3.02$, $p < .05$. Connecting lines illustrate significant pairwise comparisons.

(a)



(b)

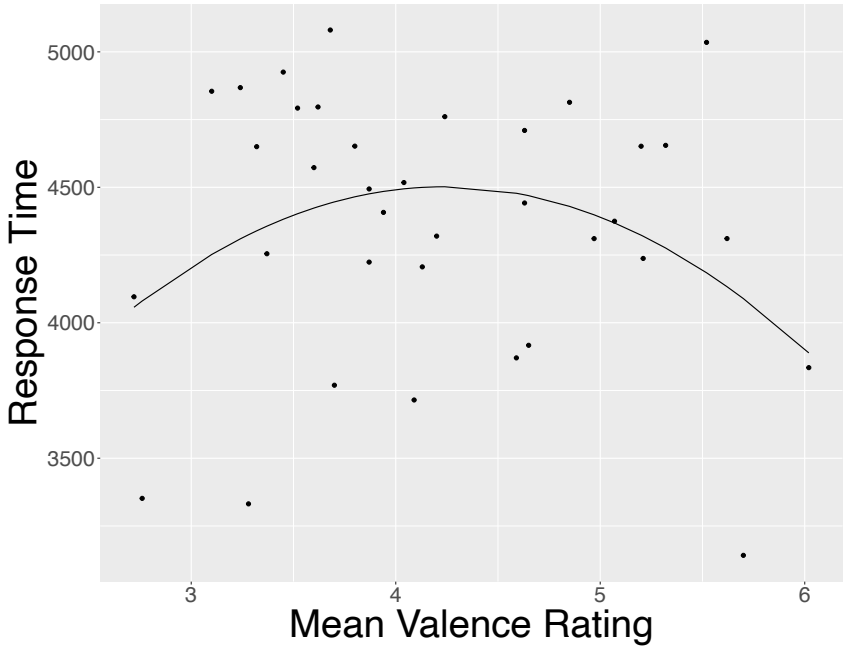


Figure 5. The relation between mean valence rating and (a) accuracy and (b) response time of each item. Solid lines represent model predictions from the mixed-model regression equation fit in Study 2.