The Challenges of Measuring Empathic Accuracy: A Mentalizing Versus Experience-Sharing Paradigm

Short title: *MEASURING EMPATHIC ACCURACY*

Rose Turner*[1] and Frédéric Vallée-Tourangeau[2]

[1] University of the Arts London

[2] Kingston University

*Corresponding author information: Rose Turner, Science Programme, UAL: London College of Fashion, 20 John Prince's Street, London, UK, W1G 0BJ (e-mail: R.turner@fashion.arts.ac.uk).

Abstract

Empathic accuracy, the ability to accurately infer the mental states of others, is essential to successful interpersonal relationships. Perceivers can interpret targets' emotional experiences by decoding facial and voice cues (mentalizing) or by using their own feelings as referents (experience-sharing). We examined the relative efficacy of these processes via a replication and extension of Zhou et al. (2017) who found experience-sharing to be more successful but undervalued. Participants estimated targets' emotional ratings in response to positive, neutral and negative images in mentalizing or experience-sharing conditions. Our analysis of absolute magnitudes of error showed similar levels of accuracy across process conditions (a non-replication of Zhou et al.); however, our exploratory analysis of directional variation across valence using raw scores revealed a pattern of conservative estimates for affective stimuli, which was accentuated in the mentalizing condition. Thus, our exploratory analysis lends conceptual support to Zhou et al.'s finding that experience-sharing represented the more successful process, and we replicated their finding that it was nevertheless undervalued. Extending Zhou et al., we also found that empathic accuracy was predicted by individual differences in fiction-exposure. Future research may further examine the impact of individual differences and stimulus properties in the employment of empathic inferencing strategies.

Keywords: MENTALIZING, EXPERIENCE-SHARING, SIMULATION, SOCIAL COGNITION, EMPATHIC ACCURACY, PERSPECTIVE-TAKING

Data availability statement: The data that support the findings of this study are openly available on the Open Science Framework at https://osf.io/64gs9/?view_only=748135ecf70a49069bec3c59d1da95d6 and https://osf.io/7bphr/?view_only=d3652234de8d4aac97e401d033a1822a

Acknowledgements: This research was funded by a Qualtrics Academic Bursary. R. Turner acknowledges the support of a Kingston University PhD studentship award. The authors thank Wendy Ross for helping to develop and run the linear mixed effects analysis.

The Challenges of Measuring Empathic Accuracy: A Mentalizing Versus Experience-Sharing Paradigm

The ability to accurately infer the internal experiences of another person represents a core social cognitive skill and essential component of intersubjectivity. It enables behaviour prediction (Nichols & Stich, 2003) and is associated with positive relationships (Castano, 2012), interpersonal cooperation and prosocial behaviour (Batson et al., 1981; Paal & Bereczkei, 2007). The skills involved, such as the ability to interpret voice tone, gesture and facial cues, and to integrate prior knowledge about a situation, typically develop in childhood (e.g., Perner & Wimmer, 1985) and so research has largely focused on children and groups with characteristic deficits. However, these skills vary between neurologically typical adults, continue to evolve through the lifespan (Duval et al., 2010; Happé, et al., 1998; Maylor et al., 2002), and can be enhanced through training (Teding van Berkhout & Malouff, 2016). Thus, the operationalisation, measurement and exploration of these abilities in typical adult populations represents an important area of enquiry (Turner & Felisberti, 2017).

Processes in Empathic Accuracy

There has been much debate surrounding the extent to which imaginative capabilities versus the process of instantiating another's inner state are required for empathy (e.g., Gallagher & Gallagher, 2019). Reviewing research on the neuroscience of empathy, Zaki and Ochsner (2012) modelled two paths to understanding others' internal states incorporating both domains: mentalizing (explicitly interpreting verbal and nonverbal cues) and experience-sharing (vicariously sharing in the target's experience), with each comprising a range of sub-processes. This model provides a framework for understanding situations which may be more likely to elicit mentalizing than experience-sharing and vice versa; the model also facilitates the exploration and measurement of these abilities. The first domain, mentalizing, includes theory of mind, perspective-taking and cognitive empathy. Imagine the host of a party

receiving a gift: one can work out whether they liked the gift by interpreting the host's verbal and non-verbal cues and by using prior knowledge about the recipient specifically, and gift-receiving reactions in general. Under experimental conditions, this mentalizing ability can be tested using facial expressions tasks such as the Reading the Mind in the Eyes Test (RMET, Baron-Cohen et al., 2001) or by assessing the ability to interpret a series of vignettes or brief narratives (e.g., Birch & Bloom, 2007; Shamay-Tsoory & Aharon-Peretz, 2007). An alternative path, experience-sharing (the second domain), refers to the tendency to engage the same neural systems when observing a state as when experiencing it first-hand. This path involves the perceiver using their own feelings, either about the gift itself or the host's reaction, as the basis for interpreting the host's experience; putting oneself in the target person's "shoes" and experiencing resonance (the engagement of overlapping neural systems when observing a target's emotional state; Zaki & Ochsner, 2012). This experience-sharing component can be measured via self-reports about vicariously experiencing others' affective states (e.g., the Empathy Quotient; Baron-Cohen & Wheelwright, 2004).[1] The distinction between mentalizing and experience-sharing is a question of *what* the host feels, versus *how* the host feels it: experience-sharing requires not only a functional understanding of the mental state of the target, but a matching of that state (see Smith, 2017).

Both mentalizing and experience-sharing can lead to prosocial concern, the third facet of Zaki and Ochsner's (2012) model, which is associated with prosocial behaviour. Here, the party host may be disappointed by the gift, and one might wish to alleviate their negative affect by providing a distraction, offering support or a replacement gift.[2] In the laboratory, prosocial behaviour may be tested through dictator games, in which participants decide how

[1] In this conception of empathy, perceivers maintain an awareness of the source of their emotion and self/other distinction (Decety & Lamm, 2006), which distinguishes it from related constructs such as emotion contagion (which could be tested using facial expression indexing, e.g., Olszanowski et al., 2019).

[2] Some researchers have conceptualised the motivational component as a result of concern (or sympathy) but not as part of it (e.g., Batson et al., 1981). Others agree with the idea that motivational concern is preceded by a cognitive (mentalizing) component (e.g., Baron-Cohen & Wheelwright, 2004).

to distribute cash sums between themselves and other players (for an overview see Camerer, 2003) or through a measure of helpful behaviour (such as whether participants pick up a pen that the experimenter has ostensibly dropped by accident; e.g., van Baaren et al., 2004). Via this route, either mentalizing or experience-sharing could ultimately result in prosocial behaviour. This is in line with the empathy-altruism hypothesis, which suggests that feeling concern for another can initiate an altruistic response (Batson, 1987, 2011; Batson et al., 1981; for an evaluation of research see Batson et al., 2015).

This model has received neuroscientific support with mentalizing and experience-sharing processes shown to initiate prosocial behaviour (Zaki & Ochsner, 2012). However, context has been found to impact the extent to which these processes are engaged, and their effects. For example, when people respond to explicit questions about targets' internal states, activity in brain areas associated with mentalizing predicts helping (Harbaugh et al., 2007), whereas when watching a target in pain, activity in areas associated with experiencing that pain predicts helping (Hein et al., 2010). Thus, mentalizing and experience-sharing processes appear to represent two dissociable routes to understanding another's internal state which, in turn, can initiate prosocial concern and behaviour.

Zhou et al. (2017)

We use "empathic inferencing" (Ickes, 1997) to refer to the process, or set of processes, through which a person makes sense of another's thoughts and feelings. "Empathic accuracy" represents the measure of one's skills in empathic inferencing; in other words, the extent to which the inference accurately reflects the thoughts or feelings of the target.[3] Zhou et al. (2017) developed an experimental paradigm with which to index and compare mentalizing and experience-sharing processes for empathic accuracy (or "reading" versus "being"; Zhou

[3] Zaki and Ochsner's (2012) model did not incorporate the accuracy component, though we use it here in order to distinguish inferencing process from the measurable component or outcome variable.

et al., p. 482). In a series of four experiments, participants were asked to estimate the emotional ratings that target individuals ("experiencers") had previously given in response to a range of positive, negative and neutral photographs. Participants were assigned to either the "theorization" condition, where they watched short videos of the experiencers' dynamic facial responses to the photographs, a "simulation" condition, where they viewed the same photographs as the "experiencers" and were able to use their own reactions as referents, or a "simultaneous" condition in which the photographs and videos were presented side-by-side. Consequently, the theorization and simulation stimuli were considered the basis for mentalizing and experience-sharing exercises, respectively. In some experiments, participants were further split into "bound" or "free choice" conditions. In the former they were assigned to their condition, whereas in the latter they were invited to select their preferred (theorization or simulation) condition following video training on each method.

Results showed that using simulation led to higher empathic accuracy compared to theorization, yet participants tended to overestimate the insight gained through theorization compared to simulation. When financially incentivised to perform well (in Experiments 2-4 participants were informed that they would receive additional payment if their performance reached the 80th percentile) those in the free-choice condition tended to self-select into the less effective theorization group. Not only were the two processes for interpreting mental states shown to be unequal, participants also misjudged their relative utility. Zhou et al. (2017) suggested that participants' reluctance to use their own experience as a guide for estimating someone else's was analogous to findings from the field of affect forecasting where participants tend to under-appreciate the value of using another person's experience as a guide for their own (Gilbert et al., 2009). However, simulation may not represent the most useful process across all situations (Zhou et al., 2017; see also, Barrett et al., 2011). Rather than

attempt to examine an exhaustive list of potentially influential contextual factors, we now turn to three which we theorise may impact these processes.

Potential Sources of Variation

Identification with Target

Zhou et al. (2017) suggested that participants' lower confidence in their ability to infer mental states from having the same experience as a target rather than reading their facial expressions may be due to the tendency to overestimate dissimilarity between self and others. Indeed, research has shown that people are less likely to adopt experience-sharing with targets that they perceive as dissimilar to themselves (Hein et al, 2010; Zaki & Ochsner, 2012), potentiating ingroup advantages in accuracy (Adams et al., 2010; Matsumoto et al., 2009). Therefore, identification with experiencers represents a potential source of variance both in the successful engagement of experience-sharing versus mentalizing, as well as in their perceived value.

Valence

Zhou et al.'s (2017) stimuli varied by valence (participants viewed positive, negative and neutral images) although this was not a focus of their analysis. Studies have found that people show attentional bias towards emotionally salient information ("affect-biased attention"; Humphrey et al., 2012; Todd et al., 2012); that positive and negative emotion expressions are recognised at different speeds (e.g., Leppänen & Hietanen, 2003), are associated with differences in the perceiver's neural activity (Kilts et al., 2003), and result in different biases (Kauschke et al., 2019); and research has revealed interactions between emotionally-valenced stimuli and psychiatric disorders (e.g., Surguladze et al., 2004; Unoka et al., 2011). Zhou et al. used absolute scores to index empathic accuracy, alongside aggregated correlations between participant and target ratings, and so it was not possible to determine the direction of effects; in other words, whether estimates were positively or

negatively biased and whether this depended on the valence of the target emotion. Closer examination of raw (positively and negatively signed) scores would extend Zhou et al. by facilitating such an analysis.

Individual Differences in Fiction-exposure

Frequent fiction-readers are regularly exposed to the emotional states of characters whose circumstances and experiences may be very different to the reader's own. A growing body of research has indicated that some variation in empathic accuracy appears to be accounted for by lifetime exposure to fiction (for a meta-analysis, see Mumper & Gerrig, 2017). However, this field of research, wherein the parameters of theoretical models are currently being developed (e.g., Black et al., 2021), has yet to distinguish between fiction's effects on mentalizing versus experience-sharing processes. If readers become practiced at using their own mental apparatus to make sense of characters' experiences—using their own beliefs, motivations and emotions as proxy for the characters'—they may both hone and trust in their ability to recognise mental states based on their own emotional responses. As Zhou et al.'s (2017) results suggest that experience-sharing is the more beneficial approach to accurately interpreting the emotional states of targets, fiction-exposure may be associated with both the ability to accurately interpret emotions using this faculty, and the willingness to engage it.

The Present Study

According to Zhou et al. (2017, Experiments 1-2), experience-sharing represents the most effective approach for interpreting the emotions of another person, despite participants endorsing mentalizing as the better empathic inferencing method. This finding has implications for initiatives aimed at developing social cognitive skills—the focus should be on cultivating experience-sharing over mentalizing ability—and so it warrants replication. Furthermore, it remains unclear how far contextual factors may influence the respective

efficacy of these processes. Thus, we attempted to replicate Zhou et al.'s main findings that experience-sharing (simulation) leads to higher empathic accuracy compared to mentalizing (theorization) but that people tend to under-value experience-sharing comparatively. We extended the study by examining potential sources of variation (stimuli valence, identification with experiencers and individual differences in fiction-exposure). The specific aims of this study were: (a) to replicate Zhou et al.'s findings that experience-sharing leads to higher empathic accuracy compared to mentalizing, and that (b) people tend to undervalue experience-sharing, (c) to account for variance arising from properties the stimuli (valence and experiencer), and (d) to predict empathic accuracy and preference for inferencing process from individual differences in lifetime fiction-exposure. The first two aims represent replications of Zhou et al., whereas the latter two extend their research.

We diverged from Zhou et al.'s (2017) procedure in four ways (and so our replication is conceptual rather than exact): first, we drew from a different image database when constructing our stimuli (detailed under Materials). Second, our stimuli comprised trials across six target experiencers which were presented to all participants (i.e,, a repeated measures factor) in order to analyse identification with experiencers (Zhou et al.'s participants were assigned to one of twelve targets and this source of variance was not analysed). Third, we measured process preference retroactively, rather than allowing participants to self-assign into conditions. Fourth, in line with Zhou et al., we computed overall differences in empathic accuracy using absolute values, and we additionally conducted the same analyses using raw scores in order to establish differences in the direction of errors. On the one hand, converting scores on the dependent variable into absolute values would enable levels of error to be compared across conditions but would result in a compression of the variance associated with the full range of positive and negative responses. On the other hand, reflecting the full scale via raw scores would enable the direction of error

to be interpreted, elucidating any positive or negative biases, but would risk value signs cancelling out across aggregate means. Therefore, both absolute and raw values are useful in this context: absolute scores represent the magnitudes of effects, whereas raw scores reveal the general direction of effects. Our exploratory analysis using raw scores is aimed at a more nuanced understanding of empathic accuracy as a function of empathic process and valence.

Method

Participants

This study received a favourable opinion from the [redacted] ethics committee and complied with the British Psychological Society's standards for the treatment of human participants. Zhou et al. (2017) based their sample size on the simple heuristic that each "experiencer" was paired with at least two participants in each process condition. This approach yielded a large effect of condition, $t(70) = 7.26$, $p < .001$; they reported a common language effect size (CL) of 92.3% for the difference between theorization and simulation with 24-25 participants per group (Experiment 1, total $n = 73$).[4] Mumper and Gerrig's (2017) meta-analysis correlating fiction-exposure with RMET (93%) or actor-intention vignette scores (7%), $r = .21$, $p < .001$, $d = .43$, two-tailed, $\alpha = .05$, indicated that a sample size of 200 would be required to detect the relationship between fiction-exposure and empathic accuracy task performance at > 80% power ($N = 200$ for .86 power).[5] Two-hundred and thirteen participants were recruited via the Qualtrics platform in return for financial compensation. Sample size was determined a priori and no interim analyses were conducted during data collection. However, inspection of the dataset revealed that the sample consisted entirely of females due to a fault in the online screening logic and so an additional 120 male participants

[4] The CL statistic (McGraw & Wong, 1992) reflects the probability that a score sampled at random from one distribution would be greater than that of another distribution. The value here indicates that a score selected from the simulation condition would be higher than a score selected from the theorization condition in 92.3 out of 100 cases.

[5] Mumper and Gerrig (2017) drew on published and unpublished studies, finding no evidence of a "file-drawer" effect (see Rosenthal, 1979).

were recruited (total N = 324). Four participants were excluded due to reporting a technical (stimulus loading) error, three for straight-lining (selecting the same response throughout the study), and fifteen for response durations greater than three standard deviations from the mean, resulting in a final sample size of 302. This sample consisted of 100 males (33%) and 202 females (67%) aged 18-86 (M = 50.2, SD = 15.8), resident in the UK (34%), USA (37%) or Canada (29%).

Materials

Development of Empathic Accuracy Test Stimuli

In line with Zhou et al. (2017), a group of students (five females, three males, aged 23-45 [M = 31.83, SD = 9.28]) were recruited to act as target "experiencers" voluntarily or in return for course credit if applicable. They were informed that their data would be used as stimuli in future experiments aiming to "explore how adults can improve their abilities to understand what other people are thinking and feeling". The experiencers were asked to rate their emotional responses to 60 positive, negative and neutral pictures (20 each) from the Geneva Affective Picture Database (GAPED; Dan-Glauser & Scherer, 2011) on a 9-point Likert-style scale ranging from "very negative" to "very positive" (the same scale labels as Zhou et al., 2017, although they used the International Affective Picture System database; Lang et al., 2008). The images were presented in a random order, each appearing for 7 seconds. With the participants' permission, their faces were filmed throughout the task and the computer screen was simultaneously recorded. The two video streams were aligned using video editing software (Final Cut Pro) and edited into 60 5-second clips of the participants' facial reactions to each picture (each clip was extracted from the point at which the stimulus image had loaded). These clips, presented alongside a still image of the experiencer, formed the stimuli for the theorization condition, and the GAPED pictures, presented alongside the same still of the experiencer, formed the stimuli for the simulation condition (Figure 1). The

still images of the experiencers were neutral screenshots taken prior to each experiencer's exposure to the images, and they remained the same across each trial. They were included in both conditions so that participants were equally aware of the experiencers' demographic characteristics. Each stimulus set consisted of 60 trials across six experiencers (four females and two males, aged 19-45)[6], with 10 trials per experiencer, plus five trials across two "practice" experiencers (one male aged 40 and one female aged 65 for which data were not collected), in both conditions.

Figure 1. Screenshot of the theorization (mentalizing) condition practice stimuli (left panel) and the simulation (experience-sharing) condition practice stimuli (right panel). Right panel GAPED image adapted from Université de Genéve Research Material, licensed by Creative Commons (CC BY-NC-SA 3.0: https://creativecommons.org/licenses/by-nc-sa/3.0/).

Empathic Accuracy Test

Participants were asked to predict the emotional ratings of six target experiencers in response to 60 pictures (described above). The distribution of positive, negative and neutral images across the six experiences is shown in Figure 2. Process condition (theorization versus simulation) varied between subjects: participants were randomly assigned to either the simulation condition, in which they viewed the same pictures as the experiencers, or the

[6] The gender imbalance was due to the availability of participants to take the role of experiencer and the two practice experiencers being selected as such due to technological problems resulting in incomplete datasets for those experiencers.

theorization condition, in which they viewed video footage of the experiencers reacting to the pictures. Participants in both conditions were given the following instruction:

> "In a previous study, we asked the experiencers to look at some pictures. We asked them to tell us how they felt about each picture on a scale from "extremely negative" to "extremely positive". We call this their "emotional rating". Your task is to estimate the experiencer's emotional rating for each picture that they viewed, using the same scale."

After each trial, participants were asked two questions: first, "what was your emotional rating for this clip of the experiencer?/picture?" Second, "what was the experiencer's emotional rating for this picture?" To both questions, participants responded using the same scale that the experiencers had used. Differences between responses to the second question and the experiencers' actual ratings constituted the dependent variable (α = .87).

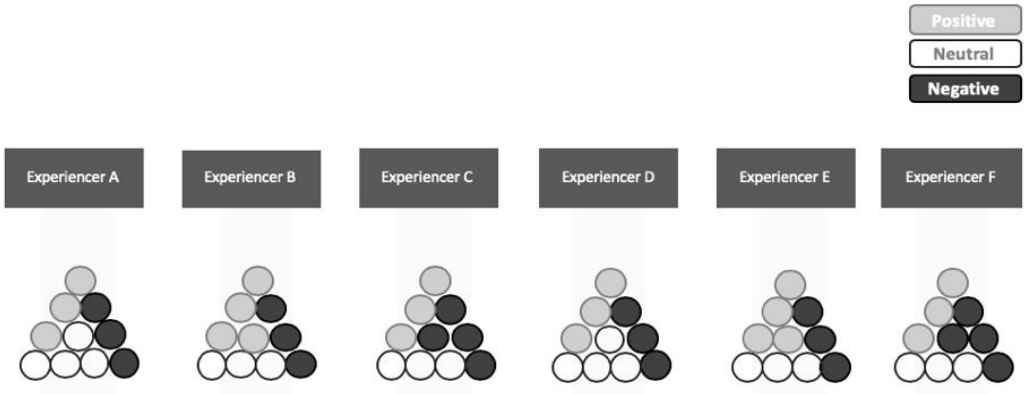

Figure 2. Distribution of positive, negative and neutral images across the six experiencers.

Identification with Experiencer

Before each experiencer block, participants viewed a still image of the experiencer and were asked to indicate how similar they considered the experiencer to be to themselves on a single item sliding scale from "we are completely different" (0) to "we are completely alike" (100).

Retroactive Group Preference

Following completion of the empathic accuracy test, participants were informed that there were two versions of the task and presented with examples of each process condition via looped videos taken from the practice experiencer sets. They were then asked the question, "which option do you think would be the most successful in enabling you to estimate the experiencers' ratings as accurately as possible?" Preference was recorded as a single dichotomous "theorization" vs. "simulation" item.

Fiction-exposure

We used a version of the Author Recognition Test (ART; Stanovich & West, 1989) to determine general exposure to fictional stories. The ART provides a proxy measure of lifetime fiction-exposure by testing participants' recognition rates of author names; familiarity with author names is assumed to reflect having read the author's work or having browsed related works. The test controls for socially desirable or indiscriminate responding by incorporating foils and participants are explicitly told that points are deducted for selecting false alarms. We used a revised and updated version of the ART (Mar et al., 2006; Turner & Vallée-Tourangeau, 2020; see supplemental materials), which incorporated ten genres across mutually exclusive fiction and nonfiction categories (five fiction, e.g., romance, thriller; and five nonfiction, e.g., science, business). The inclusion of the nonfiction dimension allows nonfiction-exposure to be controlled, as fiction and nonfiction-exposure tend to correlate (Mar et al., 2006, 2009). The current version comprised 55 fiction names, 55 nonfiction names, and 40 foils (150 items total; α = .94).

Demographic and Control Variables

Data were gathered on gender, age and level of education, and the seven-item perspective-taking subscale of the Interpersonal Reactivity Index (IRI; Davis, 1980) provided

a measure of the dispositional tendency to consider other people's points of view (e.g., "I try to look at everybody's side of a disagreement before I make a decision"; α = .77).

Procedure

The study was administered in Qualtrics. After providing consent and demographic information, participants completed the updated ART, perspective-taking scale, and empathic accuracy test (in their randomly assigned condition) and answered the group preference question. With the exception of the group preference measure, which always followed the empathic accuracy test, the order of tasks was randomized, and task items were internally randomized. In the empathic accuracy test, trials were randomized within experiencer blocks, which were randomized within conditions.

Computation and Data Analysis

Empathic accuracy scores were calculated by subtracting the target experiencer's actual rating from the participant's rating for the experiencer, for each trial, averaged across the 60 trials. Therefore, a value of zero indicates no difference between participants' estimates and experiencers' actual ratings; values closer to zero indicate higher empathic accuracy and larger values indicate lower empathic accuracy. Absolute difference values provide the magnitudes of errors (differences between participant ratings for experiencers and experiencers' own ratings) in line with Zhou et al.'s (2017) approach, and raw difference scores establish the general directions of errors. Raw negative values indicate that participants' average estimates were over-negative in comparison to the experiencers' own ratings, whereas raw positive values indicate that the participants' average estimates were over-positive. We first examined variance in difference scores using *t*-tests and analyses of variance (ANOVAs), consistent with Zhou et al.; all ninety-five percent confidence intervals were bias-corrected and accelerated using bootstrapping (N = 1000). Then, we fitted a linear mixed model (LMM) to examine effects of process, valence and experiencer, while

accounting for random stimulus and participant variation within nested clusters. Both the horizontal and vertical datasets are provided via the Open Science Framework; the latter including the R code (R Core Team, 2021) for the LMMs.

Results

Average absolute and raw difference scores, ART and perspective-taking scores are presented in Table 1. The theorization group reported marginally higher perspective-taking tendencies compared to the simulation group, $p = .047$, but accounting for the error rate associated with using both absolute and raw scores renders this effect non-significant. Participants recognised more fiction than nonfiction authors in both conditions (in line with the ART scores reported by Mar et al., 2006, although scores were lower than those reported in Mar et al.'s study), and foil-selection was low.

Table 1.

Means (and Standard Deviations) and Their 95% Confidence Intervals for Overall Scale Scores and for Each Process Condition

Measure	**Grand Mean**	**Simulation**	**Theorization**	***t***	***p***
Absolute difference scores	1.45 (0.32) [1.42, 1.49]	1.42 (0.32) [1.37, 1.47]	1.48 (0.32) [1.43, 1.53]	-.57	.19
Raw difference scores	0.01 (0.48) [-0.05, 0.06]	0.12 (0.44) [0.05, 0.19]	-0.11 (0.50) [-0.19, -0.02]	4.13	< .001
Perspective-taking	16.90 (5.07) [16.31, 17.51]	16.34 (5.29) [15.57, 17.16]	17.50 (4.77) [16.71, 18.35]	-2.0	.05
Fiction-exposure	9.38 (8.55) [8.46, 10.37]	8.69 (8.06) [7.47, 9.97]	10.11 (9.00) [8.73, 11.48]	-1.44	.15
Nonfiction-exposure	3.45 (4.06) [2.03, 3.93]	3.25 (3.97) [2.65, 3.93]	3.67 (4.14) [3.01, 4.39]	-.90	.37
Foil selection	0.41 (0.92) [0.31, 0.52]	0.35 (0.76) [0.25, 0.48]	0.47 (1.07) [0.31, 0.67]	-1.08	.29

Note. 95% confidence intervals (bias-corrected and accelerated using bootstrapping $N = 1000$) are presented in brackets.

Gender

Participant gender did not significantly impact difference scores in either the simulation or theorization conditions, all *p*s > .05. Inclusion of the perspective-taking covariate did not alter this result.

Identification with Experiencer

The hypothesis that empathic accuracy would differ as a function of identification with experiencer was not supported. Identification was not predicted by difference scores in either process condition (all *B*s < +/-.004, all *p*s > .1). Modelling the interaction effect of process condition and (aggregate) identification with experiencer scores on absolute difference scores yielded a non-significant model, $F(3, 297) = 1.75$, $p = .16$, in which both predictors and their interaction were non-significant (all *B*s < +/-.06, all *p*s > .1); furthermore, when analysing identification with each experiencer individually, all models and predictors including interaction terms were non-significant (all *F*s < 2.41, all *B*s < +/-.057, all *p*s > .067). Identification was also not associated with differences between participants' ratings for their own emotional responses and for those of the experiencers (with the exception of Experiencer A in the theorization condition, $r(145) = .22$, $p = .008$, 95% CI [.05, .38]), indicating that identification generally did not lead participants to score that experiencer's rating as similar to their own. This variable was therefore dropped from further analysis.

Empathic Inferencing Process

Independent *t*-tests using the absolute and raw difference scores addressed the hypothesis that there would be an effect of inferencing process on empathic accuracy with participants showing greater accuracy in the simulation condition compared to the theorization condition. As shown in Table 1, using absolute difference scores yielded a non-significant effect of inferencing process, indicating that the average error was similar in magnitude across the two conditions. However, the effect on raw difference scores was

statistically significant, such that simulation condition estimates tended to be over positive (M = .12, SD = .44), whereas theorization condition estimates tended to be over-negative (M = -.11, SD = .50), with a mean difference of .22, $t(300) = 4.13$, $p < .001$, 95% CI [.12, .33]. While this represents a moderately small effect in the context of the 9-point response scale, scores would have been compressed as a result of positive and negative values cancelling out across the dataset and so this statistically significant difference should not be discounted. The assessment of the direction of error using raw scores supported the prediction that accuracy would differ between process conditions, but it did not replicate Zhou et al.'s (2017) finding that accuracy was greater in the simulation condition; rather, the direction and not the magnitude of errors differed.

Interaction with Valence

Figure 3 shows interactions between inferencing process and valence. Using the absolute values revealed a significant main effect of valence, $F(2, 600) = 387.82$, $p < .001$, $\eta_p^2 = .56$, supporting the hypothesis that accuracy would differ as a function of valence, and a non-significant main effect of process condition, $F(1, 300) = 2.47$, $p = .12$, $\eta_p^2 = .01$. There was a small, significant interaction between process condition and valence, $F(2, 600) = 6.16$, $p = .002$, $\eta_p^2 = .02$; however, simple effects analysis revealed that accuracy only varied for neutral stimuli, wherein the simulation condition yielded greater accuracy with a mean difference of -0.20 95% CI [-0.30, -0.11], $F(1, 300) = 14.94$, $p < .001$, $\eta_p^2 = .05$.

Raw scores provided a more detailed picture. The main effects of valence, $F(1.26, 372) = 718.35$, $p < .001$, $\eta_p^2 = .71$, and process condition, $F(1, 295) = 17.10$, $p < .001$, $\eta_p^2 = .06$, were significant, again supporting the hypothesis that valence would have an effect on accuracy. The significant interaction, $F(1.26, 372) = 64.79$, $p < .001$, $\eta_p^2 = .18$, showed that estimates tended to be conservative for both positively- and negatively-valenced stimuli (i.e., over-negative for positive stimuli and over-positive for negative stimuli) and that this effect

was greater in the theorization condition. Across positive trials simulation scores were 0.58, 95% CI [0.39, 0.75] closer to zero (and thus more accurate) than theorization scores, $F(1, 298) = 37.67$, $p < .001$, $\eta_p^2 = .11$, and across negative trials simulation scores were 0.56, 95% CI, 95% CI [0.37, 0.73] closer to zero, $F(1, 297) = 34.62$, $p < .001$, $\eta_p^2 = .10$. For neutral trials the mean difference between process conditions was 0.65, 95% CI [0.50, 0.80], $F(1, 300) = 87.02$, $p < .001$, $\eta_p^2 = .23$, but taking the directional difference into account (theorization scores were over-negative and simulation scores were over-positive), simulation scores were 0.22 closer to zero, and thus marginally more accurate. Therefore, our examination of directions of error across the valence variable using raw scores revealed conservative biases in response to affective stimuli that were accentuated in the theorization condition and, further, that simulation led to greater accuracy overall.

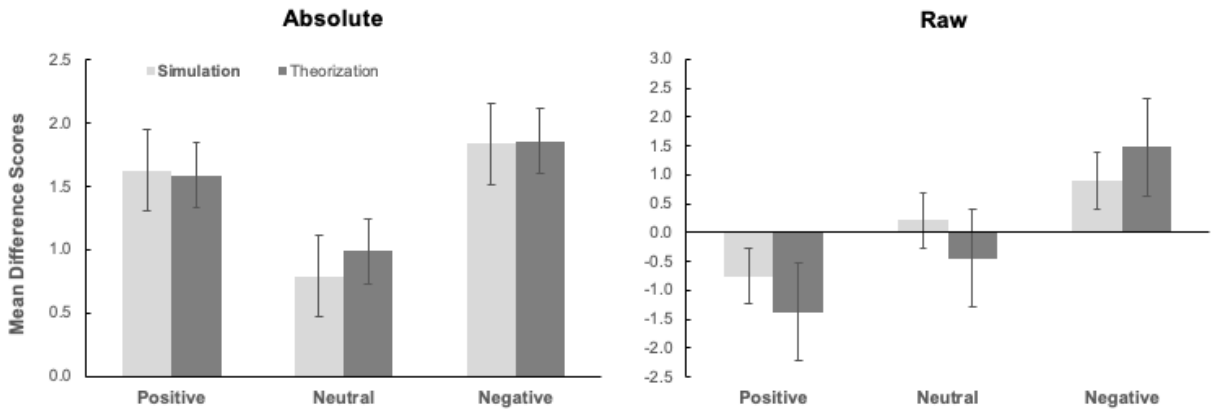

Figure 3. Absolute (left panel) and raw (right panel) difference scores grouped by inferencing process and valence. Error bars are standard errors of the mean. See supplemental tables 1-2 for exact values and associated significance.

Process-Valence-Experiencer Interactions

A further, granular analysis was conducted to assess of the direction of effects as a function of process, valence and experiencer combined. The effects of inferencing process, experiencer and valence were analysed using two 2(process: simulation, theorization) × 6(experiencer: A, B, C, D, E, F) × 3(valence: positive, neutral, negative) mixed analyses of

variance on the absolute and raw difference scores (see Figure 4 and supplemental Tables 1-2).[7] The same pattern of results was found using both sets of scores: using the absolute scores revealed significant main effects of process, $F(1, 300) = 4.42$, $p = .036$, $\eta_p^2 = .015$, valence, $F(2, 600) = 375.79$, $p < .001$, $\eta_p^2 = .56$, and experiencer, $F(4.22, 1265) = 131.79$, $p < .001$, $\eta_p^2 = .31$, and a significant three-way interaction effect, $F(7.62, 2286) = 39.60$, $p < .001$, $\eta_p^2 = 12$. In light of the three-way interaction, the data were then split by condition to assess the simple two-way interactions. Using absolute scores, the interaction between experiencer and valence was statistically significant in both the simulation, $F(7.08, 1089) = 12.34$, $p < .001$, $\eta_p^2 = .07$ and theorization, $F(6.60, 962) = 98.02$, $p < .001$, $\eta_p^2 = .40$, conditions. Using raw scores also revealed significant main effects of process, $F(1, 300) = 21.08$, $p < .001$, $\eta_p^2 = .07$, valence, $F(1.27, 379) = 745.49$, $p < .001$, $\eta_p^2 = .71$, and experiencer, $F(4.29, 1286) = 38.18$, $p < .001$, $\eta_p^2 = .11$, and a three-way interaction effect, $F(7.58, 2272) = 34.10$, $p < .001$, $\eta_p^2 = .102$. Again, simple effects analysis revealed significant interactions between experiencer and valence in both the simulation condition, $F(6.29, 968) = 116.12$, $p < .001$, $\eta_p^2 = .43$, and the theorization conditions, $F(8.49, 1240) = 317$, $p < .001$, $\eta_p^2 = .69$.

Multilevel Linear Mixed Model

Our analysis using *t*-tests and ANOVAs was in line with Zhou et al. (2017) and the three-way interaction appeared to reveal effects of experiencer (raising questions about the generalizeability of the experiencer set). Due to the nested structure of the data—trials were nested within experiencer as well as valence—we also conducted two multilevel linear mixed models (LMMs) predicting absolute and raw scores respectively. These enabled us to account for clustering effects, and to model the random effects of both participants and stimuli

[7] There were several outliers and the assumption of homogeneity of variance was violated in several conditions; however, ANOVA should be fairly robust to deviations from normality with a large sample size, and to heterogeneity with fairly equal sample sizes (Norman, 2010), and so the analysis proceeded. The assumption of sphericity was violated, which may have been due to over-sensitivity with a large sample (Weinfurt, 2000) and so Greenhouse-Geisser adjusted values are reported (Maxwell & Delaney, 2004).

(traditional ANOVAs fail to treat both as random; LMMs also offer other benefits including the ability to handle unbalanced data; Judd et al., 2012).[8]

Explanatory power was substantial for both the absolute (conditional R^2 = 0.29; marginal R^2 = 0.01) and raw models (conditional R^2 =0.52; marginal R^2 = 0.22). However, when using absolute scores, neither process (*B* = 0.06, 95% CI [-0.01, 0.13], *t*(18113) = 1.56, *p* = 0.119; *β* = 0.02, 95% CI [-0.006, 0.05], valence *B* = -0.12, 95% CI [-0.32, 0.08], *t*(18113) = -1.17, *p* = 0.241; *β* = -0.07, 95% CI [-0.20, 0.05]) nor experiencer (*B* = -0.05, 95% CI [-0.15, 0.05], *t*(18113) = -1.03, *p* = 0.303; *β* = -0.07, 95% CI [-0.19, 0.06]) were significant contributors to the model. In contrast, the model using raw difference scores yielded significant effects of process (*B* = -0.22, 95% CI [-0.33, -0.12], *t*(18108) = -4.13, *p* < .001; *β* = -0.06, 95% CI [-0.08, -0.03]) and valence (*B* = -1.11, 95% CI [-1.43, -0.80], *t*(18108) = -7.00, *p* < .001; *β* = -0.46, 95% CI [-0.59, -0.33]) but not experiencer (*B* = 0.07, 95% CI [-0.08, 0.22], *t*(18108) = 0.90, *p* = 0.368; *β* = 0.06, 95% CI [-0.07, 0.19]).

Results from the LMMs indicate that while some variation was caused by the random effects of stimuli, the experiencers that trials were associated with (or, in design terms, nested under) did not have a significant influence on difference scores. Additionally, the null effect of process in the absolute scores model in contrast to the significant effect found via the raw scores model supports our initial findings (*t*-tests and two-way [process × valence] ANOVAs) and assumption that raw scores would provide a more nuanced reflection of the data through which significant directional differences could be revealed.

[8] Mean-centring was not used for the categorical predictors (Process was coded as simulation = 0, theorization = 1; Experiencers A-F were numerically coded as 1-6 respectively; Valence was coded as negative = 0, neutral = 1, positive = 2), or the dependent variable (difference scores, where values of 0 were meaningful and represented no difference between participant and target ratings).

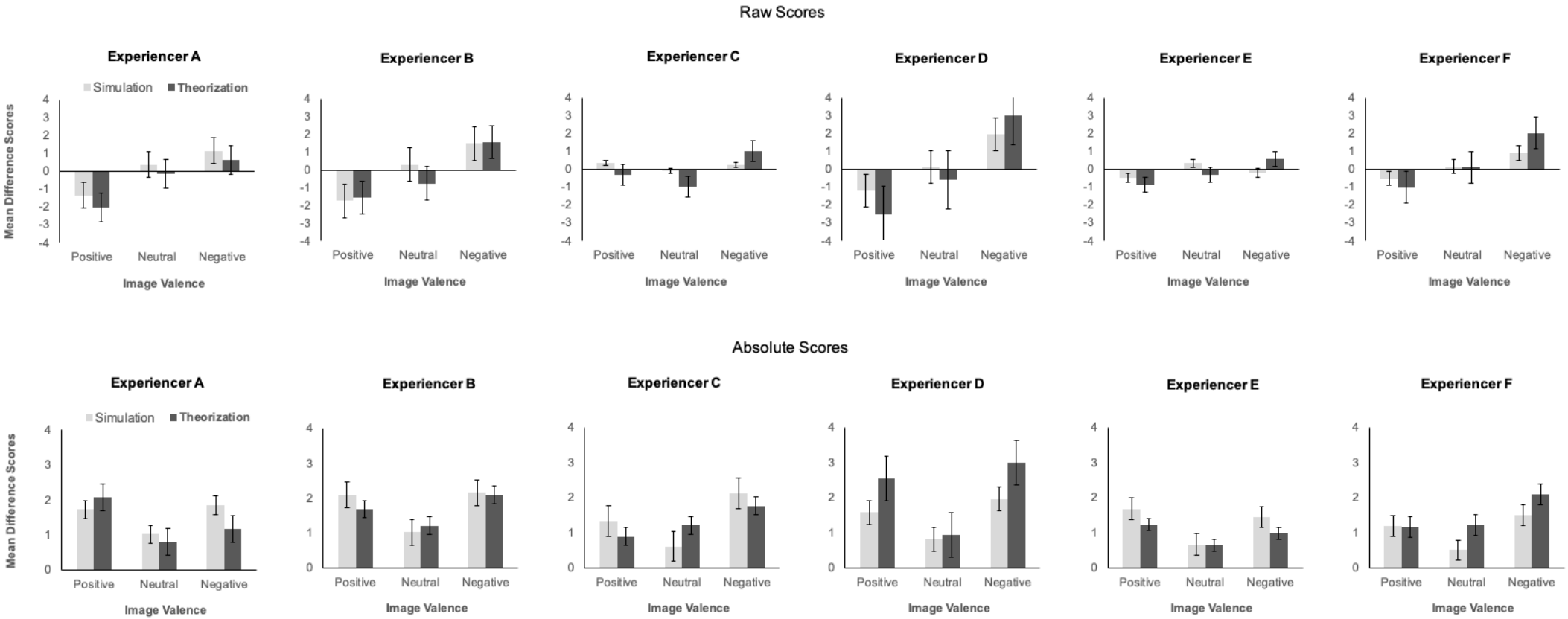

Figure 4. Mean raw (top) and absolute (bottom) difference scores in the simulation (light grey) and theorization (dark grey) groups for each of the six experiencers. See supplemental tables 1-2 for exact values and associated significance.

Fiction-exposure

A multiple linear regression was conducted to test the hypothesis that fiction-exposure would positively predict empathic accuracy. The three ART dimensions (fiction-exposure, nonfiction-exposure and foil selection) predicted 9% of the variation in absolute difference scores, $F(3, 298) = 9.32$, $p < .001$, $R^2 = .09$, adj. $R^2 = .08$. Only fiction-exposure, $\beta = -.28$, $p < .001$, significantly contributed to the prediction, such that for each fiction author recognised, empathic accuracy scores increased by .011 ($B = -.011$, p < .001, 95% CI [-.016, -.005]).[9]

The same procedure using the raw difference scores revealed similar results: this model predicted 4% of variance, $R^2 = .19$, adj. $R^2 = .035$, $F(3, 298) = 3.57$, $p = .014$. Fiction-exposure, $\beta = -.17$, $B = -.01$, $p = .031$, 95% CI [-.018, -.001], and foil selection, $\beta = .14$, $B = .07$, $p = .027$, 95% CI [.008, .13], significantly contributed to the model. The inclusion of age (which was positively associated with fiction-exposure in in both process conditions, $rs < .30$, $ps < .001$), revealed the same pattern of results, and age did not significantly contribute to either model. A moderator analysis tested the hypothesis that fiction-exposure would particularly support accuracy in the simulation condition. The interaction term did not produce a significant increase in explained variance in the absolute, $F(1, 298) = .15$, $p = .70$, or raw scores $F(1, 298) = 1.5$ $p = .23$, indicating, in contrast to our prediction, that the relationship between fiction-exposure and empathic accuracy held across process conditions.

Retroactive Group Preference

In line with Zhou et al.'s (2017) findings, participants indicated that they believed that viewing videos of the experiencers (theorization, 64%) would represent the more effective process for estimating their emotion ratings in response to pictures, compared to viewing the same pictures (simulation, 36%). Of participants who completed the task in the theorization condition, 48% selected theorization as the most effective process and 52% selected simulation. In contrast, 80% of simulation participants believed that theorization would

[9] Negative beta values represent increases rather than a decreases in empathic accuracy as operationalised as an inversion of the difference scores.

represent the most effective process. A Chi-square test showed that the association between process group and retroactive group preference was statistically significant, $\chi^2(1) = 34.30$, $p < .001$. Binomial logistic regression was used to predict process preference from ART scores and process condition. The model was a good fit, $\chi^2(8) = 10.49$, $p = .23$ and revealed that fiction-exposure did not predict retroactive group preference: only the condition in which participants had completed the empathic accuracy task significantly contributed to the prediction, $B = -1.50$, Exp(B) = .223, $p = .001$, 95% CI [-2.09, -1.00], $\chi^2(4) = 36.25$, $p < .001$, Nagelkerke $R^2 = .16$. Participants who had been assigned to the simulation condition for the empathic accuracy task were 18% more likely to select theorization compared to those who had been assigned to the theorization condition. Participants tended to consider the condition they had completed the task in as the least effective for estimating others' emotional experiences, and theorization was considered the best approach overall.

Discussion

This study attempted to replicate Zhou et al.'s (2017) findings that sharing in the experiences of targets (simulation) leads to higher empathic accuracy compared to mentalizing based on facial cues (theorization), and that participants tend to overvalue theorization. We aimed to extend this research by examining the impact of valence, identification with the experiencer, and general exposure to fictional stories. In order to detect nuanced (positive and negative) variation in empathic accuracy, we conducted our analyses using raw scores, to examine the direction of effects, as well as absolute scores to examine magnitudes (the latter was in line with Zhou et al.).

Empathic Accuracy in Theorization Versus Simulation Conditions

The absolute scores showed that the average magnitudes of error (differences between participants' ratings for the experiencers and the experiencers' actual ratings) were similar across the two conditions, and so we failed to replicate Zhou et al.'s first finding. However, the raw scores revealed group differences that had been obscured by the compression of

variance associated with using the absolute values: estimates in the theorization condition tended to be over-negative whereas estimates in the simulation condition tended to be over-positive. When viewing experiencers' facial expressions, participants' estimates were negatively biased, whereas when viewing the same set of pictures, participants' estimates were positively biased. Therefore, while the initial analysis did not replicate Zhou et al.'s finding that simulation represents the most effective process for accurately estimating the affective states of others—rather, each process resulted in similar levels of error—our exploratory analysis revealed that the direction of that error differed between process conditions.

Furthermore, our more granular examination of the data established interactions between process condition and stimuli valence (Figure 3). Absolute scores were similarly inaccurate for positive and negative images, but more accurate for neutral images. In other words, participants were more prone to error when estimating reactions to emotional content than neutral content. Interestingly, this did not appear to differ between the simulation and theorization conditions; only for neutral trials did absolute scores differ between process conditions with simulation participants showing greater accuracy. This suggests that for neutral content, the more successful approach was to view the same pictures as the experiencers, rather than trying to interpret their facial expressions. A similar pattern emerged from the raw scores where estimates for affective stimuli tended to be conservative (i.e., estimates for positive images tended to be more negative than the experiencers' own ratings and estimates for negative images tended to be more positive). This effect was stronger in the theorization condition, indicating that reading facial expressions led to the tendency to err towards even more conservative estimates. For neutral trials, participants in the simulation condition showed a small positive bias and in the theorization condition this bias was negative. Therefore, this analysis of interactions between inferencing process condition and valence lent support to Zhou et al.'s finding that simulation represents the more successful

process for interpreting the emotional responses of others. Without establishing the effects of valence, the pattern of conservative biases could not have been identified, and so examining both the raw and absolute scores provided additional information about the extent and direction of bias when "inferring perspective versus getting perspective" (Zhou et al., 2017, p. 482).

Retroactive Group Preference

Zhou et al. (2017) found that participants over-valued theorization, despite simulation being the more effective approach to empathic accuracy, and theorization also represented the most popular choice with our sample. The only other influence was the condition in which participants had completed the task: they tended to select that which they had not previously taken part in. While self-assignment to a group can lead to retroactive positive bias towards the choice made, arbitrary or random assignment can lead to favouring the unassigned choice (see Mather et al., 2003; Stoll Benney & Henkel, 2007) and this may account for some variance in our data. However, participants remained significantly more likely to value the theorization condition in general, which replicated Zhou et al. and supported their suggestion that while the effectiveness of process may vary across contexts, people's tendency to overvalue their ability to read faces and undervalue insight gained from sharing in another's experience is systematic. This knowledge could support people to better understand and employ the empathic inferencing tools at their disposal.

Sources of Inferencing Bias

There are several possible explanations for the finding that, despite participants perceiving the theorization condition as the most likely to be successful, participants in the simulation condition actually performed better (as established through our examination of raw scores across levels of valence). First, as simulation may be systematically undervalued compared to theorization, people's overconfidence in their ability to read facial expressions may lead to bias (the robust over-confidence effect; e.g., Pallier et al., 2002) whereas under-

confidence in experience-sharing may reduce it. The directional effects of this bias may take the form we observed in our data. As negativity bias (e.g., Baumeister et al., 2001; Rozin & Royzman, 2001) is thought to serve an adaptive, predictive function, enabling perceivers to quickly identify a potential threat (Hansen & Hansen, 1988), participants who viewed the facial expressions of target experiencers may have been particularly predisposed to this bias. However, negativity bias has also been documented with images (e.g., Pritsch et al., 2017) and yet it was not present in the simulation condition where, on the whole, participants showed more positive biases. Here, participants' own affective responses served as referents for the experiencers', and so reappraisal processes could have moderated the initial negative bias (Baumeister et al., 2001; Petro et al., 2018). Future research may examine the extent to which directional bias depends on the modality of a stimulus (e.g., Kauschke et al., 2019).

Individual differences in the perceiver may play a role in the ability to interpret positive, negative and neutral facial expressions. For example, individuals vary in "valence focus", the tendency to emphasize positive and negative information in verbal reports, which has been linked to an increased ability to process affective stimuli, and heightened sensitivity towards negative facial expressions (Barrett & Niedenthal, 2004). The mood of the perceiver can also bias recognition of incongruent emotions (e.g., negative bias in participants primed with a sad mood and positive bias in participants primed with a happy mood; Schmidt & Schmidt Mast, 2010).

Individual differences in perceivers' lifetime fiction-exposure positively predicted some of the variance in empathic accuracy even when nonfiction-exposure was controlled, aligning with previous research showing higher levels of empathic accuracy in individuals who have read more fiction (see Mumper & Gerrig's, 2017, meta-analysis). This effect did not interact with condition, however, suggesting that while fiction-exposure may support the development of empathic accuracy over time, it does not uniquely contribute to one empathic inferencing process over the other. Fiction can develop empathic accuracy through readers

learning about social content, or through the process of regularly simulating the experiences of characters (Mar, 2018) and so both mentalizing and experience-sharing processes may be enhanced through fiction (Mar & Oatley, 2008). Future research may contribute to an understanding of how far different forms of fiction, including genre and media presentation (e.g., Black & Barnes, 2015; Fong et al., 2013; Kidd & Castano, 2013; Turner & Felisberti, 2018), support empathic accuracy using mentalizing and experience-sharing processes.

When modelling the random effects of stimuli and participants across clusters via our LMM, we found no effect of experiencer on empathic accuracy. This suggests that, in general, individual differences across our target experiencers did not systematically impact perceivers' accuracy. Furthermore, there was no evidence of an ingroup advantage (e.g., Adams et al., 2010; Matsumoto et al., 2009) as identification with experiencers did not relate to empathic accuracy (a question of *inter*-individual differences). It may be that emotional expressiveness—the ability to accurately convey feelings nonverbally—of our whole experiencer set contributed to the pattern observed across empathic accuracy scores (emotional expressiveness shows stability across contexts, Allport & Vernon, 1933, and is associated with other personality traits [positively with extraversion and negatively with neuroticism; Riggio & Riggio, 2002] as well as gender [women tend to be more nonverbally expressive than men; Hall, 1990]). If the experiencers' facial expressions did not reflect the extent of their emotional responses to the pictures—if they were muted—this could partly explain the pattern of more conservative estimates in the theorization condition data. While our experiencers knew that they were being filmed and so self-consciousness could have impacted their expressiveness, Zhou et al. (2017) adapted their procedure for their third experiment so that experiencers were not aware of being filmed and obtained results consistent with their previous experiments. Therefore, self-consciousness in the lab setting does not offer a sufficient explanation of our finding that estimates were generally conservative and more so in the theorization condition.

Taken together, our findings suggest that success in estimating the intensity of emotional experiences in others is greater when sharing in the same experience compared to reading facial expressions, due to estimates being less conservative. Nevertheless, perceivers tend to overvalue insight gained when reading facial cues. Individual differences also play a role, with lifetime fiction-exposure supporting both processes. These findings demonstrate the importance of process, individual differences in perceivers, and the emotional context of stimuli, in inferring the emotions of others.

Limitations and Recommendations for Future Research

Measuring complex and contested empathic processes (e.g., De Vignemont & Singer, 2006) is challenging. Existing behavioural paradigms tend to test a single component, and while studies of theory of mind and facial emotion perception regularly employ behavioural tasks (for an overview see Turner & Felisberti, 2017), measures of experience-sharing tend to be physiological or self-report (Zhou et al., 2003), with the latter often focusing on dispositional rather than situational empathy. To our knowledge Zhou et al.'s (2017) approach, which we aimed to replicate and extend, provided the first behavioural paradigm designed to enable comparison between "reading" versus "being" empathic inferencing processes. However, in conducting and interpreting the results of the present study, we acknowledge the untested assumption of criterion validity: that the test and conditions we, and Zhou et al. (2017) employed, map onto the empathic accuracy construct. It is important to establish the validity and reliability of a novel tool, particularly in cases where, as in the present study, results diverge significantly between conditions. This would require analysis of performance on this tool in relation to performance on established measures of experience-sharing and mentalizing which neither we, nor Zhou et al., conducted. Nonetheless, alignment between Zhou et al.'s findings and ours, suggests that the present approach to examining the respective efficacy of two identified routes to empathic accuracy (see Zaki & Ochsner, 2012) is promising.

The outcome measure was one of target-rater (experiencer-participant) agreement. However, operationalization of empathic accuracy differed between conditions since the stimuli on which participant estimates for emotional ratings were based were accessed directly (by viewing the pictures themselves in the simulation condition) or indirectly (via real-time facial reactions in the theorization condition). This approach was aimed at addressing an important question for social cognition research: how far do mentalizing versus experience-sharing processes serve empathic accuracy? As Zhou et al. put it: "If you want to know whether someone likes a jellybean, should you watch the person eating it or taste it yourself? […] If you really want to understand the mind of another person, should you try to read that person to infer his or her perspective or try to be that person by putting yourself in that person's experience and getting his or her perspective directly?" (Zhou et al., 2017, pp. 482-483). Our results do not imply that sharing in another's experience is always the most accurate way to understand their feelings; rather, they align with Zhou et al.'s conclusion that it can be a more effective approach compared to reading facial expressions, but tends to be undervalued in comparison. As suggested above, future research employing concomitant tests known to probe mentalizing and experience-sharing processes could further support the confidence in this approach and the conclusions derived from it.

The experimental paradigm was predicated on the assumption that the experiencers' perspectives on their emotional responses were accurate, which is an inherent limitation of such approaches. Employing dynamic facial reactions as stimuli represents a more ecologically valid method for assessing empathic accuracy compared to static photographs or schematic faces (e.g., Dobs et al., 2018), though the trade-off is that experiencers' self-reports could be subject to bias (e.g., certain experiencers might systematically exaggerate or downplay their emotional responses, or they may have difficulty labelling their emotions accurately, resulting in apparently inflated or conservative estimates for some experiencers). This issue could also be usefully addressed in future research by employing validated

behavioural measures alongside newly developed (and ecologically valid) test measures (Turner & Felisberti, 2017) that can more easily generalize across contexts.

Data from male participants were collected after data from female participants (due to a fault in the study's online screening logic), and so gender differences across participants cannot be reliably interpreted. Furthermore, it is not possible to rule out effects of experiencer gender, and other differences in our experiencer set, as our sample of experiencers was small. Research has, for example, indicated that a target's gender can moderate the effect of valence on emotion-labelling (Garrido and Prada, 2017, found that participants correctly classified more "happy" expressions in images of women and more "angry" expressions in men). We could not make inferences based on gender, though we observed that participants showed generally higher levels of accuracy when rating the emotions of Experiencer E (male), particularly for neutral images (Figure 4; Supplemental Tables 1-2). Future research could explore both the individual and inter-individual effects of gender across mentalizing and experience-sharing processes.

Our replication was conceptual rather than exact, and future pre-registered replications would be helpful in lending support to the original findings of Zhou et al. (2017), our extended findings, and clarifying the point of disparity in results using absolute scores (namely, that Zhou et al. found in favour of simulation whereas we obtained a null result using absolute values). We replicated Zhou et al.'s finding that participants showed a preference for the theorization condition, but while Zhou et al.'s participants self-assigned into process groups, our assessment of group preference was retroactive, and so may have been biased by participants' experiences of being completing the task in the process condition assigned to them. Indeed, process condition was the only significant predictor of this preference (although it only accounted for a portion the variance) and future studies could assess preference before the task is completed to obtain a purer measure.

Both absolute and raw difference values were employed in our analysis, which was

aimed at providing a nuanced perspective of the variance associated with mentalizing versus experience-sharing processes. Therefore, it is important to consider the inflation of familywise error associated with multiple testing. Bonferroni-adjustments for the two sets of dependent variables (raw and absolute scores) did not significantly alter the pattern of results, but our approach did facilitate an understanding of the directional differences associated with inferencing process and stimulus valence.

Summary and Conclusion

The accurate interpretation of others' internal experiences is essential to making sense of the social world. Through a conceptual replication of Zhou et al. (2017, Experiments 1-2), we tested the respective efficacy of mentalizing (theorization) versus experience-sharing (simulation) processes for interpreting others' affective states, and participants' perceptions of their efficacy. Examining magnitudes of error using absolute values, we found no effect of process and so failed to replicate Zhou et al.'s finding that simulation was more successful than theorization when inferring the emotional states of others. However, our exploratory analysis using the raw (positively and negatively signed) values—reflecting the full rating scale used by participants—showed that estimates for affective (positive and negative) trials were generally over-conservative, and that this pattern was accentuated in the theorization condition. Thus, when the raw data were examined across levels of valence, simulation was found to represent the more successful inferencing process. Despite the relative value of simulation, we replicated Zhou et al.'s finding that participants tended to show a preference for the theorization condition. Therefore, in line with Zhou et al., we found that interpreting emotions by reading others' facial expressions was overvalued compared to being "in their shoes" by experiencing the same stimuli; participants' intuitions about how best to interpret the emotional experiences of others did not reflect their actual success in doing so. Individual differences in lifetime fiction exposure also contributed to empathic accuracy, with more frequent readers better able to accurately detect emotions in others, regardless of inferencing

process. Future research may support and expand the current paradigm to examine the role of stimulus modality and inter-individual differences in empathic accuracy, and explore how particularly cultivating experience-sharing processes could enhance in this critical social skill.

References

Adams Jr, R. B., Rule, N. O., Franklin Jr, R. G., Wang, E., Stevenson, M. T., Yoshikawa, S., Nomura, M., Sato, W., Kveragam K., & Ambady, N. (2010). Cross-cultural reading the mind in the eyes: An fMRI investigation. *Journal of Cognitive Neuroscience, 22*, 97-108. http://doi.org/10.1162/jocn.2009.21187

Allport, G. W., & Vernon, P. E. (1933). The problem of consistency in expressive movement. In G. W. Allport & P. E. Vernon (Eds.), *Studies in expressive movement* (pp. 3-35). MacMillan.

Baron-Cohen, S., & Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with Asperger Syndrome or high functioning autism, and normal sex differences. *Autism and Developmental Disorders, 34*, 163-175. https://doi.org/10.1023/B:JADD.0000022607.19833.00

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science, 20*, 286-290. http://dx.doi.org/10.1177/0963721411422522

Barrett, L. F. & Niedenthal, P. M. (2004). Valence focus and the perception of facial affect. *Emotion, 4*, 266-274. http://dx.doi.org/10.1037/1528-3542.4.3.266

Batson, C. D. (1987). Prosocial motivation: Is it ever truly altruistic? In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 20, pp. 65–122). Academic Press.

Batson, C. D. (2011). *Altruism in humans*. Oxford University Press.

Batson, C. D., Duncan, B. D., Ackerman, P., Buckley, T. & Birch, K. (1981). Is empathic emotion a source of altruistic motivation? *Journal of Personality and Social Psychology, 40*, 290-302. http://doi.org/10.1037//0022-3514.40.2.290

Batson, C. D., Lishner, D., & Stocks, E. (2015). The empathy-altruism hypothesis. In D. Schroeder, & W. Graziano (Eds.), *The Oxford handbook of prosocial behavior.* New York, NY: Oxford University Press.

Baumeister, R. F., Bratlavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5,* 323-370. https://doi.org/10.1037//1089-2680.5.4.323

Birch, S. A. J., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*, 382-386. http://dx.doi.org/10.1111/j.1467-9280.2007.01909.x

Black, J. E., & Barnes, J. L. (2015). Fiction and social cognition: The effect of viewing award-winning television dramas on theory of mind. *Psychology of Aesthetics, Creativity, and the Arts, 9*, 355-494. http://doi.org/10.1037/aca0000031

Black, J. E., Barnes, J. L., Oatley, K., Tamir, D. I., Dodell-Feder, D., Richter, T., &Mar, R. (2021). Stories and their role in social cognition. In D. Kuiken, & A. M. Jacobs (Eds.) *Handbook of Empirical Literary Studies* (pp. 229-250). Berlin: De Gruyter. https://doi.org/10.1515/9783110645958

Camerer, C. F. (2003). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Castano, E. (2012). Anti-social behavior in individuals and groups: An empathy-focused approach. In K. Deux, & M. Snyder (Eds). *The Oxford handbook of personality and social psychology* (pp. 419-445). Oxford University Press.

Dan-Glauser, E. S., & Scherer, K. R. (2011). The Geneva affective picture database (GAPED): a new 730-picture database focusing on valence and normative significance. *Behavior Research Methods, 43*, 468-477. http://dx.doi.org/10.3758/s13428-011-0064-1

Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *JSAS Catalog of Selected Documents in Psychology, 10*, 85. Retrieved from: http://www.uv.es/~friasnav/Davis_1980.pdf

Decety, J. & Lamm, C. (2006). Human empathy through the les of social neuroscience. *The Scientific World Journal, 6,* 1146-1163. https://doi.org/10.1100/tsw.2006.221

De Vignemont, F. B. M., & Singer, T. (2006). The empathic brain: How, when and why? *Trends in Cognitive Sciences, 10,* 435-441. https://doi.org/10.1016/j.tics.2006.08.008

Dobs, K., Bülthoff, I., & Schulz, J. (2018). Use and usefulness of dynamic face stimuli for face perception studies–a review of behavioural findings and methodology. *Frontiers in Psychology,* 9, 1-7. https://doi.org/10.3389/fpsyg.2018.01355

Duval, C., Piolino, P., Bejanin, A., Eustache, F., & Desgranges, B. (2010). Age effects on different components of theory of mind. *Consciousness and Cognition 20,* 627-642. http://dx.doi.org/10.1016/j.concog.2010.10.025

Fong, K., Mullin, J. B., & Mar, R. A. (2013). What you read matters: The role of fiction genre in predicting interpersonal sensitivity. *Psychology of Aesthetics, Creativity and the Arts, 7,* 370-376. http://doi.org/10.1037/a0034084

Gallagher, S., & Gallagher, J. (2019). Acting oneself as another: An actor's empathy for her character. *Topoi: An International Review of Philosophy, Online First,* 1-12. https://doi.org/10.1007/s11245-018-9624-7

Garrido, M. V., & Prada, M. (2017). KDEF-PT: Valence, emotional intensity, familiarity and attractiveness ratings of angry, neutral, and happy faces. *Frontiers in Psychology, 8,* 1-9. https://doi.org/10/3389/fpsyg.2017.02181

Gilbert, D. T., Killingsworth, M. A., Eyre, R. N., & Wilson, T. D. (2009). The surprising power of neighborly advice. *Science, 323,* 1617-1619. https://doi.org/10.1126/science.1166632

Hall, J. A. (1990). *Nonverbal sex differences: Accuracy of communication and expressive style.* Johns Hopkins University Press.

Hansen, C. H., & Hansen, R. D. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology, 54*, 917-924. http://dx.doi.org/10.1037/0022-3514.54.6.917

Happé, F. G. E., Winner, E., & Brownell, H. (1998). The getting of wisdom: theory of mind and old age. *Developmental Psychology, 34*, 358–362. http://dx.doi.org/10.1037/0012-1649.34.2.358.

Harbaugh, W. T., Mayr, U., & Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science, 316,* 1622-1625. https://doi.org/10.1126/science.1140738

Hein, G., Silani, G., Preuschoff, K., Batson, D. C., & Singer, T. (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron, 68,* 149-160. http://dx.doi.org/10.1016/j.neuron.2010.09.003

Humphrey, K., Underwood, G., & Lambert, T. (2012). Salience of the lambs: A test of the saliency map hypothesis with pictures of emotive objects. *Vision, 12.* 1-15. http://dx.doi.org/10.1167/12.1.22

Ickes, W. (Ed.). (1997). *Empathic accuracy.* New York, NY: Guilford Press.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54-69, https://doi.org/10.1037/a0028347

Kauschke, C., Bahn, D., Vesker, M., & Schwarzer, G. (2019). The role of emotional valence for the processing of facial and verbal stimuli–positivity or negativity bias? *Frontiers in Psychology, 10,* 1-15. https://doi.org/10.3389/fpsyg.2019.01654

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science, 342*, 377-380. http://doi.org/10.1126/science.1239918

Kilts, C. D., Egan, G., Gideon, D. A., Ely, T. D., & Hoffman, M. (2003). Dissociable neural pathways are involved in the recognition of emotion in static and dynamic facial expressions. *NeuroImage, 18.* 156-168. http://dx.doi.org/10.1006/nimg.2002.1323

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). *International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual* (Technical Report A-8). University of Florida.

Leppänen, J. M., & Hietanen, J. K. (2003). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research, 69,* 22-29. https://doi.org/10.1007/s00426-003-0157-2

Mar, R. A. (2018). Evaluating whether stories can promote social cognition: Introducing the Social Processes and Content Entrained by Narrative (SPaCEN) framework. *Discourse Processes, 55,* 454-479. https://doi.org/10.1080/0163853X.2018.1448209

Mar, R. A., & Oatley, K. (2008). The function of fiction is the abstraction and simulation of social experience. *Perspectives on Psychological Science, 3*, 173-192. http://dx.doi.org/10.1111/j.1745-6924.2008.00073.x

Mar, R.A., Oatley, K., Hirsch, J., dela Paz, J. & Peterson, J.B. (2006). Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality, 40,* 694-712. http://dx.doi.org/10.1016/j.jrp.2005.08.002

Mar, R. A., Oatley, K., & Peterson, J. B. (2009). Exploring the link between reading fiction and empathy: Ruling out individual differences and examining outcomes. *Communications, 34,* 407-428. http://dx.doi.org/10.1515/COMM.2009.025

Mather, M., Shafir, E., & Johnson, M. K. (2003). Remembering chosen and assigned options. Memory & Cognition, 31, 422 433. http://dx.doi.org/10.3758/BF03194400

Matsumoto, D., Olide, A., & Willingham, B. (2009). Is there an ingroup advantage in recognizing spontaneously expressed emotions? *Nonverbal Behaviour, 33,* 181-191. http://dx.doi.org/10.1007/s10919-009-0068-z

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analysing data: A model comparison perspective* (2nd ed.). Psychology Press.

Maylor, E. A., Moulson, J. M., Muncer, A. M., & Taylor, L. A. (2002). Does performance on theory of mind tasks decline in old age? *British Journal of Psychology, 93,* 465-485. https://doi.org/10.1348/000712602761381358

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111,* 361-365. http://doi.org/10.1037/0033-2909.111.2.361

Mumper, M. L., & Gerrig, R. J. (2017). Leisure reading and social cognition: A meta-analysis. *Psychology of Aesthetics, Creativity and the Arts, 11*, 109-120. http://dx.doi.org/10.1037/aca0000089

Nichols, S., & Stich, S. P. (2003). *Mindreading: An integrated account of pretence, self-awareness, and understanding other minds.* Clarendon Press.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15,* 625-632. http://dx.doi.org/10.1007/s10459-010-9222-y

Olszanowski, M., Wróbel, M., & Hess, U. (2019). Mimicking and sharing emotions: a re-examination of the link between facial mimicry and emotional contagion. *Cognition and Emotion.* Advance online publication. https://doi.org/10.1080/02699931.2019.1611543

Paal, T., & Bereczkei, T. (2007). Adult theory of mind, cooperation, Machiavellianism: The effect of mindreading on social relations. *Personality and Individual Differences, 43,* 541-551. http://dx.doi.org/10.1016/j.paid.2006.12.021

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberta, R. D. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology 129,* 257-299. https://doi.org/10.1080/00221300209602099

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that": Attribution of second-order beliefs by 5- to 10-year old children. *Experimental Child Psychology, 39,* 437-471. http://dx.doi.org/10.1016/0022-0965(85)90051-7

Petro, N. M., Tong, T. T., Henley, D. J., & Neta, M. (2018). Individual differences in valence bias: fMRI evidence of the initial negativity hypothesis. *Social Cognitive and Affective Neuroscience, 13,* 687-698. http://dx.doi.org/10.1093/scan/nsy049

Pritsch, C. Telkemeyer, S., Mühlenbeck, C., & Liebal, K. (2017). Perception of facial expressions reveals selective affect-based attention in humans and orangutans. *Scientific Reports, 7*(7782), 1-12 http://dx.doi.org/10.1038/s41598-017-07563-4

R Core Team. (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing. https://www.R-project.org/.

Riggio, H. R., & Riggio, R. E. (2002). Extraversion, neuroticism, and emotional expressiveness: A meta-analysis. *Journal of Nonverbal Behavior, 26,* 195-218. http://dx.doi.org/10.1023/A:1022117500440

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin 86,* 638-641. http://dx.doi.org/10.1037/0033-2909.86.3.638

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5,* 296-320. https://doi.org/10.1207/S15327957PSPR0504_2

Schmidt, P. C., & Schmidt Mast, M. (2010). Mood effects on emotion recognition. *Motivation and Emotion, 34,* 288-292. http://dx.doi.org/10.1007/s11031-010-9170-0

Shamay-Tsoory, S. G., & Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. In *Neuropsychologia*, 45, 3054–3067. http://dx.doi.org/10.1016/j.neuropsychologia.2007.05.021

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly, 24*, 402-433. https://doi.org/10.2307/747605

Stoll Benney, K. & Henkel, L. A. (2006). The role of free choice in memory for past decisions, Memory, *14*, 1001-1011, http://dx.doi.org/10.1080/09658210601046163

Surguladze, A. A., Young, A. W., Senior, C., Brébion, G., Travis, M. J., & Phillips, M. L. (2004). Recognition accuracy and response bias to happy and sad facial expressions in patients with major depression. *Neuropsychology, 18,* 212-218. http://dx.doi.org/10.1037/0894-4105.18.2.212

Teding van Berkhout, E., & Malouff, J. M. (2016). The efficacy of empathy training: a meta-analysis of randomized controlled trials. *Counselling Psychology, 63,* 32-4. http://dx.doi.org/10.1037/cou0000093

Todd, R. M., Cunningham, W. A., Anderson, A. K., & Thompson, E. (2012). Affect-based attention as emotion regulation. *Trends in Cognitive Sciences, 16,* 365-372. http://dx.doi.org/10.1038/s41598-017-07563-4

Turner, R. & Felisberti, F. M. (2017). Measuring mindreading: A review of behavioral approaches to measuring "theory of mind" in neurologically typical adults. *Frontiers in Psychology, 8,* 1-7. http://doi.org/10.3389/fpsyg.2017.00047.

Turner, R., & Felisberti, F. (2018). Relationships between fiction media, genre, and empathic abilities. *Scientific Study of Literature, 8,* 261-292. https://doi.org/10.1075/ssol.19003.tur

Turner, R. & Vallée-Tourangeau, F. (2020). Fiction effects on social cognition: Varying narrative engagement with cognitive load. *Scientific Study of Literature, 10,* 96-130. https://doi.org/10.1075/ssol.19008.tur

Unoka, Z., Fogd, D., Füzy, M., & Csukly, G. (2011). Misreading the facial signs: Specific impairments in error patterns in recognition of facial emotions with negative valence in borderline personality disorder. *Psychiatry Research, 189*, 419-425. http://dx.doi.org/10.1016/j.psychres.2011.02.010

van Baaren, R. B., Holland, R. W., Kawakami, K., & van Knippenberg, A. (2004). Mimicry and prosocial behavior. *Psychological Science, 15*, 71–74. http://dx.doi.org/10.1111/j.0963-7214.2004.01501012.x

Weinfurt, K. P. (2000). Repeated measures analyses: ANOVA, MANOVA, and HLM. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 317-361). American Psychological Association.

Zhou, H., Majka, E. A., & Epley, N. (2017). Inferring perspective versus getting perspective: Underestimating the value of being in another person's shoes. *Psychological Science, 28,* 482-493. https://doi.org/10.1177/0956797616687124

Zhou, Q., Valiente, C., & Eisenberg, N. (2003). Empathy and its measurement. In S. J. Lopez & C. R. Snyder (Eds.), *Positive psychological assessment: A handbook of models and measures* (pp. 269–284). American Psychological Association. https://doi.org/10.1037/10612-017

Supplemental Materials.

The Challenges of Measuring Empathic Accuracy: A Mentalizing Versus Experience-Sharing Paradigm

Updated Version of the Revised Author Recognition Test (ART-R).

The ART-R was published in *Research in Personality*, *40,* Mar, R. A., Oatley, K., Hirsh, J., dela Paz & Peterson, J. B., "Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds", 694-671. Copyright Elsevier (2006). Adapted with permission from Elsevier.

Fiction				
Romance	**Sci-Fi/Fantasy**	**Suspense/Thriller**	**Domestic fiction**	**Foreign (translation)**
Sidney Sheldon	Robert Jordan	Dean Koontz	John Updike	José Saramago
Danielle Steel	Douglas Adams	John LeCarré	W. O. Mitchell	Yukio Mishima
Jackie Collins	Anne McCaffrey	Robert Ludlum	Alice Munro	Gabriel García Márquez
Judith Krantz	William Gibson	Clive Cussler	Maeve Binchy	Albert Camus
Nora Roberts	Terry Brooks	Sue Grafton	Carol Shields	Umberto Eco
Iris Johansen	Terry Goodkind	Ian Rankin	John Irving	Milan Kundera
Diana Palmer	Piers Anthony	P. D. James	Toni Morrison	Paulo Coelho
Catherine Anderson	Arthur C. Clarke	John Saul	Amy Tan	W. G. Sebald
Joy Fielding	Ray Bradbury	Patricia Cornwell	Rohinton Mistry	Italio Calvino
Nicholas Sparks	Ursula K. Le Guin	Ken Follett	Sinclair Ross	Thomas Mann
E. L. James*	Kim Stanley Robinson*	Paula Hawkins*	Jodi Picoult*	Haruki Murakami*
Nonfiction				
Science	**Philosophy/Psychology**	**Political/Social-commentary**	**Self-help**	**Business**
Stephen Hawking	Roland Barthes	Noam Chomsky	Jack Canfield	Faith Popcorn
Stephen J. Gould	John Searle	Mary Beard*	Philip C. McGraw	Jim Collins
Richard Dawkins	Jean Baudrillard	Michael Moore	M. Scott Peck	Napoleon Hill
Thomas Kuhn	Michel Foucault	Eric Schlosser	Robert Fulghum	Robert T. Kiyosaki
Ernst Mayr	Bertrand Russell	Bob Woodward	Emma Bombeck	Stephen C. Lundin
Douglas Rushkoff	Antonio Damasio	Pierre Berton	Jean Vanier	Peter S. Pande
Amir D. Aczel	Daniel Goleman	Naomi Klein	Stephen R. Covey	Kenneth H. Blanchard
Matt Ridley	Jeffrey Gray	Naomi Wolf	Melody Beattie	Peter F. Drucker
John Maynard Smith	Joseph LeDoux	Robert D. Kaplan	Deepak Chopra	Barry Z. Posner
Diane Ackerman	Oliver Sacks	Susan Sontag	Marianne Williamson	Spencer Johnson**
Yuval Noel Harari*	Sam Harris*	Cordelia Fine*	Sarah Knight*	Sheryl Sandberg*
Foils				
Lauren Adamson	John Coundry	Martin Ford	James Morgan	Dale Blyth
Eric Amsel	Edward Cornell	Harold Gardin	Scott Paris	Robert Emery
Margarita Azmitia	Carl Corter	Frank Gresham	Richard Passman	Franklin Manis
Oscar Barbarin	Diane Cuneo	Robert Inness	David Perry	Alister Younger
Reuben Baron	Denise Daniels	Frank Keil	Miriam Sexton	Hilda Borko
Gary Beauchamp	Geraldine Dawson	Reed Larson	K. Warner Schaie	Frances Fincham
Thomas Bever	Aimee Dorr	Lynn Liben	Robert Siegler	Morton Mendelson
Elliot Blass	W. Patrick Dickson	Hugh Lytton	Mark Strauss	Steve Yussen

Updates to ART-R (Mar et al., 2006). "M. D. Johnson Spencer" was amended to "Spencer Johnson". New names (10%) were added to each dimension (one in each genre). These were authors whose works have been published or re-published within 5 years of scale construction. Each critical dimension consisted of 55 names and there were 40 foils in total. As a recipient of the Pulitzer Prize for both fiction and nonfiction, Norman Mailer was replaced with a nonfiction author of the same genre (Mary Beard).

Table 1. Mean absolute difference scores presented for each level of experiencer and valence within each inferencing process condition.

		Grand Mean	Simulation (i)	Theorization (j)	Mean difference (i–j)	*p*
Experiencer	A	1.39	1.49	1.29	0.20 [0.07, 0.32]	.005
	B	1.74	1.80	1.67	0.13 [0.04, 0.23]	.01
	C	1.39	1.43	1.34	0.09 [-0.008, 0.18]	.06
	D	1.71	1.40	2.05	-0.65 [-0.78, -0.52]	.001
	E	1.15	1.31	0.98	0.33 [0.23, 0.43]	.001
	F	1.33	1.11	1.56	-0.45 [-0.55, -0.33]	.001
Valence	Positive	1.61	1.63	1.59	0.04 [-0.06, 0.16]	.40
	Neutral	0.89	0.79	0.99	-0.20 [-0.29, -0.11]	.001
	Negative	1.85	1.84	1.86	-0.02 [-0.14, 0.11]	.73
Experiencer*	A*positive	1.90	1.73	2.08	-0.35 [-0.58, -0.14]	.007
Valence	A*neutral	0.92	1.03	0.80	0.23 [0.08, 0.38]	.002
	A*negative	1.51	1.85	1.16	0.70 [0.51, 0.89]	.001
	B*positive	1.90	2.09	1.69	0.39 [0.24, 0.54]	.001
	B*neutral	1.12	1.03	1.21	-0.17 [-0.32, -0.03]	.03
	B*negative	2.13	2.17	2.09	0.09 [-0.12, 0.30]	.41
	C*positive	1.12	1.34	0.90	0.44 [0.31, 0.58]	.001
	C*neutral	0.91	0.62	1.22	-0.60 [-0.74, -0.47]	.001
	C*negative	1.95	2.12	1.77	0.35 [0.19, 0.51]	.001
	D*positive	2.06	1.58	2.56	-0.98 [-1.20, -0.76]	.001
	D*neutral	0.88	0.82	0.94	-0.12 [-0.26, 0.18]	.10
	D*negative	2.48	1.97	3.01	-1.04 [-1.34, -0.74]	.001
	E*positive	1.46	1.69	1.23	0.47 [0.33, 0.61]	.001
	E*neutral	0.66	0.67	0.65	0.02 [-0.12, 0.17]	.78
	E*negative	1.23	1.45	0.99	0.46 [0.32, 0.60]	.001
	F*positive	1.19	1.20	1.17	0.04 [-0.12, 0.20]	.64
	F*neutral	0.86	0.51	1.23	-0.72 [-0.86, -0.58]	.001
	F*negative	1.79	1.50	2.10	-0.60 [-0.78, -0.41]	.001

Note. 95% bias-corrected and accelerated confidence intervals (using $N = 1000$ bootstrapping) for mean differences are presented in brackets.

Table 2. Mean raw difference scores presented for each level of experiencer and valence within each inferencing process condition.

		Grand Mean	Simulation (i)	Theorization (j)	Mean difference (i–j)	*p*
Experiencer	A	-0.17	0.10	-0.46	0.56 [0.38, 0.74]	.001
	B	-0.19	-0.001	-0.39	0.39 [0.24, 0.51]	.001
	C	0.09	0.17	0.015	0.16 [-0.01, 0.32]	.46
	D	0.11	0.30	-0.10	0.41 [0.25, 0.56]	.001
	E	-0.20	-0.14	-0.27	0.13 [-0.01, 0.26]	.69
	F	0.40	0.26	0.55	-0.29 [-0.44, 0.16]	.001
Valence	Positive	-1.06	-0.76	-1.37	0.60 [0.44, 0.76]	.001
	Neutral	-0.10	0.21	-0.44	0.65 [0.51, 0.78]	.001
	Negative	1.18	0.90	1.48	-0.58 [-0.78, -0.38]	.001
Experiencer*	A*positive	-1.67	-1.33	-2.03	0.69 [0.43, 0.97]	.001
Valence	A*neutral	0.13	0.36	-0.13	0.49 [0.28, 0.70]	.001
	A*negative	0.92	1.16	0.66	0.51 [0.20, 0.79]	.003
	B*positive	-1.46	-1.35	-1.57	0.22 [-0.002, 0.41]	.055
	B*neutral	-0.21	0.30	-0.76	1.05 [0.84, 1.28]	.001
	B*negative	1.53	1.50	1.56	-0.06 [-0.31, 0.19]	.65
	C*positive	0.02	0.34	-0.31	0.64 [0.44, 0.85]	.001
	C*neutral	-0.52	-0.09	-0.98	0.89 [0.67, 1.10]	.001
	C*negative	0.61	0.24	1.00	-0.76 [-1.06, -0.49]	.001
	D*positive	-1.84	-1.17	-2.54	1.37 [1.14, 1.60]	.001
	D*neutral	-0.21	0.16	-0.60	0.76 [0.57, 0.94]	.001
	D*negative	2.48	1.97	3.01	-1.04 [-1.35, -0.70]	.001
	E*positive	-0.66	-0.47	-0.85	0.38 [0.15, 0.60]	.006
	E*neutral	0.02	0.34	-0.32	0.66 [0.49, 0.82]	.001
	E*negative	0.18	-0.18	0.56	-0.74 [-0.98, -0.51]	.001
	F*positive	-0.75	-0.50	-1.00	0.50 [0.27, 0.71]	.001
	F*neutral	0.13	0.16	0.12	0.04 [-0.14, 0.22]	.67
	F*negative	1.46	0.90	2.04	-1.14 [-1.40, -0.88]	.001

Note. 95% bias-corrected and accelerated confidence intervals (using $N = 1000$ bootstrapping) for the mean difference (simulation-theorization) values are presented in brackets. Experiencer D rated each negative image at -4 (the most negative rating possible). Therefore, participants' estimates could never be over-negative and so values or D*Negative could only be positive (and therefore identical to the absolute values).