Accurate inferences of others' thoughts depend on where they stand on the empathic trait continuum

Abstract

This research explores the possibility that a person's (perceiver's) prospects of making a correct inference of another person's (target's) inner states depends on the personal characteristics of the target, potentially relating to how readable they are. Twenty-seven targets completed the Empathy Quotient (EQ) and were classified as having low, average or high EQ. They were unobtrusively videoed while thinking of an event of happiness, gratitude, anger and sadness. After observing targets thinking of such a past event, fifty-two perceivers (participants) in Study 1 were asked to infer what the target was thinking, and fifty perceivers in Study 2 were asked to rate the target's expression – positive or negative. Results suggested that (1) perceivers' accuracy in detecting targets' thoughts depended on which EQ group the target belonged to, and (2) target readability is not a proxy measure for level of target expressiveness. In other words, something about EQ status renders targets more or less easy to read in a way that is not simply explained by expressive people being more readable. We conclude with discussion of the importance of the target's trait as well as situation they experience in determining how accurately a perceiver might infer their inner states.

Key Words: mindreading; retrodiction; accuracy; empathic trait; spontaneous behaviour

Accurate inferences of others' thoughts depend on where they stand on the empathic trait continuum

Mindreading (known otherwise as mentalizing, empathic accuracy) refers to people's (perceivers') ability to infer what another person (the target) might think, feel, and know for the purpose of interpreting and predicting their behavior (Premack & Woodruff, 1978; Flavell, Miller, & Miller, 1993). Past research on mindreading has explored people's ability to infer others' mental states (e.g., Cassidy Ropar, Mitchell, & Chapman, 2013, 2015; Ickes, Stinson, Bissonnette, & Garcia, 1990; Pillai, Sheppard, & Mitchell, 2012; Pillai et al., 2014; Sheppard, Pillai, Wong, Ropar, & Mitchell, 2016; Wimmer & Perner, 1983); but much of this research largely ignores the characteristics of the target – the person we are making inferences about (Andrews, 2008; Rai & Mitchell, 2004; Wu, Sheppard, & Mitchell, 2016a, 2016b) -- as if we only need to focus on the features of the situation in order to explain mindreading. The empirical work reported here is novel in seeking to explore the possibility that some aspects of target traits might affect how accurately we make mental state inferences. Specifically, further investigation is needed that focuses on our accuracy in interpreting signals in natural, spontaneous target behaviour, taking into account that the target behaviour (and therefore the signal available for mindreading) will depend on individual differences in the targets, potentially measurable by where they stand on a trait continuum. This research will thus illuminate how accuracy in attributing inner states to others depends on considering their personality traits – something that has been largely overlooked to date.

Previous studies have suggested that perceivers are able to infer which situation caused a target's reaction (Cassidy et al., 2013, 2015; Pillai et al., 2012, 2014; Sheppard et al., 2016; Teoh, Wallis, Stephen, & Mitchell, 2017; Kang, Anthoney, & Mitchell, 2017) even though the particular situation experienced by the target provokes a range of reactions across different targets. Worldly events occurring in a given situation (e.g. something that happened to the target, something that the target witnessed or heard) evoke a mental state which in turn gives rise to a signal in the target that is potentially observable to a perceiver (Sheppard et al., 2016; Teoh et al., 2017; Valanides, Sheppard, & Mitchell, 2017). According to Teoh et al (2017), the information available to the perceiver is the target's behaviour (which is signalling something about the target's mind) and from this the perceiver makes a backwards inference to the underlying target mental state (the proximal cause of the target's behaviour – Kang, Schneider, Schweinberger, & Mitchell, 2018) and the perceiver then makes a further backwards inference to the event that evoked the target mental state (the distal cause). This process of 'retrodictive mindreading' (Gallese & Goldman, 1998; Teoh et al., 2017) confers considerable benefits in that we can exploit our ability to read others' minds to know various things in the world, including some things that cannot be apprehended through our ordinary senses. The current study thus was built on the framework of 'retrodiction', by which we explored accuracy in thought inferences from spontaneous target behaviour, in relation with the characteristics of the targets (where they stand on the empathy trait continuum).

Note, however, there is no precise correspondence between the particular form of target behavior and the event that triggered the reaction (Zaki & Ochsner, 2011; Russell, Bachorowski, & Fernandez-Dols, 2003). It is not the case, for instance, that when targets listened to an unfortunate story they reliably looked concerned (sometimes they looked amused, sometimes indifferent, sometimes bored, Pillai et al., 2012, 2014). The range of target reactions is linked causally with the situation or state, and while seldom acknowledged in previous research, it seems the particular reaction within that range is explained by the characteristics of the target. Thus, accounts of mindreading would be more comprehensive and useful if they recognised that perceivers (1) have to work with individual differences in how a target's signalled mind is displayed while (2) appreciating that the particular domain of inner state being experienced by the target nevertheless constrains the range of their reactions.

A small number of recent studies have begun exploring how characteristics of the target impact upon the perceiver's accuracy in mindreading. Studies conducted by Zaki, Bolger and Ochsner (2008, 2009) suggest that the target's level of expressivity is a significant predictor of perceiver performance in inferring how the target felt. Another recent study was conducted by Sheppard et al (2016), in which perceivers (participants) were asked to identify which of four events the target had experienced after viewing a short mute video of the target. Results suggested that that perceivers were more effective in detecting the minds of neurotypical targets than targets with autism spectrum disorder (ASD); though they rated ASD targets equally expressive as neurotypical targets, suggesting targets with ASD were expressive in a different way,

a way that was difficult for perceivers to interpret. In short, the behaviour that reflects the signalled mind might be easier to 'read' in some targets than in others. Yet, to our knowledge, no study has directly examined how individual differences in target characteristics determine perceiver effectiveness in detecting specific target states of mind.

Relevant to this matter, Wu et al (2016a) discovered that it was easier for perceivers, after watching a brief sample of behaviour, to identify targets located at the extremities of the continuum of empathic trait than it was to identify targets located in the middle of the continuum. Wu et al speculated that targets located at various points along the continuum might possess minds that vary in their level of readability (how easily a perceiver could infer their inner states). For example, a person who is unusually low in empathy (an extreme case being autism) might signal mental states quite differently than those closer to the middle of the empathic trait continuum (Brewer et al, 2016; Faso, Sasson, & Pinkham, 2015; Sheppard et al, 2016). According to Wu et al (2016a, 2017), targets located at empathic trait and big-five trait extremities were easy to identify as being low or high on trait continua. Accuracy in inferring another's mental states might depend on characteristic aspects of targets (Andrews, 2008; Zaki et al., 2008). The purpose of the current research was to test whether or not targets vary in how readable they are depending on where they stand on the empathy continuum.

We adapted a procedure of 'retrodictive mindreading,' that was used previously (Pillai et al., 2012, 2014; Cassidy et al., 2013, 2015; Teoh et al., 2017; Valanides et

al., 2017; Kang et al., 2017), in which the perceiver "makes a backward inference from the observed action to a hypothesized goal state" (Gallese & Goldman, 1998, p.497). The target was asked to think of something in the past that caused them to experience a particular state, where we assume the target's visible behaviour is an externalization of their inner thoughts (Faso et al., 2015; Valanides et al., 2017). Perceivers were then asked to infer what the targets had been instructed to think about (Valanides et al., 2017). Importantly, we the researchers knew independently what targets had been asked to think (one of four kinds of event), allowing us to compare perceiver judgments of the target's inner state against an objective fact, thus satisfying West and Kenny's (2011) 'truth condition'. The accuracy of perceivers' inferences of targets' inner states can thus be measured objectively as a matter of fact.

Study 1

Method

Based on the procedure developed by Valanides et al (2017) in which targets were cued to think about either positive or negative events they had experienced, in Study 1 targets were filmed while thinking of four autobiographical events, including those that led to positive feelings and those that led to negative feelings. Targets were classified into three groups according to their empathic trait measurable with the Empathy Quotient (EQ, Baron-Cohen & Wheelwright, 2004; Baron-Cohen, 2012): Low EQ, Average EQ and High EQ. We persevered with the trait of empathy in this research (1) to be consistent with the previous findings in empathic trait judgment

(Wu et al., 2016a) and (2) because extremities of this trait might be associated with a state that is less easy to read (Sheppard et al., 2016).

Perceivers were tasked with inferring which of the four events (a happy event, an event that provoked gratitude, a sad event, and an event that provoked anger) the target was thinking about after watching a short silent video of the target. The study tested: (1) how well perceivers inferred the thoughts of the targets; (2) whether accuracy in inferring the target thoughts varies depending on which EQ group the target belonged to.

Participants

Fifty-two college students (25 males; *M* = 20.67 years) in Guangzhou and Zhanjiang China participated as perceivers in exchange for monetary compensation. Sample size was calculated using the software G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2009), affording 95% power to detect a medium effect on the within-subjects factors and 94% power to detect a large effect on the interaction . Perceivers were shown photographs of the targets and were included only if they reported not having seen any of the targets previously. Two additional females were acquainted with one or more targets and were excluded.

Materials

Video stimuli collection and editing. Videos were collected from 27 college students (targets, 15 females, *M* = 21 years), recruited in exchange for monetary compensation. All had responded to a call to do a screen test advertising the university and to complete questionnaires, and they also were informed they needed

to talk of some experiences about themselves before the screen test. One additional male target was excluded due to a technical problem.

Targets were individually videoed in a quiet laboratory with a Sony Handycam HDR-SR12 video camera mounted on a tripod placed approximately 1.5 meters away to record the target's face and the top part of their body. The target sat at a desk facing the camera and the researcher sat opposite but out of view of the camera. Unknown to the target, the camera automatically began recording as soon as the target entered the room. At the end, before leaving the laboratory, all targets were fully debriefed and gave written informed consent to use the videos for research purposes.

On arrival, targets were issued with a consent form and an information sheet that outlined the tasks they would perform, and were informed they would only be videoed while doing the screen test. Once inside the laboratory, after they read the information sheet and signed the consent form, the researcher began with a brief conversation. After that, the target was asked to think of a specified past event and then talk about the experience. Each target repeated this exercise for six past experiences in total, including a happy experience, an experience that led to a feeling of gratitude, an angry experience, a sad experience, an experience of having breakfast and doing a routine activity during the weekend – the latter two were filler activities. The focal experiences (happy, gratitude, anger, sadness) included two of positive valence and tow of negative valence, but other than that the experiences were not pre-validated with respect to emotional distinctiveness from each other. The order of the experiences was counterbalanced across the targets. The target was asked to spend

about 1 minute silently recalling each experience before talking about it. Subsequently, the target was asked to read the script of promotional material to the camera after the researcher ostensibly switched to 'record mode'. This 'cover story' of examining whether the target might be talented in promoting the university gave legitimacy to the presence of the camera.

Four separate video clips of each target including thinking of the four emotional events (happy, grateful, angry and sad) were used in this study, making 108 videos in total (27 targets × 4 videos per target). The average duration of the video clips was 21.33 s (*SD* = 10.24; ranging from 7 s to 38 s) for the Happiness, 23.85 s for the Gratitude (*SD* = 5.88, ranging from 6 s to 30 s), 21.26 s (*SD* = 8.36; ranging from 7 s to 34 s) for the Anger, and 22.33 s for the Sadness (*SD* = 9.10; ranging from 6 s to 35 s). A one-way repeated-measures ANOVA (F (3, 78) = .70, p = .554) did not detect any difference between the mean duration of the videoclips of the four events.

Empathy Quotient (EQ). Following a short break for a couple of minutes, the target filled in the Empathy Quotient (EQ, Baron-Cohen & Wheelwright, 2004). The EQ questionnaire offers a comprehensive measurement of the trait structure of empathy. It comprises 40 items (along with 20 filter items) pertaining to a range of behaviours associated with empathizing, with an overall rating that is useful in determining individual differences in empathic trait. All targets completed the Chinese translated version of the EQ questionnaire (adopted from the website: http://www.autismresearchcentre.com/arc/default.asp).

Target EQ scores ranged from 12 to 64 (M = 37.52, SD = 14.45). A score in the range of 0-32 is low EQ and 11 targets were in this category, 33-52 is average and 10 targets were in this category, 53-63 is above average and 5 targets were in this category, and 64-80 is high and 1 target was in this category (Baron-Cohen, 2012). Following Wu et al (2016a) we combined the 'above average' and 'high' categories into one range from 53 to 80 that was re-labeled as a category of high EQ. We then grouped the targets into three EQ categories, with 11 in the Low EQ Group (4 males), 10 in the Average EQ group (5 males), and 6 in High EQ group (3 males).

Procedure

Perceivers were tested individually. A set of 108 target videos (27 targets each contributing 4 videos) was displayed in random order to each perceiver using E-Prime Version 2.0.8.22. In each trial, following a fixation cross ('+') presented for 800 ms, one video clip was displayed; after that, a response screen appeared, presenting a four-forced choice in a fixed order as response options ((1) an angry event, (2) a happy event, (3) a sad event and (4) a grateful event). The perceiver registered his/her inference of the target's thoughts by using the keyboard to select the number '1, 2, 3 or 4' for the corresponding options. After the perceiver made the choice the screen moved to the fixation cross in preparation for the next trial. Responses were automatically recorded by the software for later retrieval. Perceivers typically needed about 45 minutes to complete the task.

Results

Given that signal detection theory (SDT) allows assessment of accuracy and sensitivity that is immune to response bias (the tendency to select one category more frequently than another; Macmillan, 2002; Macmillan & Creelman, 2005), it is widely applied to measure performance across various tasks, such as accuracy in trait judgments (Wu et al., 2016; Wu et al., 2017) and mental state inferences (Pillai et al., 2012, 2014; Valanides et al., 2017; Kang, et al., 2017). We thus adopted SDT to compute participant accuracy (sensitivity) in inferring the thoughts of targets.

According to guidelines on calculating SDT (Macmillan, 2002; Macmillan & Creelman, 2005), a correct judgment that a target thought about a particular event counted as a 'hit' while an incorrect judgment that a target recalled the same event counted as a false alarm. Performance of participants across the different target EQ groups over a total of 27 trials for each state was characterised as single values for each perceiver in the form of d-prime (d') for assessing perceiver accuracy in inferring each state. Following Macmillan and Creelman (2005), where the number of hits (or false alarms) was 0, 0.5 was added and the hit rate (or false alarm rate) was then calculated; where the participant made the maximum number of hits or false alarms for a given state, 0.5 was subtracted from the number of hits or false alarms prior to calculating the hit rate or false alarm rate. The d' was then calculated by subtracting the z-score of the false alarm rate from the z-score of the hit rate (d' = Z (hit rate) – Z (false alarm rate), where function Z (p), $0 \leq p \leq 1$). In addition, according to SDT outlined by Macmillan and Creelman (2005), we represent the base-rate as the 'criterion' (c) for choosing any particular response category with the statistic c: the

more negative the value of c, the more perceivers were in favour of choosing this particular category, irrespective of whether correct; but when c is more positive, it implies perceivers were against choosing the particular category, meaning they were conservative in this case. Criterion c was calculated by -0.5 x (Z (false alarm rate) + Z (hit rate)).

Table 1 shows the means of hit rate (M_{HR}), false alarm rate (M_{FAR}), d-prime ($M_{d'}$) and criterion (M_c) of each mental state in each target EQ group, along with *t* values of one-sample *t* tests of each $M_{d'}$ where the comparison value is zero: If perceivers were unable to infer each of the four target thoughts, this would yield a $M_{d'}$ of zero for that thought. According to the results of one-sample *t* tests for each $M_{d'}$ presented in Table 1, perceivers were able to detect what targets were thinking when they were recalling either a happy or sad event across the three target EQ groups. Yet, perceivers were not equally effective in inferring a given thought, as shown in Table 1 and Figure 1. Specifically, perceivers were notably accurate in inferring the states of happiness and gratitude for targets with low EQ but had difficulty in inferring these two positive states when the targets had high EQ. In addition, perceivers were effective in inferring sadness in the high EQ group and inferring happiness in the average EQ group.

Table 1 & Figure 1 here

A repeated-measures ANOVA (with the three target EQ groups and the four mental states as the within-subjects factors) confirmed the results displayed in Figure 1: There were main effects related with the three target EQ groups (F (2, 102) = 9.94, $p < .001$, Cohen's $f = .44$) and the four mental states (F (3, 153) = 4.58, $p = .004$,

Cohen's f = .30), and a significant interaction between the two factors (Greenhouse-Geisser adjusted F (4.96, 253.06) = 14.50, p < .001, Cohen's f = .53).

Simple-effects analyses for the interaction between Target EQ Group and the States revealed the following results. Firstly, the main effects of the four states were found in both the low (F (3, 153) = 14.07, p < .001, Cohen's f = .52) and the high EQ groups (F (3, 153) = 14.39, p < .001, Cohen's f = .53) but not in the average EQ group (F (3, 153) = 1.15, p = .330). According to post hoc LSD tests, in the low EQ group, perceivers were most accurate in detecting the thought of happiness compared with the other target states ($ps \leq .003$), while in the high EQ group, perceivers were more accurate in inferring sadness compared with the two positive states (ps < .001).

Secondly, except for the thought of anger (F (2, 102) = .84, p = .435), main effects associated with the three other states were significant across the three target EQ groups (Happiness: Green-house Geisser adjusted F (1.67, 85.33) = 38.70, p < .001, Cohen's f = .74; Gratitude: F (2, 102) = 12.52, p < .001, Cohen's f = .49; Sadness: Green-house Geisser adjusted F (1.70, 86.43) = 3.28, p = .050, Cohen's f = .25). Post hoc LSD tests revealed the following: (1) perceivers were most accurate in inferring happiness when the targets were low in EQ (ps < .001) and least accurate when the targets had high EQ (ps < .001); (2) perceivers were least accurate in detecting gratitude in the high EQ group (ps < .001); (3) perceivers more accurately inferred sadness in the high EQ group than in the low EQ group (p = .041). In summary, how accurately perceivers inferred target thoughts depended on the EQ scales the targets belonged to and on what targets had been asked to think about.

As demonstrated in Table 1, it seemed perceivers adopted different criteria (M_c) when inferring what events the targets were thinking. A repeated-measures ANOVA (with the four states as the within-subjects factor) for the M_c across the three target EQ groups confirmed the results in Table 1: F (3, 153) = 14.84, $p < .001$, Cohen's f = .53. Post hoc LSD revealed the mean c associated with sadness was significantly lower than the mean c associated with the other three states ($ps < .001$), suggesting that generally perceivers were inclined to judge targets were thinking about a sad event when observing the target recalling any given autobiographic emotional experience.

Study 2

Method

Study 1 demonstrated that perceivers were generally able to detect the thoughts of happy and sad events, and the accuracy in inferring target thoughts depended on where the target stood on the empathic trait continuum and on which event the target was cued to think about. While the targets were recalling experiences, signals to their inner states perhaps leaked out to a greater or lesser degree, such as smiling or frowning. According to Soscia (2007), the happy and grateful events should arouse positive inner states, and the angry and sad events should arouse negative inner states. Thus, one might ask whether perceivers (in Study 1) were merely classifying target expressions as positive or negative (Kang et al., 2018) as a rather simplistic way of attributing specific thoughts to them. To investigate this possibility, Study 2 explored how perceivers explicitly rated target expressions (positive or negative) to determine

if the pattern of such judgments could reductively explain their inferences of target inner states. If not, then presumably perceivers are doing something more than merely classifying target expressions when asked to infer target inner states. Specifically, if perceivers merely classified target expressions as a strategy for making judgments without needing to infer target inner states, they would identify a positive expression when the target thought about either a happy or a grateful event, and identify a negative expression when the target recalled either an angry or a sad experience. If so, then perceivers' ratings of target expressions would be indistinguishable from their inferences of targets' inner states (Kang et al, 2018). The purpose of Study 2 was to investigate this possibility.

Participants

Fifty college students (22 males; *M* = 20 years) in Zhanjiang China voluntarily participated as perceivers. None had participated in Study 1. The sample size was determined using the G*Power 3, affording 95% power to detect a medium effect on the within-subjects factors. None of the perceivers had prior acquaintance with any of the targets. Four additional females were excluded for quitting in the middle of the task.

Procedure

The procedure was similar to Study 1 except after viewing each target video, the perceiver rated the target's expression on a five-point scale (from negative to positive). The perceiver registered his/her judgment by using the keyboard to select

the number '1, 2, 3 4 or 5' for the corresponding responses. Perceivers typically needed about 45 minutes to complete the task.

Results and Discussion

Table 2 summarizes perceivers' mean ratings of target expressions for each of the four states (Happy, Grateful, Angry, Sad) in each of the three target EQ groups, along with the corresponding one-sample *t* tests (comparing the means of expression ratings against the neutral point 3). The data show that perceivers generally rated target expressions positively when targets had been thinking of a time they felt happy; perceivers generally rated targets neutral when targets had been thinking of a time they felt grateful, and perceivers generally rated target expressions negatively when targets had been thinking of events that made them feel sad and angry. A one-way repeated-measures ANOVA revealed a significant difference in ratings of target expressions among the four events targets were cued to think about: Greenhouse-Geisser adjusted F (2.22, 108.91) = 1991.22, $p < .001$, Cohen's $f = 6.31$. Post hoc LSD tests suggest targets were rated most positively when thinking of something happy ($ps < .001$) and most negatively when thinking of something sad ($ps < .001$); target expressions were rated more positively when thinking of a time they felt grateful than when thinking of a time they felt angry ($p = .002$).

Table 2 about here

As revealed in Table 2 and Figure 2, perceivers generally rated targets as having positive expressions when they (the targets) were thinking of something happy and

rated targets as having negative expressions when they (the targets) were thinking of something sad, regardless of target EQs. When targets were thinking of a time they felt grateful, perceivers rated those targets with high EQ positively but rated those with either low or average EQ negatively. Surprisingly, perceivers rated targets with average EQ as having positive expressions when those targets were thinking of an event that made them feel angry. In short, ratings of target expressions were influenced by target EQ status as well as the kind of event the target was thinking about – but the pattern formed by these ratings was quite different than would have been expected if perceivers were making a simplistic link between the valence and strength of target expressions and what targets were thinking.

Figure 2 about here

To confirm the above results, we carried out a 3 × 4 repeated-measures ANOVA, with the three target EQ groups and the four kinds of event targets were thinking about (Happy, Grateful, Sad, and Angry) as within-subjects factors. Results showed main effects related to the three target EQ groups (Greenhouse-Geisser adjusted F (1.59, 77.70) = 2627.28, $p < .001$, Cohen's $f = 7.11$) and what targets had been asked to think about (Greenhouse-Geisser adjusted F (2.22, 108.91) = 1991.22, $p < .001$, Cohen's $f = 6.19$), and a significant interaction between the two factors (Greenhouse-Geisser adjusted F (2.02, 99.17) = 1019.67, $p < .001$, Cohen's $f = 4.43$).

Simple-effect analyses for the interaction between the EQ groups and the four kinds of target thought revealed the following results. Firstly, in each EQ group,

perceivers rated the valence of target expressions differently according to what targets had been asked to think about: for the low EQ group, Greenhouse-Geisser adjusted F (1.50, 73.58) = 681.74, $p < .001$, Cohen's $f = 3.62$; for the average EQ group, Greenhouse-Geisser adjusted F (1.60, 78.48) = 1182.54, $p < .001$, Cohen's $f = 4.77$; for the high EQ group, Greenhouse-Geisser adjusted F (2.03, 99.37) = 1862.83, $p < .001$, Cohen's $f = 5.99$. Secondly, for each kind of target thought, perceivers rated target expressivity differently between the three EQ groups: for happiness, Greenhouse-Geisser adjusted F (1.73, 84.81) = 861.31, $p < .001$, Cohen's $f = 4.07$); for gratitude, Greenhouse-Geisser adjusted F (1.54, 75.62) = 3840.31, $p < .001$, Cohen's $f = 8.59$; for sadness, Greenhouse-Geisser adjusted F (1.32, 64.63) = 144.15, $p < .001$, Cohen's $f = 1.66$; for anger, (F (2, 98) = 475.40, $p < .001$, Cohen's $f = 3.02$.

To examine how perceiver ratings of target expressions for each kind of target thought depends on target EQs, post hoc LSD tests were carried out on the main effects associated with kind of target thought. Results were as follows: (1) when targets were thinking of a time they felt happy, perceivers rated their expressions more positively if they were in the average EQ group than if they were in the low and high EQ groups ($ps < .001$); (2) when targets were thinking of a time they felt grateful, perceivers rated their expressions most positively when those targets were in the high EQ group ($ps < .001$), and more positively when they were in the low EQ group than in the average EQ group ($p = .001$); when targets were thinking of a time they felt sad, perceivers rated their expressions most positively when they were in the high EQ group ($ps < .001$), and rated equally those in the low EQ and the average EQ

groups; when targets were thinking of a time they felt angry, perceivers rated them more positively when those targets were in the average EQ group than when they were in the other two groups (*ps* < .001), and more positively in the high EQ group than in the low EQ group (*p* < .001).

General Discussion

Study 2 revealed that perceivers rated target expressions most positively when those targets were thinking of a time they felt happy; they rated targets most negatively when those targets were thinking of a time they felt sad. Consistent with this, Study1 showed that perceivers were generally able to detect the thoughts of targets when they were recalling either a happy or a sad event. Taken in isolation, these associations raise the possibility that perceivers based their judgments of target inner states on their classification of target facial expressions (but see Kang et al, 2018).

However, perceivers' ratings of target expressions (Study 2) were rather different than their inferences of target thoughts (Study 1) in many other respects. For example, in spite of rating target expressions positively when those targets had high EQ and were cued to think of a time they felt happy and a time they felt grateful, perceivers were inaccurate in inferring thoughts of happiness and gratitude (Study 1); but perceivers were more accurate in inferring that targets were thinking of a time they felt grateful if those targets had low EQ. In addition, perceivers rated expressions most positively for those targets who were in the average EQ group when thinking of

a time they felt happy; but they were most accurate in inferring happy thoughts in the low EQ target group (in Study 1). Perceivers generally rated expressions negatively when targets were thinking of a time they felt sad and yet were accurate in inferring the thoughts of sadness specifically in targets located in the high EQ group. In summary, nuances in the pattern of perceivers' accurate inferences of four kinds of target inner states across three target groups who differed in their EQ is far from fully illuminated by perceivers' ratings of target expressions. In short, it seems perceiver inferences of target inner states amounts to more than merely rating expressions as positive or negative, a conclusion which is highly consistent with that drawn by Kang et al. (2018). Presumably, then, the quality rather than the valence/ strength of target expression is what signals their inner states. Precisely what form these signals take is beyond the scope of the design and methods of the current study and remains something to pursue in future research.

Nevertheless, the results offer new information concerning people's ability to read others' minds and we shall summarise the highlights. Although there was an equal number of each of the four target events presented, perceivers did not impute an equal number of states; rather, they were biased to judge that targets were recalling a sad event. According to West and Kenny (2011), many findings in mindreading research are unclear in cases where biased responding might be an issue. According to them, the problem can only be solved by satisfying the 'truth condition' such that a measure of mindreading accuracy can be separated from response bias. The 'truth condition', as they define it, is satisfied if we can compare the perceiver's judgment

against an objective fact and our method was designed to do this. Specifically, when the perceiver judged, for example, that the target had been asked to think of something that made them feel grateful, we can then compare this judgment against the objective fact of whether or not the target was actually asked to think of something that made them feel grateful. Using the method of SDT for coding data controls for biased responding and uneven base rates; it is then possible to focus on mindreading accuracy that stands apart from issues with base-rate bias.

Using an unbiased measure (SDT) of mindreading accuracy, the results revealed notable performance in that by observing a short silent video of targets, the perceivers were systematically able to determine whether those targets were thinking of something happy and something that made them feel sad. The targets were merely sitting quietly while thinking: They were not asked to act in any way, they were not communicating and they were not engaging with anything external. The results are thus striking in showing that perceivers can observe somebody who is sitting quietly and guess what they are thinking. In addition, because the method and data-coding allows us to separate response-bias from mindreading accuracy, the findings reported here are perhaps the strongest and clearest demonstration to date of this aspect of human ability (cf Teoh et al, 2017).

We assume that the target's thought leaked out into their behaviour, taking the form of a mind that was perceptible to the perceiver. The perceivers then presumably translated by way of inference, more or less precisely, the observable target behaviour into an internal target state (Gallese & Goldman, 1998). It could have been that

perceivers were only crudely able to discriminate between occasions when targets were thinking of something positive and something negative, but nothing more precise, as was the case in past research (North, Todorov, & Osherson, 2010; Valanides et al., 2017). Impressively, though, the results here show that perceivers demonstrated levels of accuracy in a finer-grained four-way discrimination.

The finding that mindreading accuracy varies depending on the EQ status of the target supports Wu et al.'s (2016a & 2016b) general prediction that targets located at various points along the continuum might possess minds that vary in their level of readability. However, Wu et al. had not considered the possibility that target readability depends on their EQ status in combination with the particular content of thought targets were experiencing. Hence, the results reveal a complexity in the demands placed on perceivers that had not previously been anticipated or considered. Those with low EQ were most readable while those with high EQ were least readable: Why do specifically positive thoughts leak out as an interpretable signal more lucidly in targets with low EQ than in targets with high EQ? Perhaps positive thoughts have a different content or quality in those with low EQ compared with those who have high EQ – indeed, perhaps targets differed in their willingness or ability to think of something on cue, depending on their EQ status. It will surely be a challenge for future research to detail a link between thought content, quality of signal and EQ status in targets.

According to the 'lens model' (Back, Schmukle, & Egloff, 2011), accuracy in mindreading might be decided by the clues related with different factors—the target,

the perceiver and the interaction between them. In terms of the targets, they might behave in different ways depending on how empathizing they are. For example, the targets might emit different kinds of signals, including facial expressivity and bodily movements. Low EQ targets might show more positive signals when thinking something positive while high EQ targets might emit rather strong negative signals when recalling a sad event. Further research could test such possibilities by coding targets' signals (facial expressions and bodily movements) and explore the ways by which they play a role in perceiver judgments of target minds. Another possibility is that perceivers might have adopted different strategies to interpret targets with different levels of EQ. Future research could employ eye-tracking along with behavioural measurements to examine whether or not perceivers use different strategies to observe targets with different EQ levels.

Previous studies have demonstrated perceiver abilities to detect which situation caused a target's reaction (Pillai et al., 2012, 2014; Cassidy et al., 2013, 2015; Sheppard et al., 2016; Teoh et al., 2017), to infer how others felt (Zaki et al., 2008; 2009), to infer what another person is thinking (Ickes et al., 1990; Valanides et al., 2017), and to judge where a stranger is located along trait continua (Wu et al., 2016a, 2017); the current research expanded these findings by suggesting perceiver capability in inferring specified target thoughts, and the accuracy of such mindreading, was affected by target EQ status as well as the events experienced by the target.

References

Andrews, K. (2008). It's in your nature: A pluralistic folk psychology. *Synthese, 165,* 13-29. doi:10.1007/s11229-007-9230-5

Back, M. D., Schmukle, S. C., & Egloff, B. (2011). A closer look at first sight: Social relations lens model analysis of personality and interpersonal attraction at zero acquaintance. *European Journal of Personality, 25*, 225-238. doi: 10.1002/per.790

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger Syndrome or High Functioning Autism, and normal sex differences. *Journal of Autism and Developmental Disorders, 34,* 163-175. doi:10.1023/B:JADD.0000022607.19833.00

Baron-Cohen, S., (2012). *Zero Degrees of Empathy: A New Theory of Human Cruelty and Kindness*. London: Penguin Books.

Brewer, R., Biotti, F., Catmur, C., Press, C., Happé, F., Cook, R., & Bird, G. (2016). Can neurotypical individuals read autistic facial expressions? Atypical production of emotional facial expressions in Autism Spectrum Disorders. *Autism Research, 9,* 262–271. doi: org/10.1002/aur.1508

Cassidy, S., Ropar, D., Mitchell, P., & Chapman, P. (2013). Can adults with autism spectrum disorders infer what happened to someone from their emotional response? *Autism Research, 7*, 112-123. doi:10.1002/aur.1351

Cassidy, S., Ropar, D., Mitchell, P., & Chapman, P. (2015). Processing of spontaneous emotional responses in adolescents and adults with Autism Spectrum Disorders: effect of stimulus type. *Autism Research, 8*, 534-544. doi:10.1002/aur.1468

543 Faso, D. J., Sasson, N. J., & Pinkham, A. E. (2015). Evaluating Posed and Evoked
544 Facial Expressions of Emotion. *Journal of Autism and Developmental Disorders*, *45*,
545 75-89. doi: 10.1007/s10803-014-2194-7
546 Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses
547 using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research*
548 *Methods*, *41*, 1149-1160.
549 Flavell, J. H., Miller, P. H., & Miller, S. A. (1993). *Cognitive development (3rd ed.).*
550 Englewood Cliffs, NJ: Prentice-Hall.
551 Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of
552 mindreading. *Cognition, 2(12),* 493-501. doi:10.1016/S1364-6613(98)01262-5
553 Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social
554 cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social*
555 *Psychology, 59 (4),* 730-742. doi:10.1037/0022-3514.59.4.730
556 Kang, K., Anthoney L., & Mitchell, P. (2017). Seven- to 11-year-olds' developing
557 ability to recognize natural facial expressions of basic emotions. *Perception, 46*,
558 1077-1089. doi: 10.1177/0301006617709674
559 Kang, K., Schneider, D., Schweinberger, S.R. & Mitchell, P. (2018). Dissociating
560 neural signatures of mental state retrodiction and classification based on facial
561 expressions. *Social Cognitive and Affective Neuroscience*. 933-943. doi:
562 10.1093/scan/nsy061
563 MacMillan, N. A. (2002). Signal detection theory. In H. Pashler (Ed.), *Stevens'*
564 *handbook of experimental psychology* (3rd ed.). J. Wixted (Ed.), *Vol. 4: Methodology*

in experimental psychology (pp. 43–90). New York, NY: John Wiley & Sons.

MacMillan, N. A.,&Creelman, C. D. (2005). *Detection theory: A user's guide (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

North, M. S., Todorov, A., & Osherson, D. N. (2010). Inferring the preferences of others from spontaneous, low-emotional facial expressions. *Journal of Experimental Social Psychology, 46,* 1109-1113. doi:10.1016/j.jesp.2010.05.021

Pillai, D., Sheppard, E., & Mitchell, P. (2012). Can people guess what happened to others from their reactions? *PLoS ONE, 7(11),* e49859. doi:10.1371/journal.pone.0049859

Pillai, D., et al. (2014). Using other minds as a window onto the world: Guessing what happened from clues in behaviour. *Journal of Autism and Development Disorders, 44,* 2430-2439. doi:10.1007/s10803- 014-2106-x

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *The Behavioral and Brain Sciences, 4,* 515-526. doi: 10.1017/S0140525X00076512

Rai, R. & Mitchell, P. (2004). Five-year-olds' difficulty with false belief when the sought entity is a person. *Journal of Experimental Child Psychology, 89,* 112-126. doi: 10.1016/j.jecp.2004.05.003

Russell, J. A., Bachorowski, J. A., & Fernandez-Dols, J. M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology, 54,* 329-349. doi:10.1146/annurev.psych.54.101601.145102

585 Sheppard, E., Pillai, D., Wong, G. T., Ropar, D., & Mitchell, P. (2016). How easy is it
586 to read the minds of people with Autism Spectrum Disorder? *Journal of Autism and*
587 *Developmental Disorders*, *46*, 1247-1254. doi:10.1007/s10803-015-2662-8
588 Soscia, I. (2007). Gratitude, delight, or guilt:The role of consumer's emotions in
589 predicting postconsumption behaviors. *Psychology & Marketing*, *24 (10)*, 871-894.
590 doi:10.1002/mar.20188
591 Teoh, Y., Wallis, E., Stephen, I.D., & Mitchell, P. (2017). Seeing the world through
592 other minds: Inferring social context from behaviour. *Cognition*, *159*, 48-60.
593 doi:10.1016/j.cognition.2016.11.003
594 Valanides, C., Sheppard, E., & Mitchell, P. (2017). How accurately can ohter people
595 infer your thoughts—And does culture matter? *PLoS ONE, 12(11)*: e0187586.
596 doi:10.1371/journal.pone.0187586
597 West, T. V. & Kenny, D. A. (2011). The truth and bias model of judgment.
598 *Psychological Review*, *118*, 357-378. doi:10.1037/a0022936
599 Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and
600 constraining function of wrong beliefs in young children's understanding of
601 deception. *Cognition*, *13*, 103-128. doi:10.1016/0010-0277(83)90004-5
602 Wu, W., Sheppard, E., & Mitchell, P. (2016a). Being Sherlock Holmes: Can we sense
603 empathy from a brief sample of behaviour? *British Journal of Psychology, 107 (1)*, 1-
604 22. doi:10.1111/bjop.12157

Wu, W., Sheppard, E., & Mitchell, P. (2016b).The game is afoot: A response to three insightful commentaries. *British Journal of Psychology, 107 (1),* 33-35. doi:10.1111/bjop.12168

Wu, W., Sheppard, E., & Mitchell, P. (2017) Judging personality from a brief sample of behavior: It is relatively easy to detect who is unique on a trait. *European Journal of Personality, 31,* 685-700. doi:10.1002/per.2116

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

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion, 9,* 478-487. doi:10.1037/a0016551

Zaki, J. & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry, 22,* 159-182. doi:10.1080/1047840X.2011.551743

Tables

Table 1. Means and standard deviations (*SD*) of hit rates (M_{HR}), false alarm rates (M_{FAR}), d-prime ($M_{d'}$), criterion (M_c) of each mental state across the three target EQ groups and within each target EQ group, along with *t* values of one-sample *t* tests for d' (comparing with 0), 95% confidence intervals (95% CIs) of each $M_{d'}$ and Cohen's *d* in Study 1

Target EQs	States	M_{HR}	M_{FAR}	$M_{d'}$	M_c	95%CIs	*t*	Cohen's *d*
Across	H	.27 (.10)	.22 (.08)	.17 (.29)	.73 (.28)	[.09, .25]	4.24***	.57
three	G	.25 (.12)	.24 (.08)	0 (.34)	.74 (.32)	[-.09, .10]	.04	0
EQs	A	.22 (.08)	.21 (.07)	.03 (.29)	.83 (.24)	[-.06, .11]	.63	.10
	S	.35 (.10)	.31 (.10)	.11 (.26)	.45 (.24)	[.04, .19]	3.13**	.42
Low EQ	H	.29 (.13)	.17 (.10)	.48 (.42)	.82 (.38)	[.36, .59]	8.25***	1.14
	G	.30 (.18)	.23 (.09)	.17 (.53)	.68 (.37)	[.02, .32]	2.31*	.32
	A	.21 (.12)	.20 (.08)	-.03 (.51)	.89 (.30)	[-.17, .12]	-.69	-.06
	S	.36 (.15)	.35 (.14)	.03 (.47)	.40 (.33)	[-.10, .16]	.42	.06
Average	H	.28 (.14)	.22 (.09)	.18 (.45)	.71 (.28)	[.06, .30]	2.91**	.40
EQ	G	.26 (.15)	.22 (.11)	.08 (.51)	.77 (.38)	[-.06, .22]	1.10	.16
	A	.23 (.11)	.21 (.08)	.05 (.35)	.82 (.32)	[-.05, .15]	1.03	.14
	S	.34 (.13)	.31 (.09)	.10 (.39)	.48 (.26)	[-.01, .21]	1.76	.26
High EQ	H	.23 (.15)	.30 (.11)	-.25 (.53)	.70 (.36)	[-.40, -.11]	-3.47***	-.47
	G	.17 (.13)	.26 (.11)	-.33 (.55)	.86 (.33)	[-.48, -.18]	-4.31***	-.60
	A	.25 (.16)	.21 (.07)	.08 (.50)	.79 (.34)	[-.06, .22]	1.20	.16
	S	.35 (.15)	.26 (.13)	.25 (.49)	.57 (.37)	[.11, .39]	3.67***	.51

Notes: $p^* < .05$, $p^{**} < .01$, $p^{***} \leq .001$; Cohen's d = 0.2, 0.5 and 0.8 respectively represents small, medium and large size; A = Anger, H = Happiness, S = Sadness, and G = Gratitude.

Table 2. Means of perceiver ratings (M_R) and standard deviations (SD) of each mental state in each target EQ group and across the three EQ groups, along with t values of one-sample t tests (comparing with the neutral point 3), 95% confidence intervals (95% CIs) of each M_R and Cohen's d in Study 2

Target EQs	States	M_R (SD)	95% CIs	t	Cohen's d
Across three EQs	H	3.10 (.03)	[3.09, 3.11]	27.16*	-3.33
	G	3.00 (.02)	[3.00, 3.01]	.39	0
	A	2.96 (.02)	[2.96, 2.97]	-49.59*	-2.00
	S	2.86 (.02)	[2.86, 2.87]	-17.57*	-7.00
Low EQ	H	3.04 (.03)	[3.04, 3.05]	11.28*	1.33
	G	2.84 (.03)	[2.83, 2.85]	-35.20*	-5.33
	A	2.84 (.02)	[2.84, 2.85]	-52.70*	-8.00
	S	2.82 (.03)	[2.81, 2.83]	-37.19*	-6.00
Average EQ	H	3.23 (.03)	[3.22, 3.24]	52.85*	7.67
	G	2.81 (.04)	[2.80, 2.82]	-33.48*	-4.75
	A	3.04 (.03)	[3.03, 3.05]	10.14*	1.33
	S	2.83 (.05)	[2.81, 2.84]	-22.45*	-3.40
High EQ3	H	3.04 (.04)	[3.03, 3.05]	6.68*	1.00
	G	3.35 (.02)	[3.34, 3.35]	107.30*	17.50
	A	3.00 (.04)	[2.99, 3.01]	.03	0
	S	2.94 (.01)	[2.94, 2.94]	-30.26*	-6.00

Note: $p^* < .001$

Figures

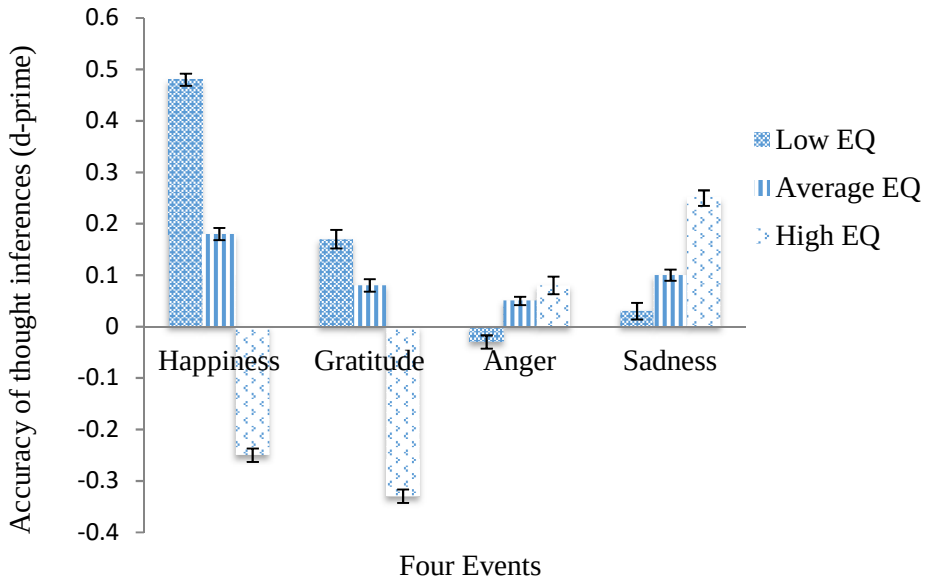

Figure 1. Mean accuracy (d-prime) of perceivers' inferences of the target mental states (happiness, gratitude, anger, and sadness) in each target EQ group in Study 1. Error bars represent standard error of the mean with 95% confidence interval.

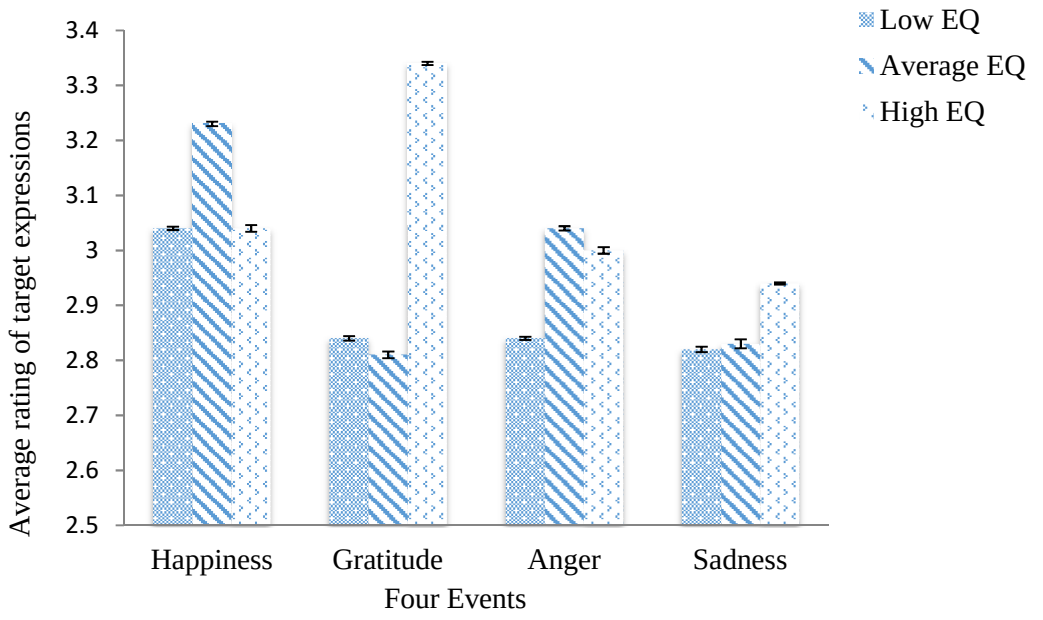

Figure 2. Mean perceiver ratings of target expressions as targets were thinking of the four events in each target EQ group in Study 2. Error bars represent standard error of the mean with 95% confidence interval.

685