Running head: *CAN WE SENSE EMPATHY?*

Being Sherlock Holmes: Can we sense empathy from a brief sample of behavior?

Wu, W., Sheppard. E. & Mitchell, P. (2016). Being Sherlock Holmes: Can we sense empathy from a brief sample of behavior? *British Journal of Psychology, 107*, 1-22

Abstract

Mentalising (otherwise known as ‘theory of mind’) involves a special process that is adapted for predicting and explaining the behavior of others (targets) based on inferences about targets’ beliefs and character. The current research investigated how well participants made inferences about an especially apposite aspect of character, empathy. Participants were invited to make inferences of self-rated empathy after watching or listening to an unfamiliar target for a few seconds telling a scripted joke (or answering questions about him/herself or reading aloud a paragraph of promotional material). Across three studies participants were good at identifying targets with low and high self-rated empathy but not good at identifying those who are average. Such inferences, especially of high self-rated empathy, seemed to be based mainly on clues in the target’s behavior, presented either in a video, a still photograph or in an audio track. However, participants were not as effective in guessing which targets had low or average self-rated empathy from a still photograph showing a neutral pose or from an audio track. We conclude with discussion of the scope and the adaptive value of this inferential ability.

Key Words: mentalising, empathy, zero-acquaintance, behavior, inference

Being Sherlock Holmes: Can we sense empathy from a brief sample of behavior?

People's minds are to some degree embodied in their behavior (Pillai, Sheppard, Ropar, Marsh, Pearson & Mitchell, 2014) and the ability to interpret behavior in order to 'read' minds is surely of great adaptive value. 'Mindreading', is the ability to predict and explain other people's behavior (Premack & Woodruff, 1978) by making inferences of what a person believes based on the situation they are in (e.g., Dennett, 1978; Wimmer & Perner, 1983) and by making inferences of the traits of the target (e.g. Jones & Davies, 1965; Andrews, 2008). Being able to infer a person's (henceforth a target's) simple factual belief *states* allows us to predict or explain the target's behavior, meaning that we can be proactive rather than just reactive in social contexts. Being able to infer aspects of a target's character, or *traits*, especially their level of empathy, is similarly adaptive in that it is useful, amongst other things, to know who is and who is not likely to be kind, supportive and caring.

This dual aspect of mindreading (inferring states and traits from clues in behavior) was fictionalized in the figure of Sherlock Holmes (Conan-Doyle, 1902) who was able to explain a target's behavior by imaginatively projecting himself into the target's perspective and then using his own mind to model what happened to the target to cause him to behave in such a way, a process we now call 'mindreading by mental simulation' (Mitchell, Currie & Zeigler, 2009). For this to be effective we must be able to infer the character of the target in order to know what kind of adjustment is needed when simulating the target's mind; and for the reason given above, perhaps one of the most important traits is empathy. Apart from helping with the process of mental simulation, knowing who is and who is not empathic could be important in determining who our friends are likely to be.

The current contribution is rather different from past research into 'theory of mind', which has tended (a) to concentrate on bias in reasoning and errors in the judgments of

people engaged in working out what others are thinking (b) to focus on people's ability to infer mental *states* while perhaps neglecting people's ability to infer *traits* (Apperly, Simpson, Riggs, Samson & Chiavarino, 2006; Keysar, Lin & Barr, 2003; Mitchell, Robinson, Isaacs & Nye, 1996). Meanwhile, research into person perception (exemplified in the special issue of the *British Journal of Psychology* published November 2011) has tended not to articulate implications for processes in mindreading. The current contribution makes a first step to bridge the gap between these two research traditions (cf. Zaki & Ochsner, 2011) by investigating how well people can detect the level of empathy (as a trait not as a state) in a target in a context that is relevant to mindreading.

A large volume of research has already investigated whether or not various traits are perceptible (e.g., Albright, Kenny & Malloy, 1988; Borkenau & Liebler, 1993; Funder, 1995; Borkenau, Mauer, Riemann, Spinath & Angleitner, 2004; McLarney-Vesotski, Bernieri & Rempala, 2006; Carney, Colvin & Hall, 2007; Todorov, Pakrashi & Oosterhof, 2009; Thoresen, Vuong & Atkinson, 2012) but the current research has a slightly different purpose. Past research only tells us that people (henceforth 'perceivers') can systematically infer personality traits; it does not tell us whether people can more easily infer who is at the extremes compared with who is in the middle of a continuum. The aim, then, was to investigate how well observers can detect that another person (the target) has *high, middle or low* empathic status: Is it just as easy to detect targets located at any point on the empathy scale or is it easier to detect who is at one part of the scale rather than who is at another part of the scale? The results will inform us how well perceivers guess different levels of empathy, knowledge that will contribute to our broader understanding of how accurate people are in "mindreading."

Research into mindreading has shown that perceivers seem to have a talent for inferring a state based on scant or fleeting information. They can to some extent accurately

judge what another person might be thinking and feeling in a brief unstructured dyadic interaction (e.g., Ickes, Stinson, Bissonnette & Garcia, 1990; Ickes, 2003, for a review; Hall & Mast, 2007); they can infer the affect of a target while observing the target talking about an autobiographical event (Zaki, Bolger, & Ochsner, 2008; 2009); they can guess fairly well what occurred to a target after viewing a video lasting only a few seconds of the target reacting in a real-life scenario (Cassidy, Ropar, Mitchell & Chapman, 2013; 2015; Pillai, Sheppard, & Mitchell, 2012; Pillai et al., 2014).

It seems people are adept in explaining behavior by making inferences about the situation that elicits a *state* of empathy in a target. For example, on seeing the target's reaction (presented in a short silent video), perceivers could guess that he or she was listening to someone recounting their difficult day (and not listening to someone telling a joke, paying a compliment or being rude; Pillai et al., 2012). Note that different targets experiencing exactly the same situation reacted in various ways and yet perceivers were nevertheless able to infer that a range of reactions were caused by the same situation. It seems perceivers understand that different people behave in different ways in a given situation, implying that they might understand that how one reacts depends not just on the situation but also on the personality of the target. This background study therefore leads us to ask how accurately perceivers can guess aspects of personality, such as empathy, from a brief sample of behavior.

Empathy is experienced as a state induced by various situations, and the capacity to have this experience varies between people as an empathic disposition that reflects relatively consistent characteristic patterns of behavior and thought pertaining to empathy (Zhou, Valiente & Eisenberg, 2003; Rumble, Van Lange & Parks, 2010). Research has well documented people's ability to infer traits based on limited information about their behavior: Perceivers spontaneously, routinely and rapidly make trait inferences in the absence of any

prompting or cuing when trying to explain or predict the behavior of others (e.g., Winter & Uleman, 1984; Todorov & Uleman, 2004; McCarthy & Skowronski, 2011). Moreover, by watching the gait of a person, perceivers could swiftly make a reliable trait judgment (Thoresen et al., 2012); by either viewing a short video when a target was performing a trivial activity, such as reading a standard weather forecast (Borkenau & Liebler, 1993), telling a joke, or singing a song (Borkenau et al., 2004) or having a get-acquainted conversation with others (Carney et al., 2007), perceivers could quickly guess some dimensions of the big-five traits of the target. The evidence thus converges in suggesting that facial expressions and behavioral manners (including speech) offer clues about mental states and psychological dispositions, and it seems that observers are effectively able to perceive and interpret this information.

Adopting an accuracy-oriented method (Funder, 1995; Funder, 2012, for a review), where judgments of traits about others are compared with a criterion of accuracy (such as self-report ratings), the current study asked perceivers to guess the target's self-rated empathy, allowing us to compare perceiver inferences against an objective criterion. Because traits give rise to fleeting patterns of behavior that transcend time and specific situations (Funder, 1991), we extracted 'thin slices' (occupying less than five minutes –Ambady, Bernieri & Richeson, 2000) from ongoing behavior. The zero-acquaintance procedure, where the perceiver is asked to judge a target's psychological traits neither with acquaintance nor prior knowledge (Norman & Goldberg, 1966; Albright et al., 1988), was used to ensure that perceivers made a judgment of empathy on the basis of the presented thin slices of behavior rather than any prior knowledge of the target.

Because of its wide application and putative reliability and validity (Lawrence, Shaw, Baker, Baron-Cohen, & David, 2004), this study adopts the 'empathy quotient' (EQ) scale, developed by Baron-Cohen and Wheelwright (2004), to measure self-rated empathy. The EQ

questionnaire offers a comprehensive measurement of the psychological structure of empathy covering both cognitive and affective factors. It comprises 40 items pertaining to a range of behaviors associated with empathizing, with an overall rating that is useful in determining individual differences in empathic tendencies. All targets completed the EQ questionnaire, and their EQs served as the criterion for gauging whether perceivers could guess how empathizing the targets were.

Perceivers in the current research observed short samples of videoed behavior of targets (for example reading aloud the script of a joke) and then guessed the targets' self-rated empathy. Given that mindreading is normally done rapidly in a context of fleeting information, we also aimed to discover how little information is sufficient for making successful inferences of self-rated empathy.

Study 1

Method

Participants

The participants (the "perceivers") were 90 students (41 males) aged 18 to 32 years (M = 21 years) recruited from the University of Nottingham Malaysia campus. Perceivers were shown photographs of the targets (taken from their videos) and were included only if they reported not seeing any of the targets previously. The perceivers were randomly assigned to view targets either in the Conversation, the Joke or the Screen Test Scenario. Details of targets appear below.

Materials

One hundred and forty-one video clips were developed as stimuli, with 47 clips in each condition where the targets were videoed during the conversation, reading a joke or

doing the screen test. Three groups of 30 perceivers viewed 47 clips showing the targets either in the Conversation, the Joke or the Screen Test Scenario. All the videos were presented using the software PsychoPy (Peirce, 2007) on a laptop.

Video stimuli collection and editing. A Sony Handycam DCR-SR60 video camera filmed the targets. Videos were collected from 47 targets recruited from the University of Nottingham Malaysia campus (24 males) aged 18 to 32 years (M = 21 years), all of whom responded to a call to do a screen test advertising the University. On arrival, targets were issued with a script of the joke and the screen test. All were individually videoed in a quiet laboratory with the camera mounted on a tripod placed approximately 1.2 meters away to record the target's face and the top part of their body. The researcher sat opposite to the target but out of view of the camera. Unknown to the target, the camera automatically began recording as soon as the target entered the room. (Subsequently, all targets were fully debriefed and gave written informed consent to use the videos for research purposes.) Once inside the laboratory, after the target read some information (including an information sheet, a script of the joke, a script of the screen test and a consent form), the researcher began with a brief conversation in which she asked a series of questions (and wrote down the responses) about the target's name, age, what course they were enrolled on, where they were from and so on. The conversation lasted approximately two minutes. The camera was then ostensibly switched to 'record mode' and the target was invited to read out the joke to the camera. After a pause of about one minute the target was then invited to read aloud a verbatim script of the screen test.

Following a short break for a couple of minutes, the target was asked to fill in the EQ questionnaire, whose scores ranged from 19 to 61 (M = 37.96, SD = 10.19). A score in the range of 0-32 is low EQ and 12 targets were in this category, 33-52 is average and 31 targets

were in this category, 53-63 is above average and 4 targets were in this category, and 64-80 is high but no targets were in this category (Baron-Cohen, 2012). In order to maintain four categories, we split the 'average' category into two ranging from 33 to 41 (20 targets) and 42 to 52 (11 targets), and combined the 'above average' and 'high' categories into one range from 53 to 80. We re-label these four categories as Scale 1 (12 targets, 6 males), Scale 2 (20 targets, 10 males), Scale 3 (11 targets, 5 males) and Scale 4 (4 targets, 3 males), where Scale 1 is lowest EQ and Scale 4 is highest EQ.

Three separate video clips were made for each target. In the Joke and the Screen Test Scenario, each video clip began when the target started the task and ended about two seconds after the target completed reading the script. The average duration of the video clips was 30.87s (*SD* = 2.56; ranging from 24s to 35s) for the Conversation, 8.94s (*SD* = 1.36; ranging from 7s to 12s) for the Joke and 29.36s for the Screen Test (*SD* = 4.48; ranging from 22s to 42s). Because the raw filming of the Conversation actually lasted around two minutes, we extracted 30-second clips from either the beginning (15 targets selected at random), the middle (16 targets selected at random) or the end (16 targets selected at random). For all three scenarios videos were edited to show active speaking parts from the target.

Procedure

Perceivers were tested individually and began by completing the EQ questionnaire, which took approximately 10 minutes. After completion they were told that the questionnaire measures empathizing and then were given an EQ information sheet containing brief definitions about empathy and the EQ as well as the meanings of the four EQ ranges:

"The questionnaire you completed is called the Empathy Quotient (EQ), which measures empathy. Empathy is the ability to understand what another person might be

thinking or how a person is feeling, and to respond to the mental and emotional states of the person with an appropriate emotion (Baron-Cohen & Wheelwright, 2004). Typical adults get an EQ score in any of the following ranges from low to high: 0-32, 33-41, 42-52, and 53-80, where 0-32 means low EQ, 33-41 and 42-52 are average EQ and 53-80 are high EQ. A low, average and high EQ respectively means low, average and high empathic capability."

The researcher explained that she did not have time to score the perceiver's EQ and would do it later. Subsequently, each of a set of 47 video clips was presented in random order to each perceiver using PsychoPy on a laptop. Following the presentation of each video clip a response screen appeared, displaying the four-point EQ scale (Scale 1 to Scale 4, from low to high EQ) as response options for the EQ rating. The perceiver registered his/her assessment of the target's EQ by using the mouse to click on the relevant point on the four-point scale. Once the perceiver made the choice the screen moved to the next video clip. Responses were automatically recorded by the software for later retrieval. Perceivers typically took about 15 minutes in the Joke Scenario and about 40 minutes in the Screen Test Scenario and the Conversation Scenario to view and rate the videos.

Results and discussion

Preliminary analysis

The average EQs of perceivers were 38.77 (*SD* = 10.63, ranging from 19 to 58) in the Conversation Scenario, 37.07 (*SD* = 8.31, ranging from 19 to 56) in the Joke Scenario, and 39.47 (*SD* = 8.40, ranging from 23 to 56) in the Screen Test Scenario. Independent-samples *t* tests did not identify any differences between the average EQs of the targets (*M* = 38.43, *SD* = 9.13, ranging from 19 to 58) and the perceivers and neither was there any evidence of difference among the three groups of perceivers in their average EQ scores according to a

one-way ANOVA. Independent-samples *t* tests did not reveal a gender difference in the EQ scores of the targets or the perceivers.

Three separate groups of perceivers rated the EQ of each of the 47 targets. Hence, each target had three mean ratings of EQ scores from perceivers who watched the conversation, perceivers who watched the target tell a joke and perceivers who watched the target do a screen test. The intercorrelations between these three sets of scores were high: r = .67 between Joke and Screen Test, r = .62 between Joke and Conversation and r = .68 between Screen Test and Conversation (p <. 001 in all cases). Hence, irrespective of how accurate perceivers were in their judgments (see below), at least it seems that samples of target behavior across three scenarios led to rather consistent ratings of EQ by the perceivers.

Main analysis – Guessing the EQ of the target

To begin, two EQ scores were coded for each target: their true self-rated EQ score (M =2.15, SD =. 91) and their mean estimated score (M =2.55, SD =. 38) as rated by a total of 90 perceivers (perceiver ratings were combined across the three scenarios). This yielded a weak but significant value (r =. 29, p <. 05) that is only partially informative because it tells us nothing about whether target EQs in one part of the EQ scale were easier to perceive than in another part of the scale. To illuminate that matter, a different kind of analysis was adopted, as explained below.

Inspection of the data revealed that base rates in perceivers' judgments of the targets' EQ varied widely across the four scales. As shown in Table 1 and Table 2, it was common for perceivers to judge that targets were in the middle two scales but less common for them to guess that targets were in the two extreme scales in each condition across the three studies. Because perceivers frequently judged targets to be in the middle scales, in absolute terms

there were a fairly large number of “correct” judgments. If giving such a correct judgment served as the index of accuracy, then we would have a false impression that perceivers were good at guessing which targets were in the middle scales.

Table 1 about here

Table 2 about here

Given that signal detection theory (SDT) allows an assessment of accuracy and sensitivity that is immune to response bias (the tendency to select one category more frequently than another; Stanislaw & Todorov, 1999; Macmillan, 2002; Macmillan & Creelman, 2005), it is widely applied to measure performance across various tasks, including those that employ multi-way forced choice procedures (Macmillan, 2002; Macmillan & Creelman, 2005), as illustrated in the following examples: SDT has been used to examine accuracy in mental state inferences (Pillai et al., 2012; 2014), eyewitness’s identification of suspects (Clark, 2012), recognition of suicide tendency (Kleiman & Rule, 2013), diagnostic decisions more generally (Swets, Dawes, & Monahan, 2000) and optimal decision making (Lynn & Barrett, 2014).

Following published guidelines on calculating SDT (Macmillan, 2002; Macmillan & Creelman, 2005), a correct judgment that a target belonged to a particular EQ scale counted as a ‘hit’ while an incorrect judgment that a target belonged to the same EQ scale counted as a false alarm; the accuracy for identifying each EQ scale over a total of 47 trials was then calculated as a single value for each perceiver in the form of d-prime (d’). Following Macmillan and Creelman (2005), the hit and false alarm calculations were corrected by adding 0.5 when there were either no hits or no false alarms; d’ is 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)*, $0 \leq p \leq 1$).

Table 3 displays the means of the hit rates, false alarm rates and mean d-prime ($M_{d'}$) in each category of the four-point scale in each condition, along with the *t* values of the one-sample *t* tests of each $M_{d'}$ where the comparison value is zero: If perceivers were unable to detect targets in any particular EQ scale ranges, this would yield a $M_{d'}$ of zero for that scale. According to the one-sample *t* tests shown in Table 3, perceivers were not uniformly effective in guessing EQ in each EQ scale: Perceivers made systematically correct judgments (indicated by $M_{d'}$ well above zero) in the cases of low (Scale 1) and high (Scale 4) EQ, while in all but one case (Scale 2 in the Joke scenario) there was no evidence of systematic judging for targets who were in the middle of the EQ range (Scales 2 and 3). Figure 1 demonstrates three U-shaped curvilinear trends associated with the three scenarios that reflect the pattern of *t* test results.

Table 3 about here

Figure 1 about here

To examine whether the perceivers differed across scales and scenarios in guessing EQ, a 3 × 4 mixed design ANOVA was computed, with the 3 scenarios as the between-subjects factor and the 4 EQ scales as the within-subjects factor; the dependent variable was d'[1]. There was a main effect associated with the scales, F (3, 261) = 53.08, $p < .001$, Cohen's f = 1.33, an interaction between Scale and Scenario, F (6, 261) = 3.01, p = .007, Cohen's f = .32, but no main effect of scenario, F (2, 87) = 2.66, p = .076.

Simple-effects analyses for the interaction between Scale and Scenario revealed two things. First, there were significant differences in judgmental accuracy among the four scales

and this trend was apparent for each scenario (Conversation: F (3, 87) = 20.82, p < .001, Cohen's f = .83; Joke: F (3, 87) = 26.70, p <. 001, Cohen's f = .94; Screen Test: F (3, 87) = 8.58, p < .001, Cohen's f = .53). Post hoc LSD analyses revealed no difference in $M_{d'}$ between Scales 2 and 3 and confirmed higher $M_{d'}$ in Scales 1 and 4 compared with Scales 2 and 3 in each of the three scenarios: In the Conversation Scenario, $M_{d'}$ in Scale 1 was greater than that in Scale 2, p = .029; $M_{d'}$ in Scale 4 was greater than in Scales 1, 2 and 3, ps < .001; in the Joke Scenario, $M_{d'}$ in Scale 1 was greater than in Scales 2 and 3, $ps \leq .004$, and $M_{d'}$ in Scale 4 was greater than in Scales 1, 2 and 3, $ps \leq .003$; in the Screen Test Scenario, $M_{d'}$ in Scale 1 was greater than in Scale 2 (p = .01) and Scale 3 (p = .004) while $M_{d'}$ in Scale 4 was greater than in Scales 2 and 3, $p \leq .002$.

Second, there was a significant difference among the three scenarios only in Scale 2 (F (2, 87) = 3.15, p = .048, Cohen's f = .32) and Scale 4 (F (2, 87) = 4.93, p = .009, Cohen's f = .41). As we see from Table 3 and Figure 1, in Scale 2 $M_{d'}$ was lower in the Joke Scenario than in the other two scenarios, and $M_{d'}$ in Scale 4 was lower in the Screen Test than in the other scenarios; these descriptive results were confirmed by post hoc LSD tests: For Scale 2, $M_{d'}$ in the Joke was lower than that in the Conversation (p = .029) and in the Screen Test (p = .037); for Scale 4, $M_{d'}$ in Screen Test was lower than in the Conversation (p = .006) and in the Screen Test (p = .01).

In summary, perceivers made systemically accurate judgments of the targets' EQs in different situations after watching a short video. They were especially good at identifying those who had extreme EQ but not so good in guessing those who had average EQ. Generally, perceivers performed better in Scale 4 in the Conversation and Joke Scenarios than in the Screen Test Scenario, and performed worse in Scale 2 in the Joke Scenario than in the other scenarios.

Why did perceivers fail to identify targets within Scales 2 and 3? Is this an artifact in the way Scale 2 and Scale 3 were derived from but a single EQ category or does it genuinely reflect perceivers' limitations in recognizing targets who were average in self-rated empathy? As reported in the method section, in order to maintain a four-point scale, we split the original average EQ scale into Scale 2 and Scale 3; is there a possibility that such sub-categorization has no psychological value and that perceivers are thus unable to make a distinction (between Scale 2 and Scale 3) that does not really exist? If this explanation is correct, then perceivers should be able to detect average EQ if we combined Scale 2 and Scale 3 to make a single average EQ scale. To examine this possibility, we counted the number of hits and false alarms in Scales 2 and 3 combined and used these to calculate d' for a composite Scale 2 and 3. The hits and false alarms for the two extreme scales were not affected and thus the values of d' for Scales 1 and 4 remained. The combined $M_{d'}$ was -.12, (*SD* = .71), a value that was not significantly above zero, t (89) = -1.60, p = .11.

Is it possible that the U-shaped trends in Figure 1 (and in the other figures) are an artifact of a 'range effect'? The dependent variable represented the number of times the perceiver guessed correctly (for example, accurately judging that the target was in EQ Category 1) weighted by the number of times the target used a given EQ category in error (a false alarm, such as judging that the target was in EQ category 1 when in fact he or she was in a different EQ category). Arguably, there were more opportunities for such false alarms when judging EQ in Scales 2 and 3 than when judging EQ Scales 1 and 4. When judging EQ Scale 1, for example, there is only one adjacent category (Scale 2) but when judging EQ Scale 3, for example, there are two adjacent categories (Scale 2 and Scale 4). Hence, the probability of a false alarm might be higher in the two middle EQ categories than in the two outer EQ categories. This could artifactually lead to the value of $M_{d'}$ being lower for the

middle categories than for the outer categories, thus giving rise to a spurious U-shaped function. We addressed this matter by statistically adjusting the four scales such that the opportunity for making false alarms was equalized and then we repeated all the analyses. The false alarm rate of judgments on each EQ scale was confined to cases where correct responses would have been the neighboring EQ scale (or the mean number of times the correct response would have been one or other of the two neighboring EQ scales in the case of Scales 2 and 3). The same U-shaped trends emerged even with these adjusted false alarm rates. This was true not only in the current study but also in the other studies reported in this article. It seems, then, that perceivers were limited in guessing the EQ of targets who were average in self-rated empathy and that the data indeed formed a U-shaped pattern.

Study 2

Based on the results of Study 1, it is tempting to conclude that watching a sample of behavior spanning just a few seconds is sufficient for the perceiver to estimate the target's EQ; they were especially effective in identifying targets who were either low or high in self-rated empathy. Perhaps identifying people at the extremes of trait continua is particularly adaptive: People who are average by definition behave according to situational norms while those who are not average do not, in which case predicting behavior depends on the perceiver's ability to take into consideration the target's traits. If it is the case that perceivers are especially sensitive to targets located at the extremes of self-rated empathy, then the U-shaped trend should emerge robustly in subsequent experiments, and this was tested in Studies 2 and 3.

Another purpose of Study 2 was to cast light on how perceivers inferred the level of empathy of the targets. Specifically, could perceivers make accurate assessments even after viewing still photographs of the targets? It seems fair to assume that the target's *behavior*

reveals their self-rated empathy and a still photograph of the target may not convey enough about their behavior for perceivers to make accurate inferences. Alternatively, does a still image of the target in any pose, even a neutral pose, provide sufficient information to identify high and low EQ? Recent research (Kramer & Ward, 2010; Jones, Kramer & Ward, 2012) suggests that people can identify high from low trait values in some personality dimensions after viewing a target's neutral pose for a few seconds and it might therefore be possible that perceivers in our Study 1 made accurate judgments of EQ based only on the appearance of the targets instead of any behavioral cues. Alternatively, is it that, for example, a photograph capturing the apex of the target's expression as he or she delivers the punch line of a joke uniquely revealed those who had high and low EQ? If the former, then perceivers should perform well in identifying targets with high and low EQ whether the still photograph was at a point when the target delivered the punch line or at an earlier point in the video when we might suppose the target was less expressive. If the latter, then perceivers should be able to identify cases of high and low EQ on condition that the still picture was at a point when the punch line was delivered but not at any other point. The purpose of Study 2 was to clarify this matter.

Method

Participants

Sixty students (33 males) between 18 years and 25 years old (M = 21 years) were recruited from Monash University Sunway campus. Perceivers were randomly assigned to two groups of 30 to view an array of photographs under one or other condition as explained below. After completing the task, the perceivers were asked whether they had previously seen any of the targets in the pictures and all reported that they had not.

Materials and procedure

Because similar patterns of results emerged from the three scenarios used in Study 1 and because perceiver ratings of target EQ correlated strongly across the three scenarios, we used only the joke scenario in Study 2 for simplicity and in the interest of demonstrating how accurately perceivers perform with minimal information (the Joke videos were much shorter than those for the other two scenarios). From each target, two photographs were extracted in which the target was either at the beginning of reading the joke or at the end of the joke (the punch line). These are labelled the 'first photograph condition' and the 'last photograph condition.' Each photograph was trimmed to standardize the size using the software Windows Movie Maker. To match the endurance of the joke videos shown in Study 1, the photographs in each condition appeared for 9 seconds. All the picture stimuli were displayed in full colour on a laptop using PsychoPy. Generally, the procedure was similar to that in Study 1.

Results and discussion

Preliminary analysis

The perceivers assigned to the first photograph condition had a mean EQ of 38.17 (*SD* = 10.73, ranging from 19 to 54) while those assigned to the last photograph condition had a mean of 35.13 (*SD* = 8.90, ranging from 18 to 54). Independent-samples *t* tests did not reveal any differences between the mean EQs of the targets and the perceivers and neither was there any evidence of difference between the two groups of perceivers in their mean EQ scores. An independent-samples *t* test suggested female perceivers had higher EQ than male perceivers, $t(58) = 3.21$, $p = .002$.

Main analysis – Guessing the EQ of the target

The procedure of coding based on signal detection was the same as that used in Study 1. Table 4 presents the means of the hit rates, false alarm rates and $M_{d'}$ in each category of the four-point scale in each condition, and *t* values of one-sample *t* tests for each $M_{d'}$. As we see in Table 4 and Figure 2, perceivers were above chance in identifying targets with high EQ (Scale 4) in the first photograph condition and above chance in guessing both low (Scale 1) and high EQ (Scale 4) in the last photograph condition.

Table 4 about here

Figure 2 about here

A 2 × 4 mixed design ANOVA was carried out, with the two photograph conditions as the between-subjects factor, and the four EQ scales as the within-subjects factor; the dependent variable was *d'*. There was a main effect associated with the scales, F (3, 174) = 21.69, $p < .001$, Cohen's $f = .85$, an interaction between Scale and Condition, F (3, 174) = 4.18, $p = .007$, Cohen's $f = .37$, but no main effect of condition, F (1, 58) = .67, $p = .418$.

Simple effects analyses for the interaction between Scale and Condition revealed the following. First, a significant effect associated with the four EQ scales was found in each condition: First Photograph Condition: F (3, 87) = 6.39, $p = .001$, Cohen's $f = .46$; Last Photograph Condition: F (3, 87) = 17.87, $p < .001$, Cohen's $f = .77$. Post hoc LSD analyses in the first photograph condition revealed better performance in Scale 4 than in Scale 1 ($p = .045$), Scale 2 ($p = .001$) and Scale 3 ($p = .001$) but there was no difference among Scales 1, 2 and 3. Post hoc LSD tests for the last photograph condition confirmed greater accuracy in Scales 1 and 4 as opposed to Scales 2 and 3 ($ps < .001$), and there was no evidence of difference in $M_{d'}$ between Scales 1 and 4; nor was there any evidence of difference between Scales 2 and 3. Second, as shown in Table 4 and Figure 2, the $M_{d'}$ in Scale 1 was higher in

the last photograph condition than in the first photograph condition, t (58) = 3.26, p = .002; there was no difference between the two conditions in Scales 2, 3 and 4.

In summary, the results in the last photograph condition in Study 2 replicated the same U-shaped pattern found in Study 1. However, the evidence for such a U-shaped pattern was not compelling for the first photograph condition and in particular perceivers did not perform well in guessing which targets had low EQ. It seems therefore that information from the target's delivery of the punch line of the joke was sufficient for perceivers to infer high and low EQ; a photograph where the target was less expressive, merely reading text before he or she reached the punch line, apparently was not sufficient for perceivers to guess low EQ.

Study 3

The previous two studies demonstrated perceivers' ability in inferring target self-rated empathy on the basis of visual behavior cues. Would perceivers also be able to identify who has high EQ and who has low EQ after listening to the target talking for a few seconds? Previous studies have indicated that people can sometimes predict others' daily behaviors after hearing fragments of sound unobtrusively recording their daily lives (Holleran, Mehl, & Levitt, 2009; Mehl, Gosling & Pennebaker, 2006); people can also infer others' emotions while hearing them relating their life experiences (Zaki et al., 2009). If the talking does not relate to an individual's personal life but is merely reading aloud a few lines, as in the case of the Joke Scenario, would perceivers be able to make an accurate judgment of the EQ of the target?

Study 3 thus presented a new condition in which perceivers could only hear the soundtrack of the Joke Scenario. Does being able to perceive EQ depend on having visual access to the target or is auditory evidence sufficient? Some researchers have suggested that

the face, especially the eyes, is the principal source of psychological information (e.g., Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997). If they are right, then we should expect perceivers to be much more accurate in the video than in the auditory conditions that are described below.

Method

Participants

Sixty students (32 males) aged 18 to 23 years (M = 20 years) were recruited from the University of Nottingham Malaysia campus. These perceivers were shown photographs of the targets and asked if they had seen any of them before and only those who said they had not were included in the sample. Perceivers were randomly divided into two groups of 30 either to view video clips or listen to targets telling the joke.

Materials and procedure

The 47 video clips in the Joke Scenario were used as the set of visual stimuli. The auditory stimuli were extracted from the same video clips, thus yielding 47 samples of audio stimuli. Thirty perceivers viewed 47 video clips without sound (Video Only Condition) and another 30 heard 47 audio tracks (Audio Only Condition). The video stimuli were displayed in the size of 800 × 650 pixels on a laptop with PsychoPy, and the audio stimuli were also presented on the laptop using PsychoPy. The procedure was similar to that in the previous studies.

Results and discussion

Preliminary analysis

The perceivers' average EQ was 38.63 in the Video Only Condition (*SD* = 8.30, ranging from 22 to 54) and 36.60 in the Audio Only Condition (*SD* = 9.11, ranging from 22 to 54). Independent-samples *t* tests did not show any differences between the average EQs of the targets and the perceivers and neither was there any evidence of difference between the two groups of perceivers in their average EQ scores. An independent-samples *t* test suggested female perceivers had higher EQ than male perceivers, t (58) = 3.61, p = .001.

Main analysis – Guessing the EQ of the target

Table 5 displays the means of the hit rates, false alarm rates and $M_{d'}$ in each category of the four-point scale in each condition, and *t* values of one-sample *t* tests for each $M_{d'}$. Table 5 suggests that perceivers made systematically correct judgments in the case of low (Scale 1) and high (Scale 4) EQ in the Video Only Condition, while in the Audio Only Condition perceivers systemically estimated the EQ of targets only in the case of high EQ (Scale 4). There was no evidence of perceivers effectively estimating the EQ of targets who were in the middle categories (Scales 2 and 3). As shown in Figure 3, only $M_{d'}$ values in the Video Only Condition yielded a distinct U-shaped curve.

Table 5 about here

Figure 3 about here

To examine whether the perceivers performed differently across scales and conditions in judging EQ, a 2 × 4 mixed design ANOVA was conducted, with the two conditions (Video – Audio) as the between-subjects factor, and the four EQ scales as the within-subjects factor; the dependent variable was *d'*. There was a main effect associated with the scales, F (3, 174) = 35.52, $p < .001$, Cohen's f = 1.08, a main effect of Condition, F (1, 58) = 9.25, p = .004,

Cohen's f = .56, and an interaction between Scale and Condition, F (3, 174) = 6.96, p = .004 Cohen's f = .48.

Simple-effects analyses for the interaction between Scale and Condition revealed two things. First, as with the previous studies, there were significant differences among the four scales and this trend was evident for each condition: Video Only Condition, F (3, 87) = 26.51, p < .001, Cohen's f = .94; Audio Only Condition, F (3, 87) = 15.10, p <. 001, Cohen's f = .71. Post hoc LSD analyses confirmed higher $M_{d'}$ values in Scales 1 and 4 compared with Scales 2 and 3 in the Video Only Condition (*ps* < .001) but higher $M_{d'}$ only in Scale 4 in the Audio Only Condition (*ps* < .001). Second, as demonstrated in Table 5 and Figure 3, the $M_{d'}$ value in Scale 1 was much higher in the Video Only Condition than in the Audio Only Condition, and an independent- samples t test offered confirmation, t (58) = 5.85, p < .001; there was no difference in Scale 4 between the Video and Audio Conditions.

To summarize, Study 3 provided new evidence of perceivers' ability to make judgments of self-rated empathy: Perceivers also systematically identified targets with high EQ (but not with low or middle EQ) on merely listening to the target's voice for about 9 seconds as he or she read aloud a joke. Evidently, perceivers stood a better chance of estimating EQ in the Video Only Condition, especially when the target's EQ was low.

How could perceivers form accurate first impressions of self-rated empathy even if they only heard a soundtrack spanning less than ten seconds? In the Audio Only Condition, the content of the verbal information was merely several lines of a joke. Hence, there was no possibility for perceivers to obtain information about the life of the targets in making judgments of EQ; instead, the only available information was the target's voice characteristics like tone, pitch, as well as mannerisms, such as laughing. Nevertheless, on hearing the soundtrack, perceivers performed as well as in the Video Only Condition when

judging targets with high EQ. These results suggest that visual information is not the only channel perceivers utilize in making psychological inferences, and auditory cues can also play a key role in forming an accurate first impression of self-rated empathy, especially in the case of high empathy.

General Discussion

The three studies collectively demonstrate that perceivers can detect self-rated empathy after observing targets for a few seconds in a video, in a photograph or after hearing their voice telling a brief joke. Previous studies on mentalising have shown that people are capable of identifying empathic *states* of another person based on a brief sample of behavior, for example, making inferences of the contents of thoughts, feelings and emotions that another person experienced (e.g., Ickes, et al., 1990; Ickes, 2003; Hall & Mast, 2007; Zaki et al., 2008; 2009) and guessing that a target was listening to another person relating an empathic story (Pillai, et al., 2012). The present study has extended these findings by showing people's success in inferring empathy as measured by the EQ questionnaire.

Previous studies have indicated that an empathic disposition could be more or less revealed in empathy-related behavior, such as facial expressions, bodily movements and vocalizations (Zhou, et al., 2003; Stueber, 2013). The current research suggests that an observer can interpret these cues (e.g., a smiling face after delivering the punch line of a joke) to determine a target's self-rated empathy, especially targets who had high or low levels of empathy; perceivers were not effective in identifying targets who were average. Although perceivers often judged that targets were in the middle range of self-rated empathy, such judgments were frequently incorrect, suggesting that participants were not effective in discerning who is and who is not 'average' in their level of empathy. This could either be

because they have a poor concept of average empathy or it could be that despite having a good concept they are nevertheless ineffective in recognizing average empathy.

The data thus formed a striking U-shaped trend in which participants were successful in identifying those at the extremes of the empathy continuum but not effective in detecting those in the middle. This pattern emerged consistently and robustly across different samples of target behavior (telling a joke, having a conversation or doing a screen test) and across different forms of presentation (seeing a video or seeing a photograph of the target's behavior).

Compared with the average, targets who had extreme EQ might emit clear and more noticeable indicators of behavior, such as a smiling face, peculiar bodily movements or unique vocal characteristics. This is consistent with the finding that participants were not as effective in guessing empathy from a still photo of the target at the beginning of telling a joke as they were with a photo that captures the *behavior* of the target – at least for targets with low self-rated empathy. Interestingly, though, high levels of empathy might be evident even in a target's neutral pose, as suggested by the results of Study 2, and in this respect our findings add to those reported by Kramer and colleagues (Kramer & Ward, 2010; Jones et al., 2012). While hearing the target is not necessary to identify high and low self-rated empathy, it is nevertheless legitimate to enquire whether hearing the target is *sufficient* to identify high and low empathy. The results of Study 3 create a distinctive pattern in showing that hearing the target's voice is sufficient for recognizing those with high self-rated empathy but not for identifying those with low self-rated empathy. Perhaps the signs of low empathy are more accessible in the visible aspects of the target's behavior than in the target's speech.

Even if those with low and high self-rated empathy did not give any more clues to observers about their empathic status, compared with those with average self-rated empathy,

perceivers might nevertheless be especially well adapted to detecting high and low empathy. Researchers have argued that a capacity for empathy is associated with moral development (e.g., Hoffman, 2000); moreover, a capacity for empathy predicts people's prosocial behaviors, such as altruism (Hoffman, 1984; Eisenberg & Fabes, 1990; Batson, 1991), helping (Batson, O'Quin, Fultz, Vanderplas, & Isen, 1983) and cooperation (Rumble et al., 2010). Presumably, it is useful to know not only who is and who is not empathic but also to know about the morality of people we encounter. Taking these factors into consideration, perhaps it is plausible to suppose that it is very adaptive to recognize swiftly and efficiently those who are either strong or weak in empathizing.

Most research into mindreading has focused on mental states associated with situations and behavioral norms rather than on inferring personality traits (e.g., Andrews, 2008). In real life presumably people perform a synthesis of what they expect as normal behavior in a given situation and what they expect from an individual with an extreme trait. People who are average or normal in their traits behave according to situational norms – this is what it means to be normal. Those who are at the extremes of trait continua do not behave according to situational norms and their behavior can only be predicted or explained if one takes into consideration their traits. As far as predicting or explaining behavior is concerned, then, it is especially important to be sensitive to targets who have extreme traits; targets with normal traits behave according to situational norms such that traits do not require much consideration. This could explain why perceivers in our studies were relatively accurate in detecting targets with high or low EQ (but not those with average EQ).

In short, situational norms and the effects of traits probably interact in giving rise to behavior (e.g., Shinner, 2009; Funder, 2006), and accurate mentalising will only be possible

if the participant can process such an interaction. A priority for future research will be to investigate participants' ability to process this interaction when mindreading.

References

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgments at zero acquaintance. *Journal of Personality and Social Psychology, 55,* 387-395, doi: 10.1037/0022-3514.55.3.387.

Ambady, N., Bernieri, F., & Richeson, J. (2000). Towards a histology of social behavior: Judgmental accuracy from thin slices of behavior. In Zanna, M. P. (Ed.), *Advances in experimental social psychology* (pp. 201–272). New York, NY: Academic Press.

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

Apperly, I. A., Simpson, A., Riggs, K. J., Samson, D., & Chiavarino, C. (2006). Is belief reasoning automatic? *Psychological Science, 17,* 841-844, doi: 10.1111/j.1467-9280.2006.01791.x.

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.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: evidence from very high functioning adults with autism or Asperger Syndrome. *Journal of Child Psychology and Psychiatry, 38,* 813-822, doi: 10.1111/j.1469-7610.1997.tb01599.x.

Batson, C. D. (1991). *The altruism question: Toward a social psychological answer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Batson, C. D., O'Quin, K., Fultz, J., Vanderplas, M., & Isen, A. M. (1983). Influence of self-reported distress and empathy on egoistic versus altruistic motivation to help. *Journal of Personality and Social Psychology*, *45*, 706-718, doi: 10.1037/0022-3514.45.3.706.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65*, 546-553, doi: 10.1037/0022-3514.65.3.546.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology, 86*, 599-614, doi: 10.1037/0022-3514.86.4.599.

Carney, D. R., Colvin, C. R., & Hall, J. A. (2007). A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, *41*, 1054-1072, doi: 10.1016/j.jrp.2007.01.004.

Cassidy, S., Ropar, D., Mitchell, P. & Chapman, P. (2013). Can adults with autism spectrum disorders infer what happened t 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.* In press, doi: 10.1002/aur.1468.

Clark, S. E. (2012). Costs and benefits of eyewitness identification reform: Psychological science and public policy. *Perspectives on Psychological Science, 7,* 238–259, doi: 10.1177/1745691612439584.

Conan-Doyle, A. (1902). *The Hound of the Baskervilles*. London: George Newnes.

Dennett, D. C. (1978). "Beliefs about Beliefs" (commentary on Premack, et al.). *Behavioral and Brain Sciences, 1,* 568-570, doi: 10.1017/S0140525X00076664.

Eisenberg, N., & Fabes, R. A. (1990). Empathy: Conceptualization, measurement, and relation to prosocial behavior. *Motivation and Emotion*, *14*, 131-149, doi: 10.1007/BF00991640.

Funder, D. C. (1991). Global traits: A Neo-Apportian approach to personality. *Psychological Science, 2 (1),* 31-39.

Funder, D. C., Kolar, D. C., & Blackman, M. C. (1995). Agreement among judges of personality: Interpersonal relations, similarity, and acquaintanceship. *Journal of Personality and Social Psychology, 69*, 656–672, doi: 10.1037/0022- 3514.69.4.656.

Funder, D. C. (2006). Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality, 40,* 21-34, doi: 10.1016/j.jrp.2005.08.003.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21,* 177-182, doi: 10.1177/0963721412445309.

Hall, J. A., & Mast, M. S. (2007). Sources of accuracy in the empathic accuracy paradigm. *Emotion, 7,* 438-446, doi: 10.1037/1528-3542.7.2.438.

Hoffman, M. (2000). *Empathy and Moral Development*, Cambridge: Cambridge University Press.

Hoffman, M. L (1984). Interaction of affect and cognition in empathy. In C. E. Izard, J.

Holleran, S. E., Mehl, M. R., & Levitt, S. (2009). Eavesdropping on social life: The accuracy of stranger ratings of daily behavior from thin slices of natural conversations. *Journal of Research in Personality, 43,* 660-672, doi: 10.1016/j.jrp.2009.03.017.

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). Naturalistic social cognition: Empathic accuracy in mixed-sex dyads. *Journal of Personality and Social Psychology, 59 (4),* 730-742, doi: 10.1037/0022-3514.59.4.730.

Ickes, W. (2003). *Everyday mind reading: Understanding what other people think and feel*, Amherst, NY: Prometheus Books.

Jones, E. E, & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. *Advances in Experimental Social Psychology, 2,* 219-266.

Jones, A. L., Kramer, R. S. S., & Ward, R. (2012). Signals of personality and health: The contributions of facial shape, skin texture, and viewing angle. *Journal of Experimental Psychology: Human Perception and Performance, 38,* 1353-1361, doi: 10.1037/a0027078.

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*, 25-41, doi: 10.1016/S0010-0277(03)00064-7.

Kleiman, S. & Rule, N. O. (2013). Detecting suicidality from facial appearance. *Social Psychological and Personality Science, 4,* 453-460, doi: 10.1177/1948550612466115.

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and

health. *Quarterly Journal of Experimental Psychology, 63,* 2273-2287, doi: 10.1080/17470211003770912.

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., & David, A. S. (2004). Measuring empathy: reliability and validity of the empathy quotient. *Psychological Medicine, 34*, 911-924, doi: 10.1017/S0033291703001624.

Lynn, S. K., & Barrett, L. F. (2014). "Utilizing" signal detection theory. *Psychological Science*, 25, 1669-1673, doi: 10.1177/0956797614541991.

Macmillan, N. A. (2002). Signal detection theory. In Pashler, H. (Ed.) *Stevens' handbook of Experimental Psychology* (3nd ed.). In Wixted, J. (Ed.), *Vol. 4: Methodology in experimental psychology* (pp. 43-90). John Wiley & Sons, Inc.

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

McCarthy, R. J., & Skowronski, J. J. (2011). What will Phil do next? Spontaneously inferred traits influence predictions of behavior. *Journal of Experimental Social Psychology, 47*, 321-332, doi: 10.1016/j.jesp.2010.10.015.

McLarney-Vesotski, A. R., Bernieri, F., Rempala, D. (2006). Personality perception: A developmental study. *Journal of Research in Personality, 40*, 652–674, doi: 10.1016/j.jrp.2005.07.001.

Mehl, M. R., Gosling, S. D. & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*, 862-877, doi: 10.1037/0022-3514.90.5.862.

Mitchell, P., Currie, G., & Zeigler, F. (2009). Two routes to perspective: Simulation and rule-use as approaches to mentalizing. *British Journal of Developmental Psychology, 27*, 513-543, doi: 10.1348/026151008X334737.

Mitchell, P., Robinson, E. J., Isaacs, J., & Nye, R. (1996). Contamination in reasoning about false belief: An instance of realist bias in adults but not children. *Cognition*, *59*, 1-21, doi: 10.1016/0010-0277(95)00683-4.

Norman, W. T., & Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology*, *4*, 681-691, doi: 10.1037/h0024002.

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8-13, doi: 10.1016/j.jneumeth.2006.11.017.

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., Sheppard, E., Ropar, D., Marsh, L., Pearson, A., & Mitchell, P. (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.

Rumble, A. C., Van Lange, P. A. M., & Parks, C. D. (2010). The benefits of empathy: When empathy may sustain cooperation in social dilemmas. *European Journal of Social Psychology, 40*, 856-866, doi: 10.1002/ejsp.659.

Shiner, R. L. (2009). Persons and situations act together to shape the course of human lives. *Journal of Research in Personality, 43*, 270-271, doi: 10.1016/j.jrp.2008.12.023.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137-149, doi: 10.3758/BF03207704.

Stueber, K., "Empathy", *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/sum2013/entries/empathy/>.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, 1-26.

Todorov, A., & Uleman, J. S. (2004). The person reference process in spontaneous trait inferences. *Journal of Personality and Social Psychology, 87,* 482-493, doi: 10.1037/0022-3514.87.4.482.

Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition, 27,* 813-833, doi: 10.1521/soco.2009.27.6.813.

Thoresen, J. C., Vuong, Q. C., & Atkinson, A. P. (2012). First impressions: Gait cues drive reliable trait judgments. *Cognition, 124*, 261-271, doi: 10.1016/j.cognition.2012.05.018.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103-128, doi: 10.1016/0010-0277(83)90004-5.

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*, 237-252, doi: 10.1037/0022-3514.47.2.237.

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

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.

Zhou, Q., Valiente, C., & Eisenberg, N. (2003). Empathy and Its Measurement. In S. J. Lopez, S. J. & Snyder, C. R. (Ed.), *Positive Psychological Assessment: A Handbook of Models and Measures* (pp. 269–284). Washington, DC: American Psychological Association.

Footnote[1]

We repeated all of the analyses after coding the dependent variable differently, based on the percentage that the perceiver judged a particular EQ category correctly compared with judgments that the category was used incorrectly. We call this system of coding 'percentage correct.' This is calculated as follows in relation to a participant's score for any given EQ scale: (Number of hits + number of correct rejections) ÷ (number of targets + number of distracters). This system of coding, like signal detection theory, yields a single value to represent the perceivers' performance across multiple trials in their ability to identify targets that belong to a particular EQ category whilst also taking into account any bias (for example, bias to judge that targets are located in the middle EQ scales). Signal detection theory does exactly the same in principle except that it yields normalized values such that systematic performance corrected for bias is indicated by values that are significantly greater than zero. In short, signal detection theory is an appropriate form of coding but even if there were any lingering doubt then at least we can offer the reassurance that the same pattern of data emerges when using a different coding procedure.

Figure captions

Figure 1. Mean *d'* of perceivers' estimates of the empathy quotient (Scales 1-4) of targets performing under three conditions in Study 1: Targets either participated in conversation, told a scripted joke or did a 'screen test' in which they read aloud a paragraph of promotional material. Error bars represent standard error of the mean.

Figure 2. Mean *d'* of perceivers' estimates of the empathy quotient (Scales 1-4) after viewing photographs of targets either with a fairly neutral pose or as they delivered the punch line of a joke (First and Last Photograph Conditions) in Study 2. Error bars represent standard error of the mean.

Figure 3. Mean *d'* of perceivers' estimates of the empathy quotient (Scales 1-4) after watching videos (without audio) of targets telling a joke or after listening to them (but not seeing them) telling a joke (Video and Audio Conditions) in Study 3. Error bars represent standard error of the mean.

Table 1. Frequencies of perceiver judgments of target EQ (Scales 1-4) in Study 1

Perceiver Response (EQ scale)	True Target EQ (Scales 1-4) in Each Scenario														
	Conversation					Joke					Screen Test				
	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4	Total
1	41	61	36	1	139	71	77	63	6	217	49	56	51	10	166
2	146	200	106	17	469	155	193	122	34	504	145	208	97	39	489
3	127	241	139	62	569	115	247	121	52	535	135	225	115	46	521
4	46	98	49	40	233	19	83	24	28	154	31	111	67	25	234
Total	360	600	330	120		360	600	330	120		360	600	330	120	

Table 2. Frequencies of perceiver judgments of target EQ (Scales 1-4) in Study 2 and Study 3

Perceiver Response (EQ scale)	Study 2 True Target EQ (Scales 1-4) in Each Condition									
	First Photograph					Last Photograph				
	1	2	3	4	Total	1	2	3	4	Total
1	45	68	48	18	179	95	74	61	17	247
2	151	235	127	43	556	151	191	122	38	502
3	126	215	118	49	508	92	242	112	44	490
4	38	82	37	10	167	22	93	35	21	171
Total	360	600	330	120		360	600	330	120	

Perceiver Response (EQ scale)	Study 3 True Target EQ (Scales 1-4) in Each Condition									
	Video Only					Audio Only				
	1	2	3	4	Total	1	2	3	4	Total
1	115	75	76	9	275	70	125	77	15	287
2	141	198	120	35	494	135	202	112	35	484
3	83	231	105	48	467	104	178	99	38	419
4	21	96	29	28	174	51	95	42	32	220
Total	360	600	330	120		360	600	330	120	

Table 3. Means (and standard deviations) of hit rates (M_{HR}), false alarm rates (M_{FAR}) and d-prime ($M_{d'}$) in each EQ scale, along with values of one-sample *t* tests of each $M_{d'}$ in Study 1

	Conversation				Joke				Screen Test			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
M_{HR}	.13	.33	.42	.36	.21	.32	.37	.29	.15	.35	.34	.26
	(.13)	(.15)	(.20)	(.27)	(.20)	(.17)	(.21)	(.22)	(.14)	(.13)	(.13)	(.18)
M_{FAR}	.10	.33	.40	.15	.14	.38	.38	.10	.11	.35	.37	.17
	(.10)	(.13)	(.15)	(.14)	(.14)	(.13)	(.16)	(.09)	(.10)	(.11)	(.13)	(.11)
$M_{d'}$	.26	.01	.06	.85	.31	-.21	-.10	.82	.22	-.01	-.11	.38
	(.36)	(.43)	(.51)	(.62)	(.44)	(.37)	(.51)	(.70)	(.38)	(.30)	(.49)	(.59)
t	3.90**	.07	.60	7.49**	3.87**	-3.10*	-.106	6.34**	3.13*	-.10	-1.24	3.60**

Note: S1, S2, S3, S4 = Scale 1, Scale 2, Scale 3, and Scale 4. Three groups (n = 30 in each) of perceivers viewed targets in one of three scenarios (Conversation, Joke, Screen Test). *.$p < .01$, **. $p \leq .001$; two-tailed.

Table 4. Means (and standard deviations) of hit rates (M_{HR}), false alarm rates (M_{FAR}) and d-prime ($M_{d'}$) in each EQ scale, along with values of one-sample *t* tests of each $M_{d'}$ in Study 2

	First Photograph Condition				Last Photograph Condition			
	Scale 1	Scale 2	Scale 3	Scale 4	Scale 1	Scale 2	Scale 3	Scale 4
M_{HR}	.14	.39	.36	.18	.27	.32	.33	.23
	(.13)	(.16)	(.15)	(.08)	(.23)	(.15)	(.18)	(.14)
M_{FAR}	.13	.40	.36	.12	.15	.38	.35	.13
	(.10)	(.11)	(.12)	(.08)	(.11)	(.12)	(.12)	(.09)
$M_{d'}$	.09	-.05	-.03	.33	.43	-.19	-.11	.42
	(.36)	(.40)	(.39)	(.49)	(.44)	(.42)	(.51)	(.47)
t	1.41	-.74	-.41	3.63**	5.31**	-2.51*	-1.14	4.95**

Note: Two groups (n=30 in each group) of perceivers each viewed targets in one of two conditions (First Photograph Condition & Last Photograph Condition); *p = .01, **p < .001; two-tailed.

Table 5. Means (and standard deviations) of hit rates (M_{HR}), false alarm rates (M_{FAR}) and d-prime ($M_{d'}$) in each EQ scale, along with values of one-sample *t* tests of each $M_{d'}$ in Study 3

	Video Only				Audio Only			
	Scale 1	Scale 2	Scale 3	Scale 4	Scale 1	Scale 2	Scale 3	Scale 4
M_{HR}	.33	.32	.31	.28	.21	.34	.29	.31
	(.25)	(.15)	(.19)	(.17)	(.14)	(.13)	(.16)	(.22)
M_{FAR}	.16	.37	.33	.12	.21	.35	.29	.15
	(.13)	(.10)	(.12)	(.10)	(.13)	(.13)	(.09)	(.11)
$M_{d'}$	.59	-.18	-.13	.72	-.04	.08	.09	.58
	(.44)	(.38)	(.51)	(.51)	(.39)	(.36)	(.50)	(.56)
t	7.29**	-2.51*	-.1.38	7.69**	-.57	-1.28	-1.05	5.70**

Note: Two groups (n = 30 in each group) of perceivers each viewed targets in one of two conditions (Video, Audio); *. p < .05, **. p < .001, two-tailed.

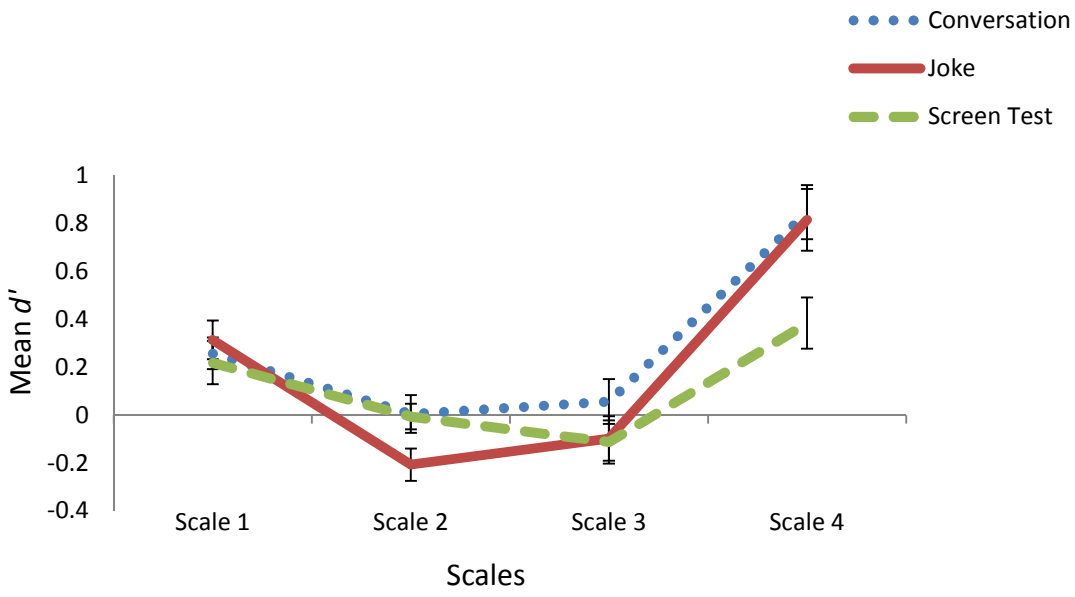
Conversation
Joke
Screen Test
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Mean d'
Scale 1
Scale 2
Scale 3
Scale 4
Scales

Figure 1

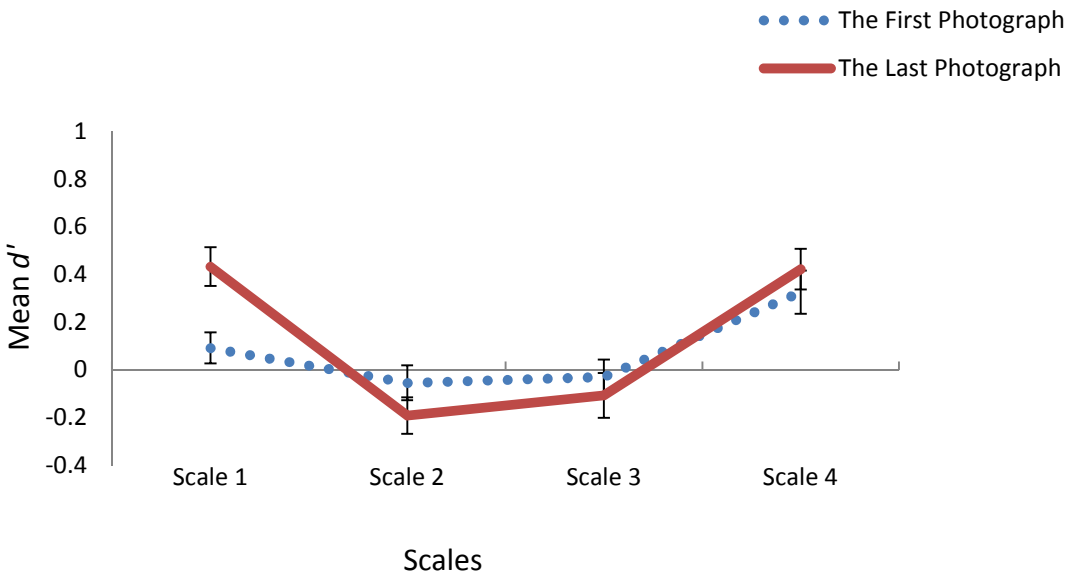
The First Photograph
The Last Photograph
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Mean d'
Scale 1
Scale 2
Scale 3
Scale 4
Scales

Figure 2

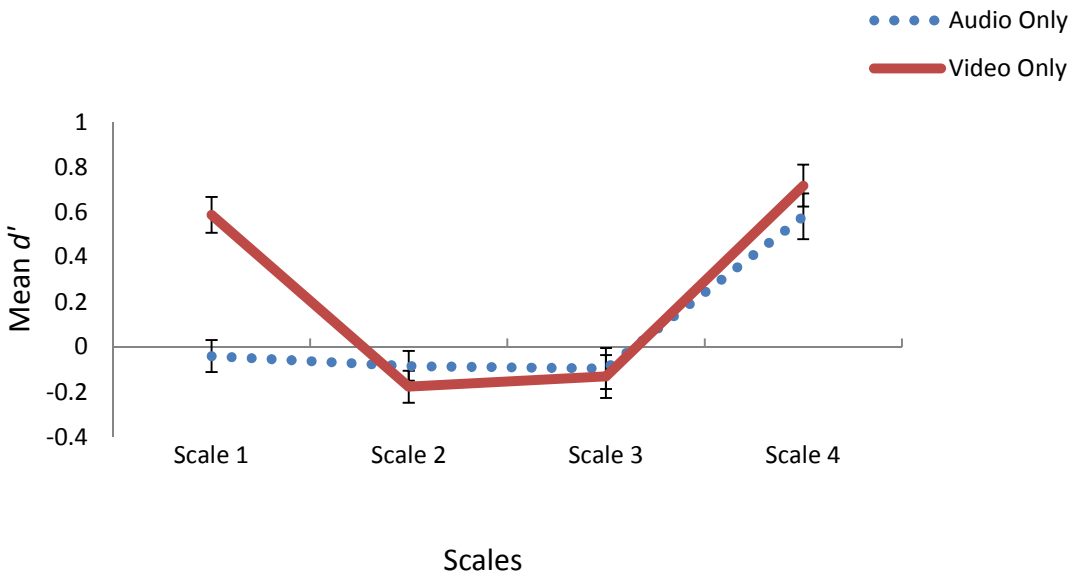
Audio Only
Video Only
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Mean d'
Scale 1
Scale 2
Scale 3
Scale 4
Scales

Figure 3