Adults' facial impressions of children's niceness, but not shyness, show modest accuracy

Jemma R. Collova[1], Clare A.M. Sutherland[1,2], Linda Jeffery[1], Ellen Bothe[1], & Gillian Rhodes[1]

[1] School of Psychological Science, The University of Western Australia, Crawley, Australia
[2] School of Psychology, University of Aberdeen, Scotland
Corresponding author: Jemma Collova, jemma.collova@uwa.edu.au
ORCID id: https://orcid.org/0000-0003-4300-2543

Collova, J. R., Sutherland, C. A. M., Jeffery, L., Bothe, E., & Rhodes, G. (in press). Adults' facial impressions of children's niceness, but not shyness, show modest accuracy. *Quarterly Journal of Experimental Psychology.*

Note: This is a copy of the Accepted Manuscript. This manuscript was accepted for publication in *QJEP*, 26 June 2020. This paper is not the copy of record and may not exactly replicate the final version.

Acknowledgements

We are grateful to the parents and children who helped make this research possible. We would also like to thank Romina Palermo for providing us the opportunity to contact her sample of parent and child participants, and to use some of her existing data. Finally, we would like to thank the examiners who provided thoughtful comments on an earlier draft of this paper presented in a thesis.

JC, CS, LJ and GR conceived the study and edited the manuscript. JC programmed the experiment, collected undergraduate participant data, performed the statistical analyses, and drafted the first manuscript. EB coordinated image collection. All authors participated in the study design, and read and approved the final manuscript.

Funding: This research was supported by an Australian Research Council (ARC) Centre of Excellence Grant award to GR [CE110001021], ARC Discovery Early Career Research Award to CS [DE190101043], ARC Discovery Award to GR and CS [DP170104602], ARC Discovery Award to LJ [DP140101743] and a Research Training Program Stipend to JC.

Abstract

Lay wisdom warns against "judging a book by its cover". However, facial first impressions influence people's behaviour towards others, so it is critical that we understand whether these impressions are at all accurate. Understanding impressions of children's faces is particularly important because these impressions can have social consequences during a crucial time of development. Here, we examined the accuracy of two traits that capture the most variance in impressions of children's faces, niceness and shyness. We collected face images and parental reports of actual niceness/shyness for 86 children (4-11 years old). Different images of the same person can lead to different impressions, and so we employed a novel approach by obtaining impressions from five images of each child. These images were ambient, representing the natural variability in faces. Adult strangers rated the faces for niceness (Study 1) or shyness (Study 2). Niceness impressions were modestly accurate for different images of the same child, regardless of whether these images were presented individually or simultaneously as a group. Shyness impressions were not accurate, either for images presented individually or as a group. Together, these results demonstrate modest accuracy in adults' impressions of niceness, but not shyness, from children's faces. Furthermore, our results reveal that this accuracy can be captured by images which contain natural face variability, and holds across different images of the same child's face. These results invite future research into the cues and causal mechanisms underlying this link between facial impressions of niceness and nice behaviour in children.

Adults' facial impressions of children's niceness, but not shyness, show modest accuracy

People spontaneously infer personality characteristics from a glimpse of a face (for a review see Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). These impressions have serious social consequences across the lifespan. For adults, impressions influence political success (Antonakis & Dalgas, 2009; Ballew & Todorov, 2007) dating preferences (Hinsz, 1989) hiring of business leaders (Antonakis & Eubanks, 2017; Graham, Harvey, & Puri, 2016; Olivola, Eubanks, & Lovelace, 2014) and sentencing decisions within the criminal justice system (Blair, Judd, & Chapleau, 2004; Wilson & Rule, 2015). Arguably, the social consequences of first impressions are particularly important to consider during childhood, as this period represents a crucial time of development. Indeed, for children, impressions influence school achievement (Salvia, Algozzine, & Sheare, 1977), popularity (for a review Langlois et al., 2000), the severity of punishment received (Berkowitz & Frodi, 1979; Zebrowitz, Kendall-Tackett, & Fafel, 1991) and teacher expectations (Clifford & Walster, 1973; Kenealy, Frude, & Shaw, 1988), and these effects can have long lasting behavioural consequences (e.g. the self-fulfilling prophecy effect: Zebrowitz, Voinescu, & Collins, 1996). The tendency for impressions to have serious social consequences is well established. However, a fundamental question about these impressions remains widely debated. That is, do impressions predict behaviour or personality at all?

Are impressions of adults' faces accurate?

To date, almost all studies have focused on the accuracy of impressions of adult faces. Consistent with lay wisdom not to "judge a book by its cover", some studies on impressions of adults have found no accuracy (Efferson & Vogt, 2013; Olivola & Todorov, 2010; Rule, Krendl, Ivcevic, & Ambady, 2013). However, a growing body of literature suggests that some impressions may provide modestly accurate signals to character and behaviour. For example, impressions of generally positive and negative traits (i.e. valence-related traits),

such as trustworthiness (Porter, England, Juodis, Ten Brinke, & Wilson, 2008; Slepian & Ames, 2016), honesty (Bond, Berry, & Omar, 1994; Zebrowitz et al., 1996), sexual unfaithfulness (Foo, Loncarevic, Simmons, Sutherland, & Rhodes, 2019; Leivers, Simmons, & Rhodes, 2015; ; Rhodes, Morley, & Simmons, 2013) and aggressiveness (Carré & McCormick, 2008; Short et al., 2012; Zilioli et al., 2015) show some accuracy (but see Rule et al., 2013).

Although accuracy is usually very modest, it is theoretically important to establish whether impressions are at all accurate, as this finding would suggest a link between appearance and behaviour, at least in adult faces. There are several theories which could account for this link (see Zebrowitz & Collins, 1997). One explanation, is that there are intrinsically valid signals of character in faces. From an evolutionary theoretical perspective, it would be adaptive to be able to accurately detect some of these signals. For example, accurately inferring how faithful someone is may help guide our behaviour towards that person as a potential mate (Rhodes et al., 2013). Another explanation, according to self-fulfilling prophecy accounts, is that long-term effects of being evaluated on the basis of one's facial appearance may encourage behaviours that increase the accuracy of that initial impression. For example, looking honest might encourage others to engage in trusting behaviours with that person, and thus encourage that person to act more honestly (Bond et al., 1994; Zebrowitz et al., 1996). Regardless of the underlying mechanism, these accounts suggest a link between facial appearance and behaviour.

In contrast to the above, other theoretical accounts suggest there is no link between appearance and behaviour. For example, in line with the overgeneralization hypothesis of facial impressions, the tendency to form impressions reflects the by-product of adaptive mechanisms. That is, some facial qualities are so important in guiding adaptive behaviour, that they are overgeneralized to people whose facial appearance resembles those cues (for a

review; Zebrowitz & Montepare, 2008). For example, a face with a neutral expression that slightly resembles a smile, may activate an adaptive mechanism that has evolved to detect facial emotion expressions. As a consequence of this emotion overgeneralization, a face may be perceived as trustworthy irrespective of whether or not that person is actually trustworthy (Oosterhof & Todorov, 2008). In this context, facial signals may or may not be related to behaviour.

Are impressions of children's faces accurate?

The question of whether or not facial impressions are accurate remains widely debated (Bonnefon et al., 2015), and relatively little is known about the accuracy of trait impressions of children's faces as compared to adults' faces. However, some studies do show a modest relationship between a child's facial appearance and their behaviour or character. For example, Zebrowitz, Hall, Murphy, and Rhodes (2002) found modest accuracy for impressions of intelligence from images of faces as young as 10 years old (for similar reviews see Jackson, Hunter, & Hodge, 1995; Langlois et al., 2000). Impressions of psychological adjustment from children's faces also show modest accuracy, with adults able to distinguish socially competent children from their more aggressive or anxious peers (Dumas, Nilsen, & Lynch, 2001; Serketich & Dumas, 1997). It is possible that people may be sensitive to actual facial cues that signal intelligence and psychological adjustment, driven by an evolved mechanism sensitive to the detection of "good" or "bad" genes (see also Zebrowitz & Rhodes, 2004). Alternatively, it is also possible that these links between appearance and character reflect a self-fulfilling prophecy effect. For example, children who are expected to be smarter because of their appearance may be given more opportunities to become smarter (Zebrowitz et al., 2002).

There is also some evidence that adults can form accurate impressions of trustworthiness (Li, Heyman, Mei, & Lee, 2017) and honesty (Zebrowitz et al., 1996) from

children's faces. Adults' impressions of trustworthiness from children's faces predicted actual trustworthiness in those children one year later (Li, et al., 2017). This result suggests there may be valid signals of trustworthiness in children's faces, whether this link arises through a self-fulfilling prophecy or is based on evolution. Of course, these two main theoretical accounts for accuracy are not mutually exclusive. For example, it is possible that people have an evolutionary preparedness to detect valid signals of trustworthiness in faces, and that this accuracy is enhanced by self-fulfilling prophecy effects. So, there are theoretical reasons to expect that impressions of children's faces would show a degree of accuracy, although research in this field is limited.

To date, most research into impressions of children's faces has focused on specific traits predicted to be important, such as traits related to fitness (e.g. intelligence and health) and adaptive behaviour (e.g. trustworthiness). Here, inspired by a recent data-driven approach to the study of trait inferences, we examine the accuracy of two important traits, niceness and shyness, that explain the most variance in adults' impressions of children's faces (Collova, Sutherland, & Rhodes, 2019). The accuracy of niceness and shyness impressions are particularly important to consider, because these traits represent the key dimensions of evaluation of children's faces and signal important information relevant to the social goals that adults associate with children (Collova et al., 2019). These social goals are more focused on adults' caregiving relationship with children, rather than the threat-focused goals associated with adult-adult relationships. Impressions of niceness are influenced by subtle facial expressions of happiness, and may stem from mechanisms designed to detect information about those we should approach and nurture, versus those we should avoid (cf. trustworthiness: Oosterhof & Todorov, 2008). Impressions of shyness are related to subtle facial expressions of fear, and may stem from mechanisms designed to signal information about social competencies or vulnerability (Collova et al., 2019). These key dimensions of

trait perception have only recently been discovered, and so it is not yet known whether adults' impressions of these dimensions are accurate.

Observing even modest accuracy in impressions of children's faces would be theoretically important as this result would suggest that adults are sensitive to valid signals of character in these faces, whether this be due to an evolutionary preparedness, self-fulfilling prophecy effects, or another mechanism (for details on these accounts and how they might be distinguished, see Zebrowitz & Collins, 1997). In contrast, other theoretical accounts suggest that there is no link between facial appearance and behaviour. For example, impressions of niceness/shyness may reflect the by-product of mechanisms that evolved to detect adaptive (but not valid) cues in faces, consistent with an overgeneralization hypothesis (Zebrowitz & Montepare 2008) and so we may not expect any accuracy. Impressions of niceness and shyness influence adults' behavioural expectations of children (Collova et al., 2019), so it is also important that we understand the degree of their accuracy from a practical perspective.

Measuring accuracy using multiple ambient images

Our main aim was to examine the accuracy of impressions of children's faces, as little is known about these impressions. In addition, we also extended previous research by testing accuracy across multiple images of the same child's face instead of a single face image. In typical accuracy studies, participants form impressions from a single standardized image of a stranger's face, and this impression is compared with actual reports of personality or behaviour (e.g. Bond et al., 1994; Carré, McCormick, & Mondloch, 2009; Foo et al., 2019; Haselhuhn & Wong, 2011; Kleisner, Chvátalová, & Flegr, 2014; Li et al., 2017; Stillman, Maner, & Baumeister, 2010; Stirrat & Perrett, 2010; Zebrowitz et al., 2002; Zebrowitz & Rhodes, 2004). Almost nothing is known about the accuracy of impressions across different images containing natural variability in one's appearance. However, impressions can vary substantially across different images of the same individual (Jenkins, White, Van Montfort, &

Burton, 2011; Sutherland, Rhodes, & Young, 2017; Todorov & Porter, 2014). For example, the same person can look trustworthy in one image but untrustworthy in another image, simply because of dynamic changes in expression or viewpoint (Sutherland, Young, & Rhodes, 2017). These impressions can be so variable that different images of the same person can be perceived as entirely different people (Jenkins et al., 2011; Laurence & Mondloch, 2016). To date, this variability has been considered as evidence against accuracy (e.g. Olivola, Funk, & Todorov, 2014). Indeed, because impressions can vary across different images of the same person, it is not certain that accurate impressions from one image would still be accurate, given other images of that same person. However, enduring facial characteristics can be present in multiple photographs of a person (e.g. consistent structural features or common expressions), and so there might still be stable differences between impressions of different people, despite within-identity variability (Todorov & Porter, 2014). If some features are consistently captured, then accuracy should hold across some, if not all images of the same person. Moreover, observing accuracy across different images of the same person could also suggest that accuracy is not image specific.

Finally, we extended previous research by using naturalistic images instead of standardized images. These images were ambient, that is, they were free to vary as faces do in every-day life (see Jenkins et al., 2011; Sutherland et al., 2013). In previous accuracy studies, standardized images have allowed for precise control over potentially confounding cues. However, it is not clear that the accuracy observed for these highly controlled images would generalize to the variable conditions in which we see faces in everyday life, that is, faces that vary in expression, pose, lighting, and so forth. Ambient images may therefore provide a more ecologically valid estimate of impression accuracy than standardized images, although very few accuracy studies have used these stimuli (for exceptions; Back et al., 2010; Reiss & Tsvetkova, 2019; Satchell, Davis, Julle-Danière, Tupper, & Marshman, 2019).

Furthermore, it is important to show that accuracy can generalize to these sorts of naturalistic images because this context reflects people's normal experience with faces. Here, we examined the accuracy in impressions of ambient images, providing an ecologically valid test of accuracy.

Current Studies

Here, in two studies we investigated the accuracy of niceness (Study 1) and shyness (Study 2) impressions from children's faces for the first time. Using naturalistic (ambient) images of a child's face, we also tested whether any accuracy in impressions could hold across different images of that same child's face. To our knowledge, we present the first study to use different ambient images of the same person's face to investigate impression accuracy.

We collected images of children's faces (aged 4-11 years) and reports of actual niceness/shyness from the parents of those same children. In Study 1, adult participants rated the children's faces for how nice they looked. To measure accuracy, we correlated niceness impressions with actual niceness. Actual niceness was indexed by parents' responses to well-established and reliable questions regarding the pro-social and anti-social behaviours of their child (from the Strengths and Difficulties Questionnaire; Goodman, 1997).

In Study 2, new adult participants rated the same children's faces for how shy they looked. We followed the same procedure and analysis as in Study 1, and correlated shyness impressions with actual shyness measures. To index actual shyness, we analysed parents' responses to questions regarding shy behaviours of their child from a well-established questionnaire (Colorado Childhood Temperament Inventory: Buss & Plomin, 1984; Rowe & Plomin, 1977).

We had no strong predictions about whether impression of niceness or shyness from children's faces would be accurate. However, from a theoretical perspective, both an

evolutionary account and a self-fulfilling prophecy account suggest that there would be a link between appearance and behaviour. Furthermore, in light of growing evidence for a degree of accuracy in impressions of adults' faces (e.g. Bond et al., 1994; Foo et al., 2019; Rhodes et al., 2013; Slepian & Ames, 2016; Stirrat & Perrett, 2010) it was possible these impressions would show some degree of accuracy.

STUDY 1: Do impressions of niceness show any accuracy?

In Study 1 we collected ambient face images of children, and parental reports of nice behaviour for those same children (from here on referred to as "actual niceness"). Adult strangers rated the faces for niceness, and these impressions were correlated with actual niceness measures.

Participants rated different images of the same child's face, presented in two contexts. First, the images were distributed individually throughout the task. Seeing distributed individual images reflects instances in everyday life where people form impressions from different glimpses of a person's face, across time, and are not necessarily aware they represent the same person. Second, the different images of the same child's face were presented simultaneously as a group. Seeing grouped images provides a novel and ecologically valid test of accuracy, as it reflects instances in every-day life where people have access to different images of a person and are aware that they represent the same person, such as on Facebook. Seeing different images of a person as a group and knowing they represent the same person, provides a larger sample of face information (e.g. knowledge about common emotional expressions) relative to any individual image. Thus, impressions from grouped images could potentially be more accurate than impressions from any of the distributed images. We also tested this possibility here.

As well as examining accuracy at the participant-group level we also examined

individual-rater level accuracy, inspired by recent approaches (Carré et al., 2009; Foo et al., 2019; Sutherland et al., 2018). Examining individual-rater level accuracy is a stricter test of accuracy than group-level accuracy, which potentially averages across individual error. Furthermore, individual-rater level accuracy provides insight into the extent to which individual participants should rely on their impressions. Finally, we also examined individual-rater *reliability* in impression accuracy by comparing accuracy for participants who repeated the same task (see Methods for more detail). Previous research has found that facial impressions are reliable (e.g. Oosterhof & Todorov, 2008), but to our knowledge, no study has examined whether the accuracy of these impressions is also reliable. Observing reliable accuracy would provide novel insight into whether forming accurate impressions is an enduring ability, and whether individual differences in accuracy can be reliably measured.

Methods

Participants and power. Forty-seven adult participants (37 female, $M = 23.7$, $SD = 9.0$, range = 18 – 52 years, 43 Caucasian) were recruited from the University of Western Australia. Participants completed a two-phase study. In Phase 1, all participants completed the distributed-image task ($N = 47$, see Procedure for task details). In Phase 2, participants either completed the distributed-image task again ($N = 23$), or completed the grouped-images task ($N = 24$: randomly assigned). To ensure reliable face ratings at the participant-group level, we aimed for a sample size of >15 participants per group, which we exceeded. This sample size has been found sufficient for good reliability of similar trait judgments of ambient images at the group level (Collova et al., 2019; Sutherland et al., 2013). Here, we also observed good reliability at the group level (ratings averaged across the five image sets in Phase 1: Cronbach's alpha = .97).

Critically, our analyses were at the face level and therefore, power was determined by the number of face identities in our experiment. We aimed for a sample size comparable to

previous similar research (Li et al., 2017: $N = 100$ children), although sample size was influenced by the number of parents willing to send in photographs. In total, we obtained images of 86 child identities (N = 430 images in total, see below for more details).

Materials.

Child face photographs. For the purpose of this study, we obtained child face photographs from a convenient sample of participants (children who were previous participants in our lab, and for whom we already had reliable measures of actual niceness and shyness). The final stimuli set consisted of five images for each of the 86 children (430 total images; 48 female, $M_{age} = 8$ years, range = 4 -11 years, 70 Caucasian). Photographs were obtained from the parents of the children. To minimise bias in photo selection, parents were naïve to the purpose of the experiment (i.e. they were not informed that children's faces would be rated on niceness and shyness). Therefore, parents did not have a particular goal in mind when choosing the face images (although other biases may have influenced their image selection, see the General Discussion). Furthermore, parents had answered the questions regarding niceness months in advance of sending the photographs, as part of a lengthy questionnaire battery. Thus, image selection was almost certainly not influenced by the specific questions relevant here.

Parents were asked to send us five photographs of their child's face, taken within the past year[1]. They were encouraged to send the five most recent photographs where their child's face was clearly visible. There were no restrictions regarding variability of expression, hair style, etc., as long as the details of the child's face were clearly visible in each image (see Figure 1 for example images). Therefore, image characteristics (e.g. lighting, angle) and facial characteristics (e.g. expression, hair style), did vary considerably across the

[1] Note, for some children (N = 4) we had questionnaire data from more than a year ago. For these cases, we requested child face photographs from the same year that the questionnaire was completed.

face photographs. If parents sent more than five photographs, we selected the highest quality images, determined by image resolution.

We only included images which (i) exceeded 100KB in resolution, (ii) showed the face from mostly frontal aspect, and (iii) were free from occlusions to the face (comparable to criteria set by Jenkins et al., 2011). Three children did not have five images that met these criteria, and therefore we excluded these identities from the experiment. Five additional ambient images of one female and male (total $N = 10$) child were included as practice stimuli.

Stimuli were cropped with an oval mask around the face to avoid any potential influence of extra-facial information (e.g. clothing; Oh, Shafir, & Todorov, 2019) on impression formation. Images were rotated so that faces were upright. Images were re-sized to a standardized size of 180pixels wide (resolution = 72KB; approx. 4cm on screen), but were otherwise left unmodified.

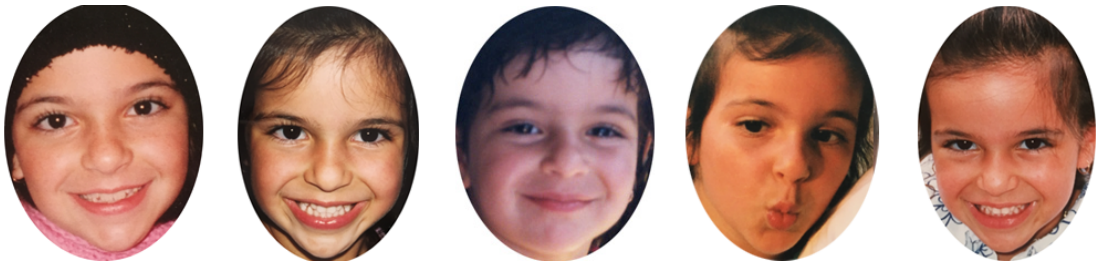

Figure 1. An example of five ambient images of one face identity. This identity was the practice stimuli in our experiment, and is representative of the test stimuli we used.

Trait measure: Niceness. To estimate actual niceness, we used parents' responses on the well-established Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997). The SDQ is a brief questionnaire that is comprised of five scales, each with five items ($N = 25$ items in total). To measure niceness, we averaged together scores from the two scales that were most relevant to the niceness dimension; the prosocial behaviour and conduct problems

scales (we never analysed results from the other scales). The prosocial scale was comprised of five statements: for example, “my child is considerate of other people’s feelings” and “my child is kind to younger children”. Higher scores reflected greater prosocial behaviours. The conduct problems scale included items such as, “my child often fights with other children or bullies them,” and “my child often lies or cheats.” Here, higher scores were indicative of greater conduct difficulties. Parents were asked to reflect on their child’s behaviour over the last six months, and responded to each statement on a three-point scale (0 = not true, 1 = somewhat true, 2 = certainly true). To create an overall “actual niceness” measure, we reverse-scored the conduct problems scale, and then averaged the two scale scores together. Therefore, scores could range from 0-10, with higher scores signalling greater niceness. The SDQ has good validity and reliability (Hawes & Dadds, 2004; Seward, Bayliss, Stallman, & Ohan, 2018; Stone, Otten, Engels, Vermulst, & Janssens, 2010), and also shows high inter-rater reliability between parent and teacher reports (Stone et al., 2010). Here, we found adequate reliability for the two scales we used; prosocial scale: $\alpha = .81$; conduct problems scale: $\alpha = .65$, and the combined scales: $\alpha = .78$.

SDQ data were already collected prior to this experiment for 73 children, as part of an unrelated study. In that study parents filled out a battery of questionnaires (which included the SDQ) and answered some demographic questions (e.g. age, sex), about their child. Due to the sheer number of questions, it was unlikely that parents knew we were specifically interested in niceness and shyness measures. Furthermore, these questionnaires were completed months prior to the request to collect photographs. Prior to the collection of face trait impressions, 13 additional children were recruited (via email and online advertising) for the purpose of this study. For these children, parents first sent in five photographs of their child and then answered the SDQ online. We purposefully chose this order to minimize bias in parental selection of photographs.

Stimuli and Procedure. Participants completed a two-phase study. In the first phase, we were interested in testing whether impressions from different images of the same child, distributed throughout the task, were accurate. Here, all participants ($N = 47$) completed the distributed-image task. In the second phase, we were primarily interested in understanding whether impressions from different images of the same person presented simultaneously (i.e. from the grouped-images task) were accurate. We were also interested in comparing accuracy between impressions of the distributed images and grouped images, therefore, in the second phase participants either completed (randomly assigned) the grouped-images task ($N = 24$), or the distributed-image task again ($N = 23$) as a control for repetition (see Figure 2 for design). This design also allowed us to examine the reliability of any impression accuracy for participants who repeated the distributed-image task.

Distributed-image task: Participants rated the individual images of children's faces for how nice they looked (Figure 2). For each image participants were asked to rate how nice the child was, on a scale of 1 (not at all nice) to 9 (extremely nice), following Collova et al (2019). Participants were encouraged not to spend too long on each image, and to go with their spontaneous first impression. The five images were randomly distributed across five different image set blocks. Each image set only contained one image of each child (following Jenkins et al., 2011: Experiment 4), although participants were not specifically told the same identities would be in each image set. Image and set order were randomized for each participant. Participants rated all 430 face images once. Faces appeared on a white background, and remained on screen until participants had responded.

Grouped-images task*:* Participants formed a single niceness impression for each group of five images of the same person, presented at the same time (as in Figure 2). Participants were informed the five photographs were of the same child, and were asked to rate how nice the child was on a scale of 1 (not at all nice), to 9 (extremely nice). The five

faces appeared in a horizontal line (face order randomized for each participant), and participants were encouraged to look at all of the face photographs when forming their impression. Participants made 86 trait judgments, one for each face identity. Faces appeared on a white background, and remained on screen until participants had responded.

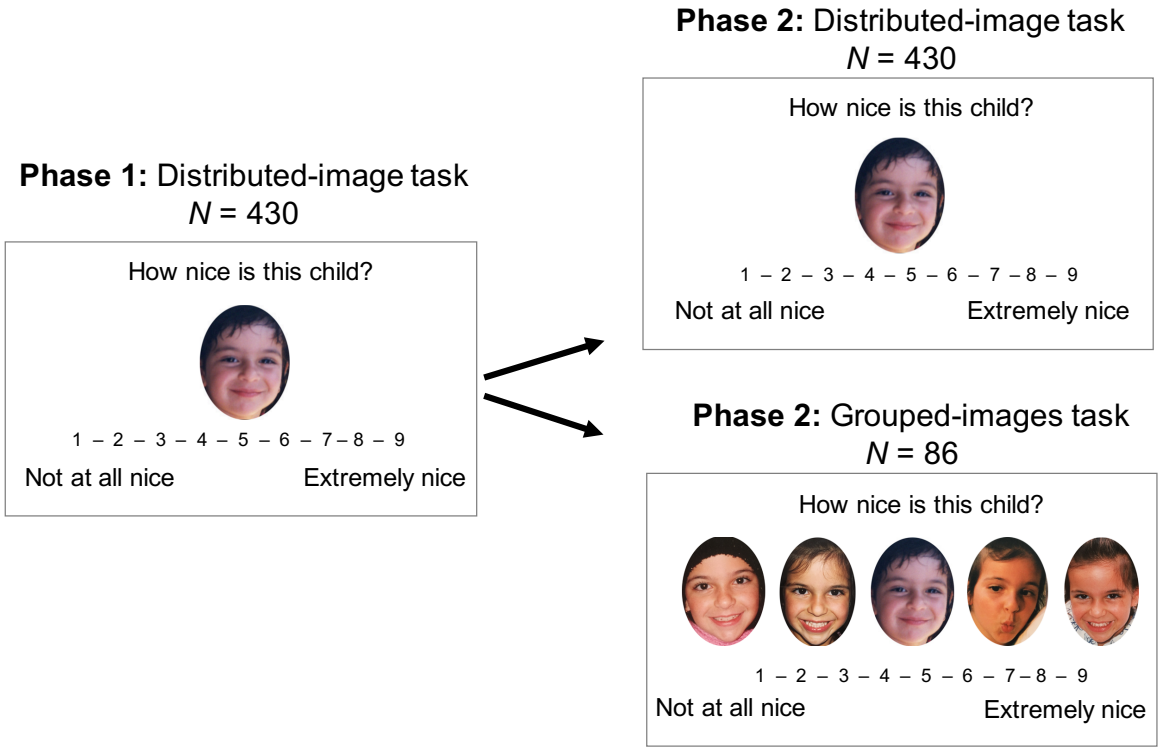

Figure 2. Study 1 procedure. In Phase 1, all participants completed the distributed-image task (N trials = 430). In Phase 2, participants were either assigned to the distributed-image task (above, N trials = 430) or the grouped-images task (below, N trials = 86).

Prior to each task, participants competed two practice trials (one female and one male face identity: data not analysed). Participants also answered basic demographic questions at the start of the experiment (e.g. age, sex, ethnicity) and provided informed consent. Participants received course credit and were debriefed after participating. Ethics approval was obtained from the University of Western Australia (RA416805).

Results and Discussion

Actual niceness. We used scores from the SDQ to index actual niceness (Table 1 for descriptive statistics). The SDQ scores from our sample were similar to norms reported elsewhere (Mellor, 2005). Scores were not normally distributed (most scores were in the higher end of the distribution; Shapiro-Wilk (86) = .858, $p < .001$), and therefore we report both parametric (Pearson's *r*: *r*) and non-parametric (Spearman's rho: *rho*) measures of association here and throughout, although they produce very similar results. We also make our data available online.

Table 1. *Descriptive statistics for actual niceness scores (SDQ) and impressions of niceness in Phase 1 (distributed-image task) and Phase 2 (distributed-image task and grouped-images task) averaged across participants. Note, criterion scores could range from 0-10, and trait ratings from Phase 1 and 2 could range from 1-9. N* = 86 face identities.

		M	*SD*	Range	Skew	Kurtosis
Criterion:	*Actual Niceness*	8.5	1.4	3.5 – 10	-1.45	2.27
Phase 1	Image Set 1	5.6	1.0	3.5 – 7.5	-0.33	-0.69
	Image Set 2	5.6	0.9	3.2 – 7.6	-0.33	-0.29
	Image Set 3	5.6	0.9	3.3 – 7.4	-0.29	-0.81
	Image Set 4	5.4	1.1	2.8 – 7.6	-0.47	-0.12
	Image Set 5	5.6	1.0	2.9 – 7.5	-0.23	-0.29
	Distributed-image Av	5.6	0.7	3.8 – 7.3	0.11	0.05
Phase 2	Image Set 1	5.4	0.9	3.2 – 7.2	-0.27	-0.25
	Image Set 2	5.4	1.0	2.7 – 7.6	-0.20	-0.14
	Image Set 3	5.4	0.8	3.5 – 7.3	-0.23	-0.60
	Image Set 4	5.3	1.1	2.8 – 7.7	-0.21	-0.49
	Image Set 5	5.4	1.0	2.8 – 7.1	-0.10	-0.62
	Distributed-image Av	5.4	0.7	3.9 – 7.3	0.23	-0.08
	Grouped-images Av	5.9	0.9	4.1 – 8.0	-0.02	-0.09

Phase 1: Distributed-image accuracy

Participant-group level accuracy for distributed images. Within each image set, we calculated a mean niceness impression for each face by averaging niceness impressions across participants (see Table 1 for descriptive statistics and Figure 3a for participant ratings). We correlated niceness impressions with actual niceness scores. Overall, impressions of

niceness showed modest accuracy. Indeed, we observed small to medium effect sizes across all image sets (Table 2a), suggesting that accuracy was robust across the different images of each child's face. Impressions were accurate across different images, despite high within-identity variability in niceness judgments (average within-identity variability = 0.54, between-identity variability = 0.52), consistent with the variability found for impressions of adult faces (Todorov & Porter, 2014).

Meta-analytical statistics (weighted by sample size) revealed almost identical results to the correlational analyses (Figure 3b): that is, impressions of niceness were accurate across the five image sets. We also examined whether impressions were accurate when averaged across ratings from the distributed images (i.e. across the five image sets). Consistent with analyses at the image set level, we observed a medium, significant correlation between perceived niceness and actual niceness (See Table 2a and Figure 3c).

Results remained significant after controlling for face sex (see Supplementary Materials Table S1 and Figure S1). We also wanted to examine whether any potential positivity bias in parents' image selection influenced our results. If our results were due to positivity bias in parents' image selection, this selection bias would likely be more strongly reflected in parents' selection of the nicest image of their child, as compared to the least nice image. We re-ran our correlation analyses selecting only the images with the highest and lowest niceness rating for each identity. Results remained significant for both of these image sets, with no differences between the sets, suggesting that parental biases could not fully account for our observed accuracy (see Supplementary Materials page 4 for details). Finally, we examined whether our results were consistent with a self-fulfilling prophecy effect, whereby impressions may become more accurate with age (i.e. for the older children) as children confirm to adults' expectations of their behaviour. However, child age did not moderate the correlation between perceived niceness and actual niceness (b = - 0.13, t = 0.95, p = .343). It

is important to note that this analysis was run post-hoc, and most children did not vary substantially with age, and so this analysis may not have provided a strong test of any self-fulfilling prophecy effect.

Table 2. *Spearman's rho and Pearson's r correlations between actual niceness and niceness impressions from a) Phase 1, the distributed-image task and b) Phase 2, the distributed-image task and grouped-images task; All N* = 86.

Participant group accuracy		Spearman's *rho*		Pearson's *r*	
		rho	*p*	*r*	*p*
a) Phase 1	Image set 1	.225	.038	.213	.050
	Image set 2	.243	.024	.308	.004
	Image set 3	.301	.005	.349	< .001
	Image set 4	.108	.323	.220	.042
	Image set 5	.224	.038	.279	.009
	Distributed-image Av	.293	.006	.368	<.001
b) Phase 2	Image set 1	.249	.021	.267	.013
	Image set 2	.250	.020	.325	.002
	Image set 3	.261	.015	.325	.002
	Image set 4	.094	.391	.191	.079
	Image set 5	.250	.020	.301	.005
	Distributed-image Av	.278	.009	.366	.001
	Grouped-images Av	.297	.005	.398	< .001

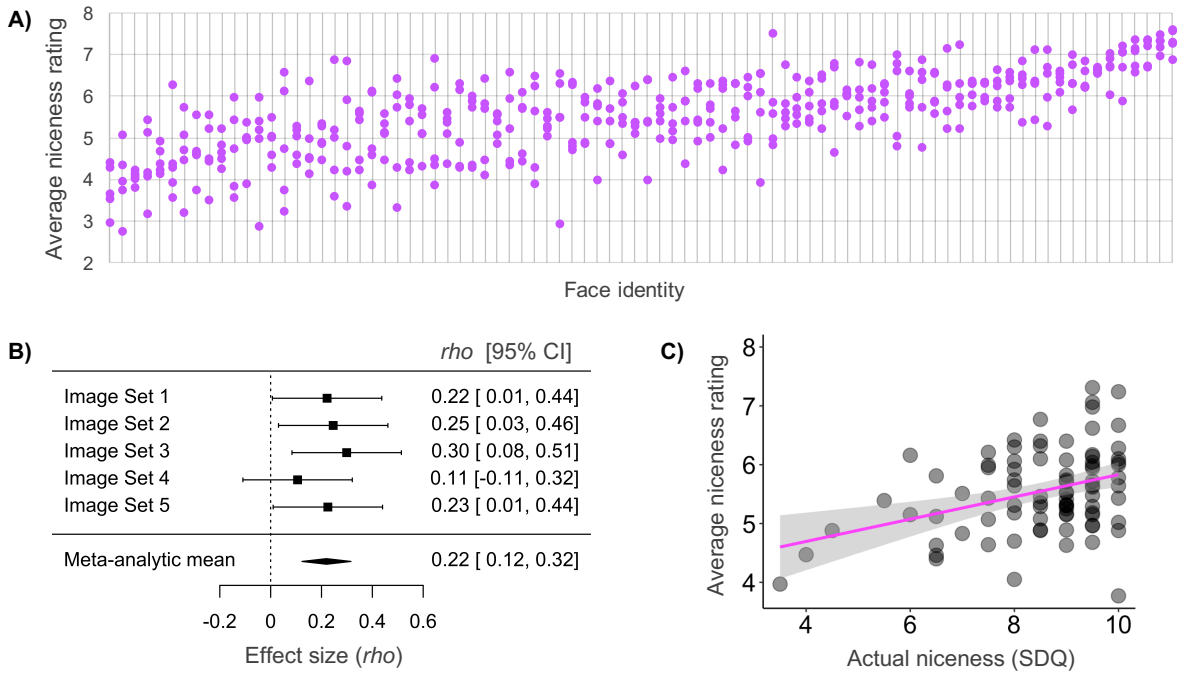

Figure 3. Niceness impressions averaged across participants' ratings in Phase 1: a) Niceness impressions for each of the five images of each face identity. Each vertical line represents one face identity ($N = 86$). Each circle represents the average rating for one of the five images of each face identity. Face identities are ordered on the *x*-axes according to their average niceness impression. b) Parameter estimates (effect size and 95% confidence intervals calculated by weighted sample size) for impression accuracy within each individual image set, and the overall meta-analytic mean effect size (Spearman's *rho*). c) Scatterplot of the association between actual niceness (on the Strengths and Difficulties Questionnaire: SDQ), and niceness impressions averaged across the five image sets. We plot the line of best fit and its confidence intervals (+- 95%).

Individual-rater level accuracy for distributed images. We calculated an accuracy score for each participant by correlating their impressions of niceness with actual niceness (Figure 4a, Table 3a). These correlations (Fisher corrected) were significantly greater than zero, for each image set and averaged across the image sets, showing accuracy at the

individual-rater level (Table 3a). As a strict test of accuracy, we also examined the proportion of participants that formed significantly accurate impressions (i.e. Spearman's *rho*: $p < .05$), based on the average distributed-image rating. Forty-nine percent of participants showed significant accuracy (Figure 4a). Therefore, significant accuracy found at the participant-group level also held at the individual-rater level, at least for half of the raters.

Table 3. Individual-rater level accuracy calculated from niceness impressions in a) Phase 1, distributed-image task, N = 47; and b) Phase 2, grouped-images task, N = 24; and distributed-image task, N = 23. We report the mean individual-rater level accuracy (rho and r) and one sample t-tests comparing the (Fisher transformed) correlations to zero.

		Spearman's *rho*			Pearson's *r*		
Rater level accuracy		*M*	*t*(df)	*p*	*M*	*t*(df)	*p*
a) Phase 1	Image Set 1	.131	10.23(46)	<.001	.117	8.77(46)	<.001
	Image Set 2	.138	10.87(46)	<.001	.171	13.15(46)	<.001
	Image Set 3	.175	13.35(46)	<.001	.199	15.16(46)	<.001
	Image Set 4	.069	4.84(46)	<.001	.134	10.29(46)	<.001
	Image Set 5	.123	7.71(46)	<.001	.165	9.56(46)	<.001
	Distributed-image Av	.200	16.03(46)	<.001	.251	19.90(46)	<.001
b) Phase 2	Image Set 1	.152	7.25(22)	<.001	.148	7.16(22)	<.001
	Image Set 2	.158	7.47(22)	<.001	.192	9.50(22)	<.001
	Image Set 3	.160	7.17(22)	<.001	.173	7.34(22)	<.001
	Image Set 4	.057	2.29(22)	.032	.127	5.23(22)	<.001
	Image Set 5	.148	5.54(22)	<.001	.184	6.45(22)	<.001
	Distributed-image Av	.190	7.07(22)	<.001	.244	8.57(22)	<.001
	Grouped-images Av	.165	7.13(23)	<.001	.221	9.13(23)	<.001

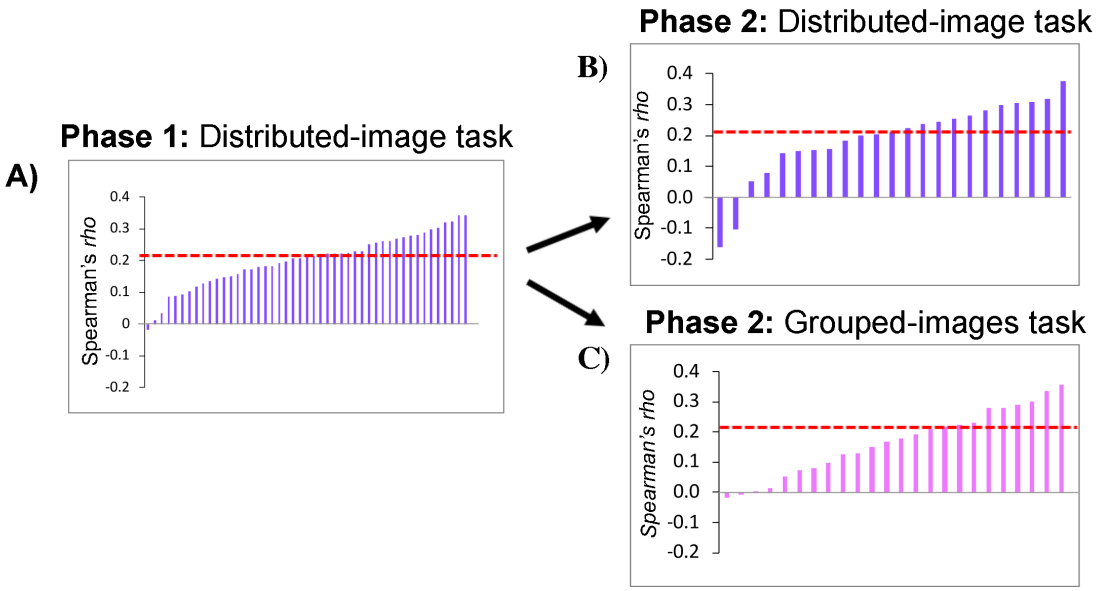

Figure 4. Participant-rater level accuracy (Spearman's *rho* correlation) for niceness impressions in a) Phase 1, distributed-image task, $N = 47$; b) Phase 2, distributed-image task, $N = 23$; c) Phase 2, grouped-images task, $N = 24$. Each rater's accuracy is represented by one vertical line. The dashed line represents the cut-off level for significantly accurate impressions at the individual level (i.e. Spearman's *rho:* $p < .05$).

Phase 2: Grouped-images accuracy

Participant-group level accuracy for grouped images. For each face identity we averaged across participants' niceness impressions and correlated these impressions with actual niceness. There was a medium, significant correlation between impressions of niceness and actual niceness (Table 2b), suggesting modest accuracy. We also compared accuracy between impressions from the grouped-images and distributed-image task. Accuracy for the grouped images was numerically greater than accuracy for any of the individual-image sets (Table 2b), but most of these differences were not significant (see Supplementary Materials Table S2). Thus, there was no strong evidence that being provided with a wider sample of faces increased the accuracy of niceness impressions.

Individual-rater level accuracy for grouped images. On average, participants formed accurate impressions of niceness (i.e. significantly greater than zero), although this accuracy was modest (Table 3b). Thirty-eight percent of participants formed significantly accurate impressions (i.e. Spearman's *rho*: $p < .05$, Figure 4c).

Individual-rater reliability in impression accuracy

As a final question, we were interested in whether individual participants who were accurate in Phase 1 were also accurate in Phase 2. To examine intra-rater reliability, we correlated mean impression accuracy from the same task (i.e. distributed-image task) in the two phases. There was a large, significant correlation, Pearson's $r = .648$, $p < .001$, suggesting that accuracy is fairly reliable for impressions of niceness.

Summary

We found that impressions of niceness were accurate for different images of the same child's face, despite within-identity variability in impression of niceness. Impressions were accurate for the different images shown individually, and shown as a group, at both the participant-group and individual-rater level. We observed small to medium effect sizes, suggesting accuracy was modest.

STUDY 2: Do shyness impressions show any accuracy?

In Study 2 a new sample of participants completed the same procedure as in Study 1, but rated the faces for how shy they looked. We correlated shyness impressions with parental reports of shy behaviour for those same children (from here on referred to as "actual shyness").

Methods

Participants. Forty-four adult participants (33 female, $M = 21.7$, $SD = 6.0$, range = 18 – 40 years, 40 Caucasian) were recruited from the University of Western Australia. In Phase

1, all participants completed the distributed-images task. In Phase 2, participants either completed the distributed-images task again ($N = 22$), or completed the grouped-images task ($N = 22$). One additional participant was tested but excluded on the basis of incomplete data (data never analysed).

Materials.

Trait measure: Shyness. Data regarding actual shyness were collected for the same children as in Study 1. To measure actual shyness, we analysed parent responses on the Colorado Childhood Temperament Inventory (CCTI: Buss & Plomin, 1984; Rowe & Plomin, 1977). Parents responded to 30 items regarding their child's personality. Here, we only analysed scores from the five items that comprised the shyness scale (Buss & Plomin, 1984: responses to the other items were never analysed). Item examples from the shyness scale include, "child tends to be shy", and, "child makes friends easily." Parents responded to each item on a 5-point scale, ranging from 1 (not at all like this / strongly disagree) to 5 (a lot like this / strongly agree). Therefore, scores could range from 5-25 with higher scores indicative of shyer behaviours. The CCTI has good validity (Webster-Stratton & Eyberg, 1982) and reliability (shyness scale $\alpha = .88$; Buss & Plomin, 1984). Here we also found good reliability for items comprising the shyness scale ($\alpha = .83$).

Stimuli and Procedure. We used the same stimuli and procedure as in Study 1, except that participants rated the children's faces for how shy they looked on a scale of 1 (not at all shy) to 9 (extremely shy).

Results and Discussion

Actual shyness. We used parent responses on the CCTI to index actual shyness (see Table 4 for descriptive statistics). The scale scores were slightly lower than norms reported elsewhere (Rowe & Plomin, 1997), indicating our sample showed less shyness than the normative sample. The scores were not normally distributed (most scores were in the lower

end of the distribution; Shapiro-Wilk (86) = .958, p = .007) and therefore, we report both parametric (Pearson's *r*) and non-parametric (Spearman's *rho*) measures of association, although they produce very similar results.

Table 4. *Descriptive statistics for actual shyness (shyness subscale from the CCTI) and shyness impressions in Phase 1 (distributed-image task) and 2 (distributed-image task and grouped-images task), averaged across participants. Note, criterion scores could range from 5-25, and trait ratings from Phase 1 and 2 could range from 1-9. N = 86 face identities.*

		M	*SD*	Range	Skew	Kurtosis
Criterion:	*Actual Shyness*	10.9	4.0	5 - 21	0.40	-0.70
Phase 1	Image Set 1	4.6	1.2	1.9 – 7.3	0.01	-0.51
	Image Set 2	4.8	1.3	2.2 – 7.1	-0.11	-1.19
	Image Set 3	4.6	1.3	1.7 – 7.2	0.03	0.26
	Image Set 4	4.7	1.3	2.2 – 7.4	0.05	0.26
	Image Set 5	4.6	1.4	1.5 – 7.5	-0.11	0.26
	Distributed-image Av	4.7	0.8	2.7 – 6.9	-0.01	0.40
Phase 2	Image Set 1	4.8	1.1	2.1 – 7.1	-0.43	0.07
	Image Set 2	5.2	1.0	3.1 – 7.1	-0.13	-0.83
	Image Set 3	5.0	1.2	2.1 – 7.7	-0.18	-0.47
	Image Set 4	5.1	1.2	2.1 – 7.3	-0.32	-0.35
	Image Set 5	4.9	1.1	2.1 – 7.5	-0.13	-0.21
	Distributed-image Av	5.0	0.7	3.4 – 7.1	0.05	0.63
	Grouped-images Av	4.8	1.1	2.3 – 8.0	0.17	0.14

Phase 1: Distributed-image accuracy

Participant-group level accuracy for distributed images. Within each image set, we averaged shyness impressions across participants for each face and correlated these impressions with actual shyness (Table 4 for descriptive statistics and Figure 5a). Impressions of shyness showed relatively large within-identity variability (within-identity variability = 1.43, between-identity variability = 0.58), and were not accurate for most image sets (Table 5a).

Next, we examined accuracy averaged across the distributed images (i.e. across the five image sets). There was a small, positive correlation between shyness impressions and actual shyness, although this correlation was not significant (Table 5a and Figure 5c). Meta-analytical statistics (weighted by sample size) revealed almost identical results to the correlational analyses (Figure 5b): that is, no compelling evidence for accuracy. Results were similar for female and male faces (see Supplementary Materials Figure S2).

Table 5. *Spearman's rho and Pearson's r correlations between actual shyness and impressions of shyness, calculated from a) Phase 1, the distributed-image task; b) Phase 2, the distributed-image task and grouped-images task. All N* = 86.

Participant group accuracy		*Spearman's rho*		*Pearson's r*	
		rho	*p*	*r*	*p*
a) Phase 1	Image set 1	-.099	.366	-.027	.806
	Image set 2	.214	.050	.226	.036
	Image set 3	-.066	.547	-.038	*.725*
	Image set 4	.057	.602	.040	*.713*
	Image set 5	.179	.098	.151	*.165*
	Distributed-image Av	.157	.148	.128	.242
b) Phase 2	Image Set 1	-.086	.430	-.038	.731
	Image Set 2	.271	.012	.301	.005
	Image Set 3	-.027	.808	.020	.854
	Image Set 4	.092	.398	.094	.390
	Image Set 5	.213	.048	.168	.123
	Distributed-image Av	.170	.118	.180	.097
	Grouped-images Av	.126	.247	.117	.284

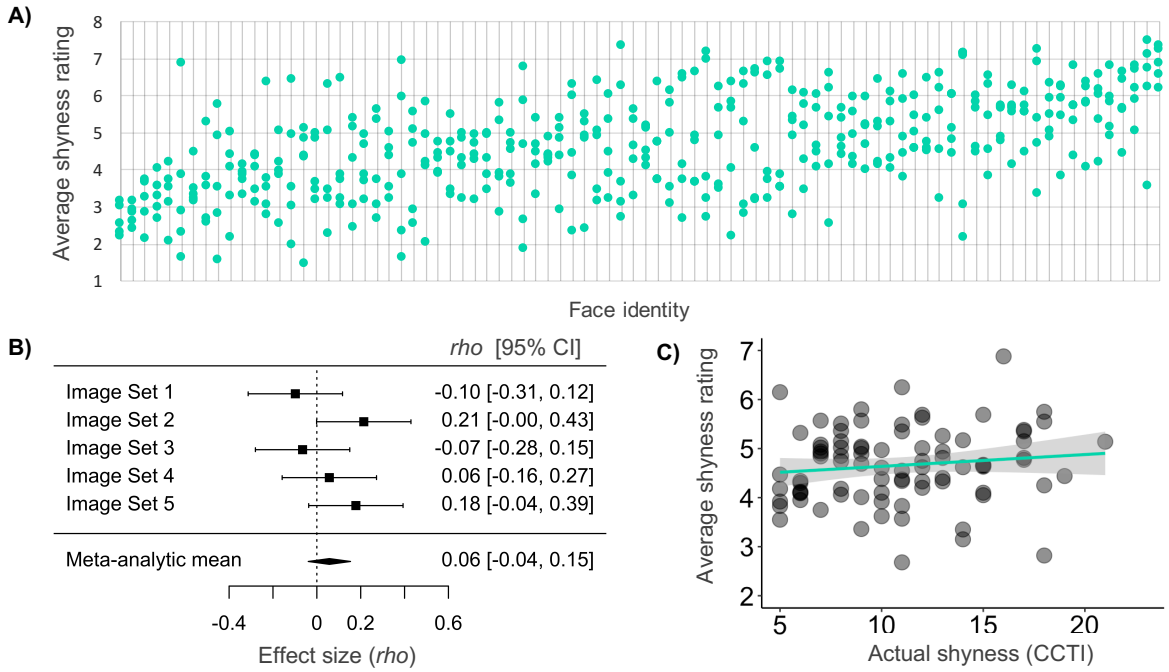

Figure 5. Shyness impressions averaged across participants' ratings in Phase 1: a) Shyness impression for each of the five images of each face identity. Each vertical line represents one face identity ($N = 86$). Each circle represents the average rating for one of the five images of each individual. Face identities are ordered on the *x*-axes according to their average shyness impression. b) Parameter estimates (effect size and 95% confidence intervals calculated by weighted sample size) for impression accuracy within each individual image set, and the overall meta-analytic mean effect size (Spearman's *rho*). c) Scatterplot of the association between actual shyness (from the Colorado Childhood Temperament Inventory: CCTI), and shyness impressions averaged across the five image sets. We plot the line of best fit and its confidence intervals (+- 95%).

Individual-rater level accuracy for distributed images. For individual raters, impressions of shyness were accurate for some image sets but not for others (Table 6a). When impressions were averaged across the image sets, accuracy was significantly greater than zero, suggesting there may be some weak accuracy at the average level (Table 6a).

However, only seven percent of participants formed significantly accurate impressions (i.e. Spearman's *rho*: $p < .05$, Figure 6a) when averaged across impressions of the distributed images. Thus, the balance of evidence suggests that if there is any accuracy for impressions of shyness, it is very modest and only for a small number of participants.

Table 6. Individual-*rater level accuracy calculated from shyness impressions in a) Phase 1, distributed-image task (N = 44); and b) Phase 2, grouped-images task (N = 22) and distributed-image task (N = 22). We report the mean individual-rater level accuracy (rho and r) and a one-sample t-test comparing the (Fisher corrected) correlations to zero*.

		Spearman's *rho*			Pearson's *r*		
Rater level accuracy		*M*	*t*(df)	*p*	*M*	*t*(df)	*p*
a) Phase 1	Image Set 1	-.039	2.98(43)	.005	-.017	1.25(43)	.220
	Image Set 2	.140	10.47(43)	<.001	.146	10.15(43)	<.001
	Image Set 3	-.033	2.18(43)	.035	-.024	1.59(43)	.119
	Image Set 4	.022	1.50(43)	.140	.022	1.44(43)	.158
	Image Set 5	.113	9.74(43)	<.001	.102	8.58(43)	<.001
	Distributed-image Av	.087	6.11(43)	<.001	.085	6.10(43)	<.001
b) Phase 2	Image Set 1	-.047	2.25(21)	.035	-.024	1.14(21)	.266
	Image Set 2	.144	7.60(21)	<.001	.158	8.29(21)	<.001
	Image Set 3	-.009	0.49(21)	.633	.008	0.36(21)	.723
	Image Set 4	.049	2.41(21)	.025	.052	2.60(21)	.017
	Image Set 5	.104	5.33(21)	<.001	.095	4.78(21)	<.001
	Distributed-image Av	.100	4.62(21)	<.001	.107	5.28(21)	<.001
	Grouped-images Av	.067	2.28(21)	.033	.072	2.54(21)	.019

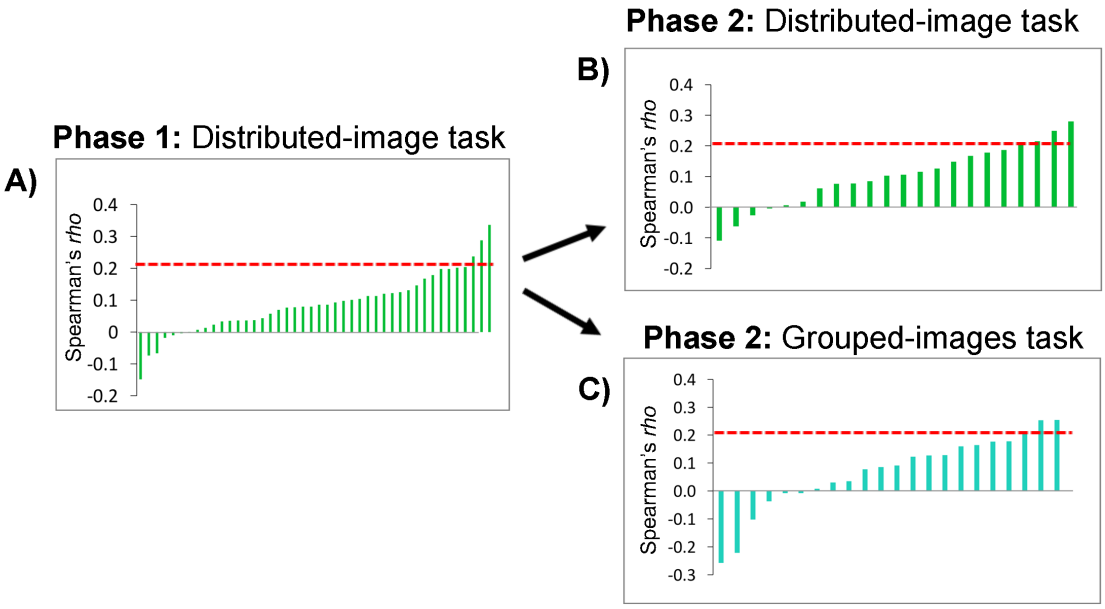

Figure 6. Rater-level accuracy (Spearman's *rho*) for shyness impressions in a) Phase 1, distributed-image task, $N = 44$; b) Phase 2, distributed-image task, $N = 22$; c) Phase 2**,** grouped-images task, $N = 22$. Each rater's accuracy is represented by one vertical line. The dashed line represents the cut-off level for accurate impressions at the individual level (i.e. Spearman's *rho*: $p < .05$).

Phase 2: Grouped-images accuracy.

Participant-group level accuracy for grouped images. There was a non-significant positive correlation between actual shyness and shyness impressions (Table 5b). Impressions from the grouped images were not significantly more accurate than impressions from each of the image sets in the distributed-image task, and the differences in accuracy were not consistently in the expected direction (see Table 5 and Supplementary Materials Table S3).

Individual-rater level accuracy for grouped images. On average, there was a small, positive correlation between shyness impressions and actual shyness. Albeit modest, this correlation was significantly greater than zero (Table 6b). Only eighteen percent of

participants formed significantly accurate impressions (Spearman's *rho*: $p < .05$, Figure 6b).

Individual-rater reliability in impression accuracy.

To examine reliability, we correlated mean accuracy from the same task (i.e. distributed-image task) in the two phases. There was a large, significant correlation, Pearson's $r = .499$, $p = .018$, suggesting that there is some reliability in individual-rater level accuracy of shyness impressions. This result suggest that the rank order of participants' accuracy was consistent, although shyness impressions were not generally accurate.

Summary

We did not find compelling evidence for accurate impressions of shyness. Shyness impressions were not accurate for most image sets, although some individual raters showed very modest accuracy for some image sets. It is possible that impressions of shyness were not consistently accurate across the image sets, because shyness impression ratings varied across the different images of the same individual. Thus, image variability may be responsible for the lack of accuracy in shyness impressions.

General Discussion

In Study 1, we found modest accuracy for adults' impressions of niceness from children's faces. Children whose parents reported them as behaving more nicely were also perceived as looking like nicer children, as compared to their counterparts. Impressions were accurate regardless of whether different images of the same child's face were presented individually or simultaneously as a group, and at both the participant-group and individual-rater level. For the first time, these results demonstrate that impressions of niceness can be accurate for different ambient images of the same person, indicating that niceness accuracy can withstand some degree of variability in appearance (e.g. Jenkins et al., 2011; Sutherland, Young, et al., 2017; Todorov & Porter, 2014). These results contribute to the current debate

surrounding the accuracy of impressions (e.g. see Bonnefon, Hopfensitz, & De Neys, 2015), and suggest modest accuracy for impressions of niceness from children's faces.

Our results are consistent with evidence for modest accuracy for similar valence-related traits inferred from adult faces. For example, accuracy has been found in judgments of trustworthiness (Slepian & Ames, 2016; Stirrat & Perrett, 2010), honesty (Bond et al., 1994; Zebrowitz et al., 1996) and aggressiveness (Short et al., 2012; Zilioli et al., 2015) from images of adult faces taken in the lab (but see Rule et al., 2013). More recently, there has also been evidence for a degree of accuracy in impressions of trustworthiness from standardized images of children's faces (Li et al., 2017, $r = .26$). Our results contribute to this growing body of evidence for accuracy in impressions, and suggest there may indeed be a link between children's facial appearance and reported nice behaviour, in an ecologically valid set of ambient images.

In Study 2, we found no compelling evidence for accurate impressions of shyness. Impressions were not accurate for most image sets, potentially because shyness impressions varied substantially across different images of the same child's face (i.e. within-identity variability = 1.43). At most, a very small number of individual participants formed accurate impressions of shyness, but only for some images. Furthermore, the size of these effects was very small.

Why might impressions of niceness be accurate?

There are several possible explanations which could account for a degree of accuracy in niceness impressions, and teasing these accounts apart represents an important future direction. First, from an evolutionary perspective, people may have adaptive mechanisms that are sensitive to intrinsically valid cues of niceness. For example, accurately inferring niceness from stranger children's faces may be adaptive from a survival perspective. That is, it would be adaptive to keep one's offspring away from "naughty children" who might disobey rules,

seek danger, and threaten the safety of offspring. Alternatively, the ability to accurately detect niceness may be over-generalised from inferring traits related to trustworthiness in adult faces. Impressions of niceness and trustworthiness are highly similar for children's faces, and are both conceptualized as the primary valence dimension in models of first impressions (see Collova et al., 2019). It is also possible that accurately inferring niceness from children's faces would enhance one's own reproductive success through optimising mate selection for one's offspring. Among hunter-gatherer societies, parents played (and often continue to play) an important role in arranging marriages for their offspring (Apostolou, 2007, 2010; Buunk, Park, & Dubbs, 2008). Critically, these marriages were typically arranged during childhood. Even when arranged marriages were not the norm, it was necessary for people to obtain their parents' approval for partner selection. In this context, parents' influence over a child's mate choice has a strong evolutionary history (Buunk et al., 2008). Interestingly, parents show a preference for their children's partners to have characteristics that signal parental investment and in-group cooperation (Buunk et al., 2008). Here, our results could suggest that impressions of niceness from children's faces may signal these caring and protective traits. Of course, it would also be adaptive to conceal any valid signals of deceptiveness, and so there may be an upper limit to this accuracy.

A second possible explanation, is that impressions of niceness from children's faces are accurate because of self-fulfilling prophecy effects. That is, children who look like they are nice might be treated more positively, and therefore end up behaving in ways that increase the accuracy of this positive impression. Here, we did not find any evidence of a self-fulfilling prophecy effect based on age. However, our study was not designed to specifically test for this effect: we did not track appearance or behaviour over time, and moreover, the age distribution of the children in the photographs was relatively limited. Previous evidence is suggestive of self-fulfilling prophecies for similar valence-related traits

such as honesty (Zebrowitz, et al., 1996) and trustworthiness (Li et al., 2017). For example, Li et al (2017) found that children who were perceived as more trustworthy became more trustworthy one year later, as compared to their untrustworthy looking counterparts, and that this effect was mediated by the peer-acceptance of those children. If the accuracy we observed was due to a self-fulfilling prophecy effect, this result would suggest that treatment by others prior to primary school (i.e. 4-11 years old) is sufficient to affect behaviour. Any attempts to reduce such effects would therefore need to take place relatively early in development. Furthermore, a self-fulfilling prophecy effect could explain the accuracy in impressions of trustworthiness from adult faces (e.g. Bond et al., 1994). That is, children who are perceived as nice may be treated positively, and therefore grow up to become trustworthy adults.

It is also possible that impressions of niceness were accurate because of the nature of the images we used. To source the images, parents were encouraged to send in the five most recent images of their child, and most parents sent in digital images that they had already taken on their phone. It is possible that these ambient images capture more valid information about behaviour and personality than face images taken in the lab under controlled conditions, which the majority of accuracy studies have used (cf. Alaei & Rule, 2019; Foo et al., 2019; Li et al., 2017; Penton-Voak et al., 2006; Pound, Penton-Voak, & Brown, 2007; Pound et al., 2008; Rhodes et al., 2013; Stirrat & Perrett, 2010; Sutherland et al., 2018). For example, nicer children might be more likely to smile for their parent's photograph, or more likely to be naturally captured smiling, than children who are less nice. In this context, using ambient images may better capture the natural behavioural tendencies of children, and therefore increase the availability of valid signals of personality. In support of this idea, Naumann, Vazire, Rentfrow, and Gosling (2009) found that impressions of full-body photographs were more accurate when targets were photographed with a spontaneous pose

and facial expression, than when their pose and expression were controlled. However, in contrast, Dumas and colleagues (2001) found that adults' impressions of competent and dysfunctional children were equally accurate for standardized images and unstandardized images of those same children. Future research could compare the accuracy of impressions from ambient images to impressions from controlled and standardized images, to investigate whether valid signals are better captured by one of these stimuli types (we return to this point later).

Why might impressions of shyness not be accurate?

It is possible that there was no compelling evidence for accurate impressions of shyness because of the limited range of children's actual shyness scores. Our sample showed lower mean shyness and slightly less variation, than norms for this scale (c.f. Rowe & Plomin, 1977), possibly because our recruitment method favoured more outgoing participants (i.e. parents and children who were active volunteers for science experiments in our lab, and who willingly provided photographs). It is possible that with a wider sample of scores (ideally, higher shyness) impressions would have been more accurate. Nevertheless, we did observe significant accuracy for impressions of niceness, despite a similarly limited range of actual niceness scores. So alone, the range of actual shyness scores cannot fully account for the limited evidence of accuracy for shyness impressions.

The characteristics of the face stimuli used might account for why we did not observe strong evidence for accurate impressions of shyness. Shyness is an important trait related to children's behaviour with *strangers*. Here, differences in the environmental contexts in which the images were taken (e.g. a photograph taken of a child in their home versus on their first day at school), might have obscured differences in the shyness of the children themselves. Indeed, we found considerably high within-identity variability in impressions of shyness as compared to niceness, suggesting that children may have displayed varying levels of shyness

across the different images. In this context, it is possible that our results reflect detection of a temporary *state* in children, as opposed to shyness as an enduring trait. That is, adults may have accurately detected a state of shyness that was specific to the context of each image (e.g. signalled by temporary facial emotional expressions), rather than unreliably detecting shyness as a trait impression across different images of the same child.

Finally, impressions of shyness might not have been accurate because these impressions may be less likely to influence behaviour towards children as compared to niceness. That is, adults might reinforce nice behaviours in children who look nice, but not necessarily reinforce shy behaviours in children who look shy. Indeed, hypothetical behaviour is more influenced by faces manipulated for niceness than shyness (Collova et al., 2019), suggesting that judgments of niceness may be more powerful in influencing actual behaviour. Thus, while a self-fulfilling prophecy effect might account for a relationship between appearance and nice behaviour, it might not have the same effect for shyness.

Limitations

One potential limitation of our study, is that parents may have been biased when selecting the photographs or when reporting actual niceness/shyness. For example, parents who report their child as being nicer might also choose nicer looking photographs of their child, consistent with a social-desirability response bias. However, there are a number of reasons why we believe this bias is unlikely to have driven our observed accuracy for niceness. First, our results are consistent with other research which has also found modest accuracy for impressions of niceness/trustworthiness using images of children taken by a stranger, and using child-peer judgments of actual trustworthiness as the criterion (Li et al., 2017). Second, the SDQ scale which we used to measure actual niceness is well-established (Hawes & Dadds, 2004). Critically, this scale shows inter-rater agreement between parent and teacher reports (Stone et al., 2010), suggesting that parents tend to answer this

questionnaire honestly. Finally, if response biases were a problem, impressions of shyness would have correlated with actual shyness, as impressions of shyness are negatively valanced for children (Collova et al., 2019). If parents had been motivated to portray their child as less shy (both in the photographs and in the questionnaire), then we would have observed a correlation between impressions of shyness and actual shyness, which is not what we found.

In summary, it is unlikely that our results can be entirely attributed to parental response biases. Nevertheless, it is interesting to consider what our results would mean if parental biases do contribute. Of course, the effects found here would no longer reflect valid cues to niceness in the face itself. However, this effect would have significant consequences in the world as it would suggest that some parents may be systematically and consistently choosing photos of their child that might help them gain positive life outcomes. For example, the images that a parent chooses to share of their child could impact other adults' expectations about how popular or intelligent that child is (Clifford & Walster, 1973). In this context, parents' selection of positive (or negative) images could have serious social consequences for children, and may lead to self-fulfilling prophecies effects, as discussed above. Indeed, in everyday life parents often share images of their child online and so our finding is important even if some element of bias was involved.

Future directions

We found that impressions of niceness were accurate for different images of the same child's face. This finding could suggest that the faces in the images consistently contained valid cues of niceness. An important question for future research will be to consider what these valid cues might be. Potential candidates include emotional expression (e.g. happiness, sadness, anger), facial attractiveness and babyfaceness, which holistically cue impressions of niceness (Collova et al., 2019). For example, children who behave more nicely might also have a face structure that resembles a happy expression (e.g. a high set brow, or an upward

turn in their lips). Alternatively, it is possible that emotional expressions were consistently captured across different images. For example, for our sample of faces, nice children might have been more likely to be naturally captured with a happy expression across multiple images, as discussed above. Likewise, children who behave less nicely might have been more likely to be captured with an angry or sad expression. For our sample of faces, most children were smiling, and impressions were generally positive (niceness rating in Phase 1: $M = 5.5$). Thus, it is also possible that subtle variations in happy expressions signalled actual niceness, such as sensitivity to genuine versus posed smiles (e.g. see McLellan, Johnston, Dalrymple-Alford, & Porter, 2010; Miles & Johnston, 2007), or smile intensity (e.g. Oosterhof & Todorov, 2009; Schmidt, Levenstein & Ambadar, 2012).

Impressions of shyness were significantly accurate for some images, but not for other images of the same children, suggesting that valid cues may have been better captured in those images. For our sample of faces, it is possible that structural facial cues might subserve accuracy in shyness, and that these cues were not consistently captured across the different images. For example, facial width-to-height ratio (fWHR) is linked to impressions of shyness (Collova et al., 2019) and the development of shy behaviour (Arcus & Kagan, 1995; Zebrowitz, Franklin, & Boshyan, 2015) in children, and varies across different images of the same person (Kramer, 2016). If the cues that are valid signals to shyness are more influenced by variable image properties than the cues that provide valid signals to niceness (e.g. consistent genuine smiles), this difference could account for why we observed accuracy for impressions of niceness but not shyness. Our results here establish a foundation for future research into the valid cues of niceness and shyness impressions, and how these cues might be impacted by variable image characteristics.

Results from Study 2 suggest that accurate impressions may not generalize to all images of a person, because impressions can change across different images. This result is

consistent with evidence that impressions do vary across different images of the same person (Jenkins et al., 2011; Sutherland, Young, et al., 2017; Todorov & Porter, 2014) and highlight the importance of considering the stimuli used in accuracy studies. For example, if we had randomly sampled images from Image Set 3, we would have concluded there was no strong evidence for accurate impressions of shyness. However, if we had sampled images from Image Set 2, we would have found significant accuracy. To date, accuracy studies have measured impressions using a single face image of a person (e.g. Bond et al., 1994; Foo et al., 2019; Graham et al., 2016; Haselhuhn & Wong, 2011; Li et al., 2017; Porter et al., 2008; Slepian & Ames, 2016; Stillman et al., 2010; Sutherland et al., 2018; Talamas, Mavor, & Perrett, 2016; Zebrowitz & Rhodes, 2004), with mixed evidence for accuracy. Our results demonstrate that some images may better capture valid cues than other images, potentially accounting for the inconsistency in accuracy results. Going forward, the field of person perception should consider whether the accuracy observed for one image of a person can generalize to other images of that same person, particularly for ambient images. Collecting this data will shift the measurement from image accuracy to person accuracy, and generate more valid estimates of impression accuracy.

To our knowledge, we present the first study to investigate impression accuracy across multiple photographs of the same person presented as a group, at the same time. In Study 1, where accuracy for niceness impressions was robust, impressions of grouped images were numerically but not significantly greater than of any individual image alone. In Study 2, where evidence of shyness accuracy was weak, there was no evidence that providing grouped images increased accuracy relative to the individual images, possibly because any valid signals of shyness were not consistently captured across all individual face images. Together, these results provide no strong evidence that impressions from grouped images of the same person are more accurate than impressions from any single image. Nevertheless, future

research (with a larger sample size of faces) should systematically test whether providing more images of the same person enhances accuracy, which we examined for the first time here. When making judgments from faces in everyday life people are likely to use more information than is available in a single photograph. Interestingly, being provided with greater variability in face information enhances accuracy on other judgments from faces, such as identity (Baker, Laurence, & Mondloch, 2017; Ritchie & Burton, 2017), and may have a similar effect for trait accuracy.

It would also be interesting to examine impression accuracy for videos of children, which may provide more information about face variability than still images (e.g. dynamic facial expressions: for a review see Krumhuber, Kappas & Manstead, 2013). Indeed, there is evidence for modest accuracy in impressions inferred from short video clips, or 'thin slices of personality', of adults (Ambady & Rosenthall, 1992; Borkenau, Mauer & Riemann 2004) and children (e.g. Tackett, et al., 2016). In particular, videos might capture any potential valid signals of shyness that may have been missed by the still images used here. Shyness is largely behaviour dependent, and here we found that shyness impressions varied substantially across different images of the same child. It would also be interesting to test whether our finding of modest accuracy for niceness generalises to impressions of adult faces, which may vary more in appearance than child faces (e.g. because of variations in makeup, facial hair, and so on).

Finally, although we found evidence of accuracy for niceness impressions, it is important to recognize the modest size of these effects. At the participant-group level, the percentage of variance shared between perceived niceness and actual niceness ranged from 1% to 9% (*rho* = .094 to .297). At the individual-rater level, roughly 50% of participants formed significantly accurate impressions of niceness, which is substantially greater than other studies investigating individual-rater level accuracy (e.g. Foo et al., 2019: 14-18%; Sutherland et al., 2018: 0-18%). However, the average individual-rater level effect size for

accuracy was also small (*rho* = .07 to .20). Considering these very modest effects, our results certainly do not suggest that people should rely on their first impressions of children's faces. Using first impressions is especially problematic when more valid information is available, such as individual behaviour (Hooper et al., 2018) or overall social norms (Olivola & Todorov, 2010). Nevertheless, these effect sizes are comparable to those found generally across the field of psychology (see Gignac & Szodorai, 2016; Richard, Bond Jr, & Stokes-Zoota, 2003; Schäfer & Schwarz, 2019), and can be considered medium to large in the context of psychological research (Funder & Ozer, 2019). So, it is still theoretically important that there is a link between adults' impressions of niceness from children's faces, and the reported behaviour of those children.

Conclusions

We found modest accuracy for adults' impressions of niceness in an ecologically valid set of naturalistic child face stimuli. Children who were perceived as looking like nicer children also scored higher on our measure of actual niceness (parent responses to well-established, pro- and anti-social behavioural questions). Moreover, impressions of niceness were accurate across different images of the same child's face, despite considerable within-identity variability in those impressions. In contrast, there was no compelling evidence for accurate impressions of shyness. These results contribute to the current theoretical debate surrounding impression accuracy, and reveal a relationship between niceness impressions of children's faces, and the behavior of those children. These results invite future research into the cues and causal mechanisms underlying this accuracy.

References

Alaei, R., & Rule, N. O. (2019). People Can Accurately (But Not Adaptively) Judge Strangers' Antigay Prejudice from Faces. *Journal of Nonverbal Behavior*, 1-13.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological bulletin, 111(2), 256.

Antonakis, J., & Dalgas, O. (2009). Predicting elections: Child's play! *Science, 323*(5918), 1183-1183.

Antonakis, J., & Eubanks, D. L. (2017). Looking leadership in the face. *Current Directions in Psychological Science, 26*(3), 270-275.

Apostolou, M. (2007). Sexual selection under parental choice: The role of parents in the evolution of human mating. *Evolution and Human Behavior, 28*(6), 403-409.

Apostolou, M. (2010). Sexual selection under parental choice in agropastoral societies. *Evolution and Human Behavior, 31*(1), 39-47.

Arcus, D., & Kagan, J. (1995). Temperament and craniofacial variation in the first two years. *Child development, 66*(5), 1529-1540.

Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological science, 21*(3), 372-374.

Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. Cognition, 161, 19-30.

Ballew, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences, 104*(46), 17948-17953.

Berkowitz, L., & Frodi, A. (1979). Reactions to a child's mistakes as affected by her/his looks and speech. *Social Psychology Quarterly, 42*(4), 420-425.

Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological science, 15*(10), 674-679.

Bond, J., Charles F, Berry, D. S., & Omar, A. (1994). The kernel of truth in judgments of deceptiveness. *Basic and Applied Social Psychology, 15*(4), 523-534.

Bonnefon, J.-F., Hopfensitz, A., & De Neys, W. (2015). Face-ism and kernels of truth in facial inferences. *Trends in cognitive sciences, 19*(8), 421-422.

Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. Journal of personality and social psychology, 86(4), 599.

Boshyan, J., Zebrowitz, L. A., Franklin Jr, R. G., McCormick, C. M., & Carré, J. M. (2013). Age similarities in recognizing threat from faces and diagnostic cues. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 69*(5), 710-718.

Buss, A., & Plomin, R. (1984). Temperament: Early developing personality traits. Hillsdale, NJ: Erlbaum.

Buunk, A. P., Park, J. H., & Dubbs, S. L. (2008). Parent–offspring conflict in mate preferences. *Review of General Psychology, 12*(1), 47-62.

Carré, J. M., & McCormick, C. M. (2008). In your face: facial metrics predict aggressive behaviour in the laboratory and in varsity and professional hockey players. *Proceedings of the Royal Society of London B: Biological Sciences, 275*(1651), 2651-2656.

Carré, J. M., McCormick, C. M., & Mondloch, C. J. (2009). Facial structure is a reliable cue of aggressive behavior. *Psychological science, 20*(10), 1194-1198.

Clifford, M. M., & Walster, E. (1973). The effect of physical attractiveness on teacher expectations. *Sociology of education, 46*(2), 248-258.

Collova, J. R., Sutherland, C. A., & Rhodes, G. (2019). Testing the functional basis of first impressions: Dimensions for children's faces are not the same as for adults' faces. *Journal of Personality and Social Psychology,* 117(5), 900.

Dumas, J. E., Nilsen, W., & Lynch, A. M. (2001). How much does physical appearance say about the psychological adjustment of competent and dysfunctional children? *Journal of clinical child psychology, 30*(3), 385-398.

Efferson, C., & Vogt, S. (2013). Viewing men's faces does not lead to accurate predictions of trustworthiness. *Scientific Reports, 3*, 1047.

Foo, Y. Z., Loncarevic, A., Simmons, L. W., Sutherland, C. A., & Rhodes, G. (2019). Sexual unfaithfulness can be judged with some accuracy from men's but not women's faces. *Royal Society Open Science, 6*(4), 181552.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science, 2*(2), 156-168.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: a research note. *Journal of child psychology and psychiatry, 38*(5), 581-586.

Graham, J. R., Harvey, C. R., & Puri, M. (2016). A corporate beauty contest. *Management Science*.

Haselhuhn, M. P., & Wong, E. M. (2011). Bad to the bone: facial structure predicts unethical behaviour. *Proceedings of the Royal Society of London B: Biological Sciences*, rspb20111193.

Hawes, D. J., & Dadds, M. R. (2004). Australian data and psychometric properties of the Strengths and Difficulties Questionnaire. *Australian and New Zealand Journal of Psychiatry, 38*(8), 644-651.

Hinsz, V. B. (1989). Facial resemblance in engaged and married couples. *Journal of Social and Personal Relationships, 6*(2), 223-229.

Hooper, J. J., Sutherland, C. A., Ewing, L., Langdon, R., Caruana, N., Connaughton, E., . . . Rhodes, G. (2018). Should I trust you? Autistic traits predict reduced appearance-based trust decisions. *British Journal of Psychology*.

Jackson, L. A., Hunter, J. E., & Hodge, C. N. (1995). Physical attractiveness and intellectual competence: A meta-analytic review. *Social Psychology Quarterly, 58*, 108-108.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

Kenealy, P., Frude, N., & Shaw, W. (1988). Influence of children's physical attractiveness on teacher expectations. *The Journal of Social Psychology, 128*(3), 373-383.

Kleisner, K., Chvátalová, V., & Flegr, J. (2014). Perceived intelligence is associated with measured intelligence in men but not women. *PloS one, 9*(3), e81237.

Kramer, R. S. (2016). Within-person variability in men's facial width-to-height ratio. *PeerJ, 4*, e1801.

Krumhuber, E. G., Kappas, A., & Manstead, A. S. (2013). Effects of dynamic aspects of facial expressions: A review. *Emotion Review*, *5*(1), 41-46.

Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological bulletin, 126*(3), 390.

Laurence, S., & Mondloch, C. J. (2016). That's my teacher! Children's ability to recognize personally familiar and unfamiliar faces improves with age. *Journal of Experimental Child Psychology, 143*, 123-138.

Leivers, S., Simmons, L. W., & Rhodes, G. (2015). Men's sexual faithfulness judgments may contain a kernel of truth. *PloS one, 10*(8), e0134007.

Li, Q., Heyman, G. D., Mei, J., & Lee, K. (2017). Judging a Book by Its Cover: Children's Facial Trustworthiness as Judged by Strangers Predicts Their Real- World Trustworthiness and Peer Relationships. *Child development*.

McLellan, T., Johnston, L., Dalrymple-Alford, J., & Porter, R. (2010). Sensitivity to genuine versus posed emotion specified in facial displays. *Cognition and Emotion, 24*(8), 1277-1292.

Mellor, D. (2005). Normative data for the Strengths and Difficulties Questionnaire in Australia. *Australian Psychologist, 40*(3), 215-222.

Miles, L., & Johnston, L. (2007). Detecting happiness: Perceiver sensitivity to enjoyment and non-enjoyment smiles. *Journal of Nonverbal Behavior, 31*(4), 259-275.

Naumann, L. P., Vazire, S., Rentfrow, P. J., & Gosling, S. D. (2009). Personality judgments based on physical appearance. *Personality and social psychology bulletin, 35*(12), 1661-1671.

Oh, D., Shafir, E., & Todorov, A. (2019). Economic status cues from clothes affect perceived competence from faces. Nature Human Behaviour, 1-7.

Olivola, C. Y., Eubanks, D. L., & Lovelace, J. B. (2014). The many (distinctive) faces of leadership: Inferring leadership domain from facial appearance. *The Leadership Quarterly, 25*(5), 817-834.

Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in cognitive sciences, 18*(11), 566-570.

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*(2), 315-324.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087-11092

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion*, *9*(1), 128.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition, 24*(5), 607-640.

Porter, S., England, L., Juodis, M., Ten Brinke, L., & Wilson, K. (2008). Is the face a window to the soul? Investigation of the accuracy of intuitive judgments of the trustworthiness of human faces. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 40*(3), 171.

Pound, N., Penton-Voak, I. S., & Brown, W. M. (2007). Facial symmetry is positively associated with self-reported extraversion. *Personality and Individual Differences, 43*(6), 1572-1582.

Pound, N., Penton-Voak, I. S., & Surridge, A. K. (2008). Testosterone responses to competition in men are related to facial masculinity. *Proceedings of the Royal Society B: Biological Sciences, 276*(1654), 153-159.

Reiss, M. V., & Tsvetkova, M. (2019). Perceiving education from Facebook profile pictures. *New Media & Society*, 1461444819868678.

Rhodes, G., Morley, G., & Simmons, L. W. (2013). Women can judge sexual unfaithfulness from unfamiliar men's faces. *Biology letters, 9*(1), 20120908.

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331-363.

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology,* 70(5), 897-905.

Rowe, D. C., & Plomin, R. (1977). Temperament in early childhood. *Journal of personality assessment, 41*(2), 150-156.

Rule, N. O., & Ambady, N. (2008). The face of success: Inferences from chief executive officers' appearance predict company profits. *Psychological science, 19*(2), 109-111.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of personality and social psychology, 104*(3), 409.

Salvia, J., Algozzine, R., & Sheare, J. (1977). Attractiveness and school achievement. *Journal of School Psychology, 15*(1), 60-67.

Satchell, L. P., Davis, J. P., Julle-Danière, E., Tupper, N., & Marshman, P. (2019). Recognising faces but not traits: Accurate personality judgment from faces is unrelated to superior face memory. *Journal of Research in Personality, 79*, 49-58.

Schäfer, T., & Schwarz, M. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*, 813.

Schmidt, K., Levenstein, R., & Ambadar, Z. (2012). Intensity of smiling and attractiveness as facial signals of trustworthiness in women. *Perceptual and motor skills, 114*(3), 964-978.

Serketich, W. J., & Dumas, J. E. (1997). Adults' perceptions of the behavior of competent and dysfunctional children based on the children's physical appearance. *Behavior modification, 21*(4), 457-469.

Seward, R. J., Bayliss, D. M., Stallman, H. M., & Ohan, J. L. (2018). Psychometric Properties and Norms for the Strengths and Difficulties Questionnaire Administered Online in an Australian Sample. *Australian Psychologist, 53*(2), 116-124.

Short, L. A., Mondloch, C. J., McCormick, C. M., Carré, J. M., Ma, R., Fu, G., & Lee, K. (2012). Detection of propensity for aggression based on facial structure irrespective of face race. *Evolution and Human Behavior, 33*(2), 121-129.

Slepian, M. L., & Ames, D. R. (2016). Internalized impressions: The link between apparent facial trustworthiness and deceptive behavior is mediated by targets' expectations of how they will be judged. *Psychological science, 27*(2), 282-288.

Stillman, T. F., Maner, J. K., & Baumeister, R. F. (2010). A thin slice of violence: Distinguishing violent from nonviolent sex offenders at a glance. *Evolution and Human Behavior, 31*(4), 298-303.

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust male facial width and trustworthiness. *Psychological science, 21*(3), 349-354.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., & Janssens, J. M. (2010). Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4-to 12-year-olds: a review. *Clinical child and family psychology review, 13*(3), 254-274.

Sutherland, C. A., Martin, L. M., Kloth, N., Simmons, L. W., Foo, Y. Z., & Rhodes, G. (2018). Impressions of sexual unfaithfulness and their accuracy show a degree of universality. *PloS one, 13*(10), e0205716.

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105-118.

Sutherland, C. A., Rhodes, G., & Young, A. W. (2017). Facial image manipulation: A tool for investigating social perception. *Social Psychological and Personality Science, 8*(5), 538-551.

Sutherland, C. A., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology, 108*(2), 397-415.

Tackett, J. L., Herzhoff, K., Kushner, S. C., & Rule, N. (2016). Thin slices of child personality: Perceptual, situational, and behavioral contributions. Journal of Personality and Social Psychology, 110(1), 150.

Talamas, S. N., Mavor, K. I., & Perrett, D. I. (2016). Blinded by beauty: Attractiveness bias and accurate perceptions of academic performance. *PloS one, 11*(2), e0148284.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Psychology, 66*(1), 519-545.

Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological science, 25*(7), 1404-1417.

Webster- Stratton, C., & Eyberg, S. M. (1982). Child temperament: Relationship with child behavior problems and parent- child interactions. *Journal of Clinical Child & Adolescent Psychology, 11*(2), 123-129.

Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological science, 26*(8), 1325-1331.

Zebrowitz, L. A., Franklin, R. G., & Boshyan, J. (2015). Face shape and behavior: Implications of similarities in infants and adults. *Personality and Individual Differences, 86*, 312-317.

Zebrowitz, L. A., Hall, J., A., Murphy, N., A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and social psychology bulletin, 28*(2), 238-249.

Zebrowitz, L. A., Kendall-Tackett, K., & Fafel, J. (1991). The influence of children's facial maturity on parental expectations and punishments. *Journal of Experimental Child Psychology, 52*(2), 221-238.

Zebrowitz, L. A., & Rhodes, G. (2004). Sensitivity to “bad genes” and the anomalous face overgeneralization effect: Cue validity, cue utilization, and accuracy in judging intelligence and health. *Journal of Nonverbal Behavior, 28*(3), 167-185.

Zebrowitz, L. A., Voinescu, L., & Collins, M. A. (1996). " Wide-Eyed" and" Crooked-Faced": Determinants of Perceived and Real Honesty Across the Life Span. *Personality and social psychology bulletin, 22*(12), 1258-1269.

Zilioli, S., Sell, A. N., Stirrat, M., Jagore, J., Vickerman, W., & Watson, N. V. (2015). Face of a fighter: Bizygomatic width as a cue of formidability. *Aggressive Behavior, 41*(4), 322-330.

Supplementary Material

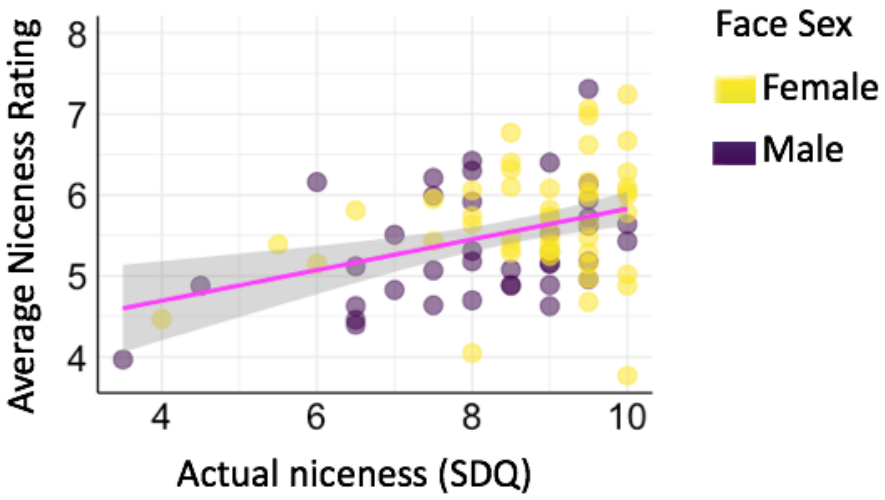

Figure S1. Study 1: Scatterplot of the association between actual niceness (on the Strengths and Difficulties Questionnaire: SDQ), and niceness impressions averaged across the five image sets for female (yellow, $N = 48$) and male (purple, $N = 38$) faces. We plot the line of best fit and its confidence intervals (+- 95%). Both female and male faces were evenly distributed across the line of best fit.

Table S1.

Study 1: Pearson's r correlations between actual niceness and niceness impressions from a) Phase 1, the distributed-image task and b) Phase 2, the distributed-image task and grouped-images task. Analyses were controlling for face sex.

		Controlling for face sex (N = 86)
a) Phase 1	Image set 1	r = .189
		p = .083
	Image set 2	r = .287
		p = .008
	Image set 3	r = .360
		p < .001
	Image set 4	r = *.233*
		p = .032
	Image set 5	r = .268
		p = .013
	Individual image Av.	r = .325
		p = .002
b) Phase 2	Grouped images Av.	r = .359
		p < *.001*
	Distributed image Av.	r = .313
		p = .004

Table S2.

Study 1: Significance tests (Fisher Z transformation; Steiger, 1980) comparing niceness accuracy (measured as Spearman's rho and Pearson's r) between the grouped images and distributed image sets, from Phase 2. All N = 86.

		Spearman's *rho*		Pearson's *r*	
Compared to Grouped-images Av		*z*	*p*	*z*	*p*
Phase 2	Image set 1	0.63	.528	1.74	.082
	Image set 2	0.58	.562	0.98	.327
	Image set 3	0.41	.682	0.88	.379
	Image set 4	2.39	.017	2.62	.009
	Image set 5	0.53	.597	1.16	.246

Study 1: Are impressions accurate for images rated as most/least nice?

We wanted to determine whether parental biases in image selection might have accounted for our observed accuracy. Any potential positivity bias in parents' selection should be least pronounced for images rated as least nice. That is, although parents may have been motivated to select some positive images of their child, they probably also chose at least one less positive image. In contrast, it is unlikely that any parents would have been motivated to select negative images. If biases were a problem, then we might have observed a larger correlation between impression of niceness and actual niceness for the images rated as most nice, as compared to the images rated at least nice. We examined whether impressions of niceness were more accurate for the image of each child which was rated as most nice, as compared to least nice. For images rated as most nice, there was a significant correlation between niceness impressions and actual niceness (Pearson's $r = .307$, $p = .004$). However, this correlation was not significantly greater than the correlation for images rated as least nice (Pearson's $r = .344$, $p = .001$; Fisher's $z = 0.39$, $p = .694$). Thus, it seems unlikely that parental biases in image selection could fully account for our observed accuracy.

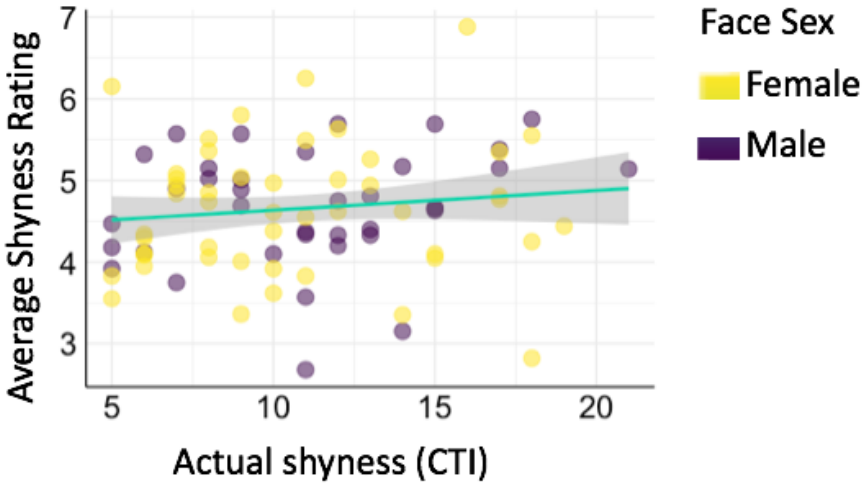

Figure S2. Study 2: Scatterplot of the association between actual shyness (from the Colorado Childhood Temperament Inventory: CCTI), and shyness impressions averaged across the five image sets for female (yellow, $N = 48$) and male (purple, $N = 38$) faces. We plot the line of best fit and its confidence intervals. (+- 95%). Both female and male faces were evenly distributed across the line of best fit.

Table S3.

Study 2: Significance tests (Fisher Z transformation; Steiger, 1980) comparing shyness accuracy (measured as Spearman's rho and Pearson's r) between the grouped images and distributed image sets, from Phase 2. All N = 86.

Compared to Grouped-images Av		Spearman's *rho*		Pearson's *r*	
		z	*p*	*z*	*p*
Phase 2	Image set 1	2.10	.040	1.44	.150
	Image set 2	-1.40	.161	-1.72	.086
	Image set 3	0.94	.347	0.95	.172
	Image set 4	0.30	.761	0.21	.831
	Image set 5	-0.80	.423	-0.48	.629