Journal of Research in Personality 75 (2018) 26–36

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Full Length Article

Cues to mental health from men's facial appearance ☆

Robert Ward *, Naomi Jane Scott

Wolfson Centre for Clinical and Cognitive Neuroscience, School of Psychology, Bangor University, United Kingdom

ARTICLE INFO

Article history:
Received 14 July 2017
Revised 9 April 2018
Accepted 20 April 2018
Available online 23 April 2018

Keywords:
Facial appearance
Mental health
Attractiveness
Masculinity

ABSTRACT

Previous work shows that mental health can be evident from neutral facial appearance. We assessed the accuracy of mental health perceptions from facial appearance, and how perceived mental health related to other appearance cues, specifically attractiveness, perceived physical health, and masculinity. We constructed composite images from men scoring high and low on autistic quotient, depressive symptoms, and schizotypy inventories, and asked observers to rate these images for mental health. We found perceived mental health reflected actual mental health in all cases. Furthermore, the accuracy of mental health inference was not fully explained by other appearance cues. We consider implications of accurate mental health detection from appearance, and the possibility that appearance could be a risk factor for mental health issues.

1. Introduction

Humans are both highly visual and highly social. Consistent with this dual nature, people make important social inferences on the basis of mere appearance, and perhaps surprisingly, social inferences based on impoverished "thin slices" of nonverbal behavior are sometimes accurate (Ambady & Rosenthal, 1992). Observers watching short video clips of a stranger quietly facing the camera could draw accurate personality inferences, that is, better than chance agreement with the stranger's self-reported personality (Borkenau & Liebler, 1992). Many thin-slices experiments include target-controllable visual cues – such as hairstyles, clothing, cosmetics, posture and expression – which can improve the accuracy of visual judgements, but such cues are not always necessary (e.g., Naumann, Vazire, Rentfrow, & Gosling, 2009). Furthermore, facial appearance can still drive accurate trait inference, even under what might be described as impoverished conditions with minimal controllable cues present. Of most relevance here, trait neuroticism can leave visual evidence in such controlled circumstances, with front-facing neutral photos, and hairstyle, clothing, and cosmetics cues removed or minimized (e.g., Little & Perrett, 2007). Although the cues underlying accurate discrimination of neuroticism are not known, some potential cues can be eliminated from the images without eliminating accuracy, such as the jawline (Kramer & Ward, 2010), and postural cues like head position (Jones, Kramer, & Ward, 2012).

Trait neuroticism is associated with a number of mental health vulnerabilities (e.g., Kotov, Gamez, Schmidt, & Watson, 2010), and is highly correlated with a general factor for psychopathology (Caspi et al., 2014). Therefore, accurate identification of neuroticism implies that observers might be able to assess, to some degree, mental health status on the basis of mere facial appearance. Indeed, even when facial images are taken under controlled conditions, accurate mental health discriminations have been found. Photographic composites of men and women with high and low scores in the dark triad (Machiavellianism, narcissism, and psychopathy) could be accurately identified (Holtzman, 2011). Similar results have been found with depressive symptoms, and furthermore, observers negatively evaluated the high-depression images, for example mistakenly identifying them as less friendly and agreeable (Scott, Kramer, Jones, & Ward, 2013).

More recently, Daros, Ruocco, and Rule (2016) demonstrated that neutral facial appearance allowed accurate discrimination of clinically-assessed borderline personality disorder (BPD), based on neutral photographs of individual targets. Accuracy was robust and maintained when evaluating the specific condition underlying

☆ **Notes**: This study was not preregistered. RW contributed to the design, analysis, and writing of this study. NJS contributed to the design and writing. We thank Emily Butler for her help in running this study, and David Cooper for his assistance in a pilot study. We thank Nick Holtzmann for helpful comments.

* Corresponding author at: Wolfson Centre for Clinical and Cognitive Neuroscience, School of Psychology; Bangor University, Brigantia Building, Penrallt Road, Bangor LL57 2AS, United Kingdom.

E-mail address: r.ward@bangor.ac.uk (R. Ward).

https://doi.org/10.1016/j.jrp.2018.04.007
0092-6566/

the experimental manipulation (i.e., descriptions of BPD); when evaluating a related mental disorder (depression); and when evaluating a general umbrella term ("mental disorder"). Daros et al. also found that judgements of mental health traits were associated with perceptions of more negative emotional states in the neutral photos. The emotion cues in the neutral photographs were evidently subtle, and nothing as obvious as a mouth curled in a smile or frown, as presenting only the lower face (or only the upper face) abolished accuracy. The attribution of negative emotions in this case might then reflect a general willingness to make negative social evaluations of facial appearances correlated with low mental health.

Scott, Kramer et al. (2013) noted the vicious cycle implicated by such observer reactions: Without saying a word or making an action, people with high risk of mental health issues, and who might be particularly vulnerable to social exclusion, could be continuously and involuntarily broadcasting a signal which is negatively evaluated by observers. Inferences of mental health from mere appearance could therefore plausibly lead to harmful outcomes.

While mental health inference from appearance may carry significant potential impact, we do not understand the bases of these judgements. Mental health is a complex construct which might plausibly be reflected in a number of visual appearance cues. Here we investigated whether the actual mental health of targets influenced their perceived mental health, above and beyond effects of three well-studied visual cues: attractiveness, perceived physical health, and masculinity. As reviewed below, it is reasonable to hypothesize that any or all of these cues may be bases for accurate mental health inference.

To begin, people with attractive faces are rated more positively than unattractive people for diverse social traits (Eagly, Ashmore, Makhijani, & Longo, 1991; Miller, 1970), including mental health (Martin, Friedmeyer, & Moore, 1977). Indeed, attractiveness and mental health may be correlated. Psychiatric patients tend to be less attractive than controls (Farina et al., 1977; Napoleon, Chassin, & Young, 1980); and facial attractiveness may affect psychological well-being and risk of depression (Datta Gupta, Etcoff, & Jaeger, 2016). A notable negative result is the large twin study of McGovern, Neale, and Kendler (1996), which found no relationship between facial attractiveness and depression. Of course, any correlation of attractiveness and mental health could result from different, non-exclusive, causal mechanisms. Mental health problems could cause reduced attractiveness (e.g., sleep deprivation, Axelsson et al., 2010). Alternatively, by an interactional model such as Coyne (1976), repeated negative reactions from observers could themselves cause mental health issues. By such an account, attractive people might find themselves in better psychological situations, owing to the more favorable responses of others (Burns and Farina, 1992; O'Grady, 1982). However, regardless of the direction of causality, a plausible hypothesis is that the accuracy of mental health judgements from facial appearance might be explained simply on the basis of attractiveness.

A very similar account could be made for perceived physical health. Mental and physical health are correlated (e.g., Colton & Manderscheid, 2006; Hays, Bjorner, Revicki, Spritzer, & Cella, 2009). For example, personality disorders of all types are associated with significantly increased risk of heart disease (Moran et al., 2007). Large-scale surveys find that psychotic disorder is associated with loss of physical health-related quality of life (Saarni et al., 2010; Strine, Chapman, Balluz, Moriarty, & Mokdad, 2008). As with attractiveness, the direction of causality relating physical and mental health is unclear. But again, regardless of causality, observer accuracy for mental health might plausibly result from cues to perceived physical health.

Accurate mental health inference might also be based on facial masculinity. Sex hormones certainly affect adult facial appearance (e.g., Tanner, 1990), and the influence of sex hormones on mental health has been proposed in multiple forms. Perhaps most notably, the extreme male brain theory of Baron-Cohen (2002) links prenatal testosterone exposure to the high autistic quotient (AQ) scores associated with mechanistic as opposed to empathetic thinking. Consistent with the possibility of a common hormonal basis for facial appearance and autistic spectrum disorder (ASD), Scott, Jones, Kramer, and Ward (2015) found that composites of high-AQ men were judged as more masculine than those of low-AQ men, and Tan et al. (2017) found increased masculinity in the facial shape of boys and girls with ASD compared to those without. Other mental health disorders also show some degree of sex-specificity, for example, women tend to show higher rates of depression (Kessler, 2003), while men seem to be more at risk from schizophrenia (Aleman, Kahn, & Selten, 2003). If facial gender were related to sex-specific risk to mental health, that would be consistent with a hormonal influence on both: for example, if high schizophrenia were associated with facial masculinity, and high depression with facial femininity.

Finally, attractiveness, perceived physical health, and masculinity are systematically related. Perceptions of physical health are closely related to attractiveness, and many facial characteristics which are attractive cross-culturally, including symmetry, averageness, and coloration, are often discussed as signals of physical health and reproductive fitness (Little, Jones, & DeBruine, 2011). Interestingly, while strong correlations between attractiveness and perceived physical health are readily found (e.g., Rhodes et al., 2007), a robust relationship between facial attractiveness and actual physical health has proven somewhat elusive (Foo, Simmons, & Rhodes, 2017). Masculinity is also related to the attractiveness of men's faces, although in a nonlinear way. Low levels of facial masculinity are not sexually attractive, and very high levels are associated with aggression and danger (Johnston, 2006). The most attractive men's faces show an intermediate level of dimorphism, which varies somewhat between studies, from somewhat more feminine than the male average (e.g., Perrett et al., 1998) to somewhat more masculine (e.g., DeBruine et al., 2006). Indeed, evolutionary approaches to signals of immunocompetence tie all three traits to male reproductive fitness (Fölstad & Karter, 1992; Gangestad & Buss, 1993; and for critical review Scott, Clark, Boothroyd, & Penton-Voak, 2013). Given these relationships, we will need to consider the correlations between attractiveness, perceived physical health, and masculinity, in order to understand how each influences perceptions of mental health.

1.1. The current study

Our primary objective is to further understand the visual bases of mental health inference. Does *perceived* mental health reflect *actual* mental health? If so, how do these perceptions relate to other well-studied facial cues? Our general method was to present observers with neutral facial images, composed from men self-reporting high or low on different mental health conditions. We measured ratings of mental health and other judgements to these images. Our analytic strategy was to first model perceived mental health on the basis of attractiveness, perceived physical health, and masculinity. We then compared this model to one additionally including the actual mental health of the images, to see whether actual mental health was incorporated into observer ratings, above and beyond the effects of the other facial cues.

We also examined other secondary issues. (1) Given that facial masculinity has nonlinear effects on attractiveness, might it have nonlinear effects on perceived mental health? (2) Do different

mental health conditions vary in how observers use different appearance cues? And (3) could the cues used for mental health judgements be localized to different facial regions? In additional exploratory analyses, we also assessed whether the influence of actual on perceived mental health might be mediated by different facial cues, including attractiveness.

2. Method

2.1. Stimulus creation

Our aim was to create stimuli reflecting any regularities in the facial appearance of men scoring high and low on traits of mental health. We wanted to use composite, or average, facial images because we were interested in systematic rather than fluctuating appearance differences – for example, systematic effects of masculinity as opposed to fluctuating asymmetries in the face.

We therefore created composite images from men with the most extreme scores on three mental health inventories, relating to depressive symptoms, schizotypy, and autistic spectrum disorder (ASD). We chose these particular disorders for multiple reasons: they represent a diversity of neuropsychiatric taxonomy (e.g., Caspi et al., 2014; Kotov et al., 2017; Crespi & Badcock, 2008); there are validated and easily applied inventories for each; and all show a clear range of variation in nonclinical populations. We have also previously worked with facial appearance relating to depressive symptoms (Scott, Kramer et al., 2013) and AQ (Scott et al., 2015), and were interested in further exploring these two conditions. We focused on men's faces, following the Scott et al. (2015) finding that AQ scores were associated with the rated masculinity of men's, but not women's, faces.

Our photographic database contained 91 male Bangor University students of self-reported white ethnicity, and who had previously completed three mental health inventories: AQ (Autistic Quotient; Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001); the self-report version of Inventory of Depressive Symptomology (IDS; Rush, Gullion, Basco, Jarrett, & Trivedi, 1996); and the Schizotypal Personality Questionnaire (SPQ; Raine, 1991), along with other personality and demographic information. Standardized alphas for the sample on these inventories were: AQ (.51), IDS (.88), and SPQ (.83). Each man had a neutral facial photo taken concurrently (face forward; hair back; neutral expression; no jewelry, beards, or glasses). This database was used previously in Scott, Kramer et al. (2013, experiment 2), and Scott et al. (2015; experiment 2), and additional details are available in those sources.

From this database, we selected the 18 men scoring highest and lowest on each of the three inventories, and made composites from their face images. Mean scores for the groups are in Table 1. As a rough guide, the low mental health (unhealthy) groups fell within the top 10% of previous reports for the IDS (Rush et al., 1996) and SPQ (Raine, 1991), and top 18% for males on the AQ (Baron-Cohen et al., 2001). From each selection of 18 men, we made a set of nine composites of four individuals. The assignment of men to composites was made so that each man appeared in two composites, but no pair of men appeared in more than one composite. Compared to previous methods in which a single high or low trait composite is created (e.g., Scott, Kramer et al., 2013, experiment 2), this method introduces stimulus variation, which in turns allowed us to include stimuli as a random factor in our analyses.

The resulting full-face images were cropped in width according to the widest part of the face, and in height from hairline to chin. They were then scaled to a standard height of 500 pixels. Each full-face image was further manipulated to produce variation in information Source (Full, Inner, or Eyes; see Fig. 1). For the inner-face images, a region around the eyes, nose, and mouth was created by occluding the top 125 pixels; the bottom 75 pixels; and then left and right occluders 85 pixels wide and 225 pixels high. For the eye-region images, the top 125 pixels and the bottom 250 pixels were occluded; the eyes and eyebrows were always included within these images.

The effect of this procedure was to create a stimulus set of 162 images, reflecting 3 mental health Traits (AQ, Depression, Schizotypy) * 2 values of Actual mental health (High or Low) * 3 Sources (Full, Inner, and Eyes) * 9 stimulus variants.

2.2. Observers

A total of 253 observers (173 women, 73 men, 7 not reported; mean reported age = 21y, SD = 4.6) were tested. An initial 203 observers were tested over a two-day mass collection. The stopping rule was simply the maximum number that could be recruited over the two days. Each of these observers rated either mental health, masculinity, or attractiveness, with the judgement varied between-observers, as described below. An additional 50 observers were subsequently recruited to rate physical health.

2.3. Procedure

Each stimulus image was presented individually and rated by observers. Stimuli were blocked by Source, with the three possible Sources (Full, Inner, Eyes) appearing in random order for each observer. Within a block, all 54 of the Trait * Value * Variant images were presented in random order. In this way, each observer rated every image for a single judgement. Between-observers, four judgements were made to each stimulus image: Mental Health, Attractiveness, Masculinity, and Physical Health.

On each trial, a description of the judgement to be made appeared below the face, with six responses, arranged from negative to positive valence. For the mental health judgement, the description was, "Rate this person's appearance of mental health. Unhealthy suggests risk for mental health problems like depression, anxiety, autism, and other social problems. Healthy suggests generally emotionally stable, content, and socially aware." The response options were labelled: very unhealthy, unhealthy, somewhat unhealthy, somewhat healthy, healthy, and very healthy. For the masculinity judgement, the question was, "Compared to other men, how masculine or feminine is this face?", with response options very feminine, feminine, somewhat feminine, somewhat masculine, masculine, very masculine. For the appearance

Table 1
Inventory scores for low and high mental health groups.

	Low		High		Full sample	
Inventory	*M*	*SD*	*M*	*SD*	*M*	*SD*
AQ	24.2	2.8	10.5	2.6	17.7	5.9
IDS	24.5	7.6	6.7	2.4	15.1	8.9
SPQ	40.9	9.1	9.5	5.3	24.5	13.4

Note. Means and SDs for the 18 targets in each of the low and high mental health groups, and for the full sample of 91 men. AQ = Autistic Quotient (Baron-Cohen et al., 2001); IDS = Inventory of Depressive Symptomology (Rush et al., 1996); SPQ = Schizotypal Personality Questionnaire (Raine, 1991).

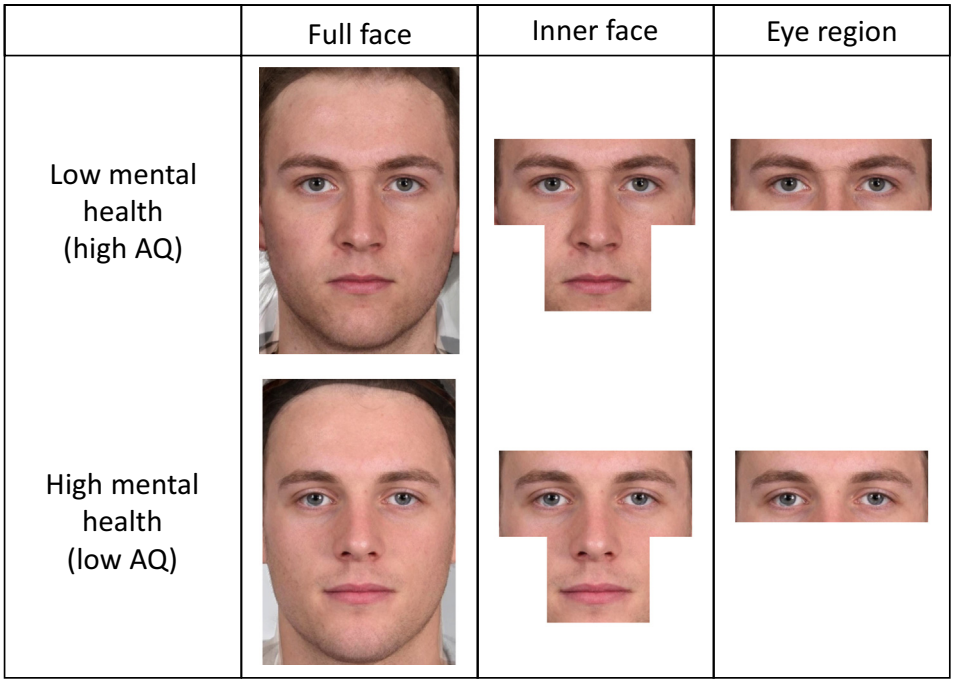

Fig. 1. Sample stimuli. For each trait (autistic quotient or AQ in this case), there were 9 composites for low and high health. The full-face stimuli were cropped as described in Methods to create the inner and eyes stimuli.

judgement, the question was, "How attractive is this face?", with response options very unattractive, unattractive, somewhat unattractive, somewhat attractive, attractive, very attractive. For the physical health judgement, the description was, "Rate this person's appearance of physical health", with response options very unhealthy, unhealthy, somewhat unhealthy, somewhat healthy, healthy, very healthy.

The first 203 observers (rating either mental health, masculinity, or attractiveness) were tested in a computer lab with approximately 3 dozen computers. Multiple observers were tested simultaneously, and multiple experimenters were present during testing to monitor behavior and answer questions. On arrival, each observer was taken through an informed consent procedure in which the task was described, and assigned a label with ID number. An experimenter would direct the observer to an available computer, and enter their ID number. The observer and experimenter would review the onscreen instructions, and observers were encouraged to raise any questions they had. The rating task took roughly 10–15 min.

The task was presented via the local intranet, with the between-observer judgement controlled centrally. Therefore, at any given time, almost all observers were performing the same judgement. However, after the ratings of 75 observers had been recorded for a particular judgement, the task was changed on the central server to begin collection for the next judgement. This change was made manually, while some observers were still finishing, and so there was some unavoidable imprecision. The order and number of observers for each judgement were: mental health (76 observers), masculinity (82), and finally attractiveness (43 observers). As we under-recruited over the two days, we had fewer observers for the final judgement. However, rather than collect data in excess of our initial stopping rule, we proceeded with data analysis, and as we will see, there seem to be no issues arising from the difference in observer numbers. Following the face judgement task, observers proceeded to a second station for another task, relating to automatic imitation. The combination of tasks was done purely for efficiency of mass recruitment; the imitation task is not relevant to this experiment and not discussed here. The final 50 observers, rating physical health, were recruited for course credit after the mass collection exercise, and tested individually in a quiet room.

Although we assessed observer accuracy on three mental health conditions, we elected to ask observers to simply rate the appearance of mental health, in the broad sense, as opposed to the appearance of specific disorders. This decision reflected several considerations, including complications due to definitional issues: it is not easy to translate standard clinical definitions into simple criteria for observers. There is also overlap of items on different mental health inventories, for example, interpersonal difficulties are relevant to both schizotypy (Raine, 1991) and AQ inventories (Baron-Cohen et al., 2001). Even if definitional issues could be addressed, by asking about mental health broadly, we were able to tap into observers' abilities to identify there is "something wrong", even if they were unsure how best to categorize the problem represented. Finally, as we are not asking whether untrained observers can make what would be essentially neuropsychiatric diagnoses from facial appearance, the broad question of mental health is entirely suitable for our question, of whether perceived mental health can be distinguished from other appearance judgements.

3. Results

Anonymized data files and fully commented R scripts for all analyses and modelling are included and will be publically available.

3.1. Full-face ratings

Mean observer ratings for full-face stimuli are presented in Fig. 2, and judgements of mental health for the three traits (AQ, Depression, Schizotypy) are shown in the left column of this figure. Intraclass correlations were satisfactory for all ratings (mental health: ICC(2,76) = .94; masculinity: ICC(2, 83) = .95; attractive-

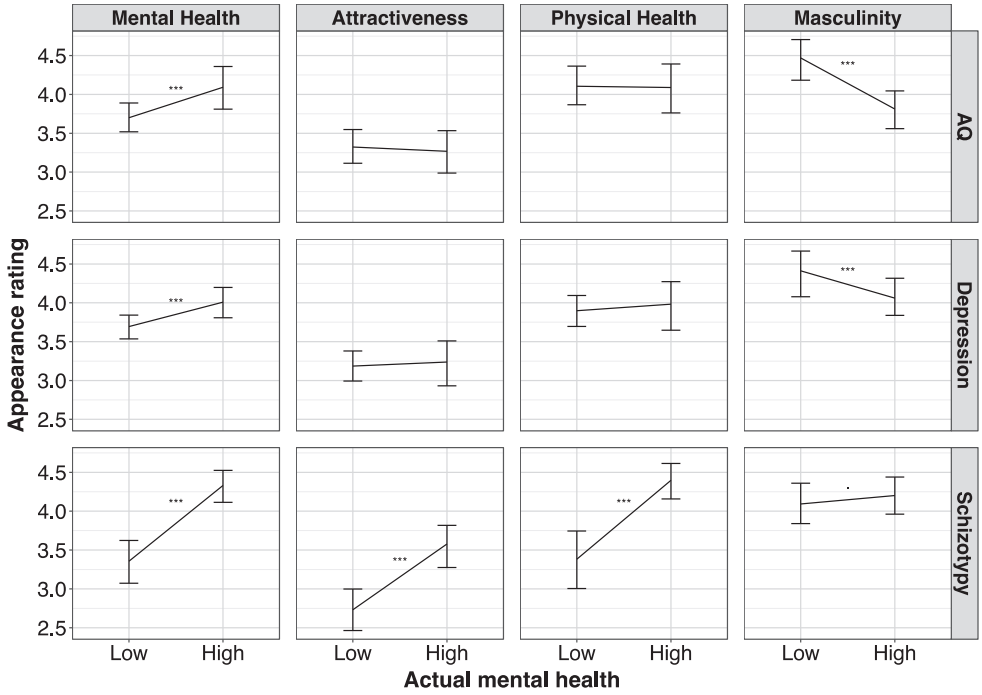

Fig. 2. Mean ratings for full face images. Columns provide rating; the mental health trait used to make the high and low composites in rows. For example, the upper right panel shows the ratings of masculinity for composites low and high in AQ. Error bars mark 95% confidence intervals over stimuli. *** = significant effect of Actual mental health, p < .0001. Otherwise, p > .05.

ness: ICC(2,44) = .89; physical health: ICC(2,50) = .93). A 2 × 3 ANOVA (Actual Mental Health: High or Low; Trait: AQ, Depression, or Schizotypy) confirmed that *perceived* mental health reflected *actual* mental health: observers gave higher mental health ratings to images of high compared to low mental health, $F(1, 225) = 240.7$, $p < .0001$. Planned comparisons confirmed this advantage for all three traits, AQ: $t(75) = 6.15$, $p < .0001$, $d = .69$; Depression: $t(75) = 5.65$, $p < .0001$, $d = .58$; Schizotypy: $t(75) = 14.4$, $p < .0001$, $d = 1.75$. Differences in the accuracy of mental health between traits were suggested by a significant interaction of Trait * Actual Mental Health, $F(2, 225) = 33.4$, $p < .0001$. However, this interaction should be interpreted with caution, and we will see below it is premature to conclude that the strength of mental health cues differed between traits. The effects of actual mental health on ratings of attractiveness, physical health, and masculinity (the other columns of Fig. 2) were more mixed. We avoid detailed consideration of these effects, pending the mixed-effects approach below, which proved to give a useful account of all variables.

3.2. Accuracy and predictors of mental health ratings

The evidence in Fig. 2 shows that observer perceptions of mental health were accurate for all traits tested, but does not tell us whether this accuracy can be attributed to attractiveness or other judgements, as opposed to actual mental health. Our primary analysis therefore used mixed-effects regression modelling to better understand how perceived mental health was related to actual mental health and these other judgements.

We first examined the correlations between variables and the possibility of multicollinearity (Table 2). The main concern was the high variance inflation factors (VIF) for attractiveness and physical health (VIF > 9), driven by the high correlation between them ($r = .94$). We replaced these two factors with a single variable in our models, the first component (PC1) from the principal component analysis (PCA) of attractiveness and physical health. This approach was attractive statistically, as (a) this single component accounted for nearly all variance for both attractiveness ($r^2 = .96$), and perceived physical health ($r^2 = .98$); (b) the revised VIF scores were low (Table 2); and (c) all models converged robustly. This approach was also attractive conceptually, as attractiveness and perceived physical health (if not actual physical health) are closely related, as we reviewed earlier. For easier exposition, while recognizing this component reflects the combination of attractiveness and physical health ratings, we will refer to it as *Latent attractiveness*.

We then ran and compared mixed-effects models predicting mental health ratings using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015), simultaneously modelling both observers and stimuli as random effects, for each of the 4104 full-face ratings of mental health (i.e., each of the 54 ratings for each of the 76 observers rating mental health).

In the "baseline" model, an observer's rating of mental health for a given image was predicted using fixed effects comprising the latent attractiveness of that image (based on attractiveness and physical health ratings from other observers), and the perceived masculinity of that image (mean rating from other observers). The "test" model was identical except for the addition of actual mental health as a fixed effect. The null or random-effects only model included no fixed effects and predicted mental health ratings only on the basis of the random effects and a global intercept. All three models shared an identical random effects structure, comprising random intercepts for observers and for image, and random slopes by observer for latent attractiveness, masculinity, and actual mental health. This structure allowed individual observers to show different overall biases in the use of the mental health scale, different sensitivities to the other ratings, and different sensitivities to actual mental health (that is, some observers may be more discriminating than others). In all models, continuous ratings variables were rescaled to zero mean and unit variance. Actual mental health was modelled as low health = −1, high health = 1.

Our analyses report the significance of actual mental health by comparing the baseline and test models (identical other than the

Table 2
Means, standard deviations, variance inflation, and correlations of ratings.

Variable	*M*	*SD*	VIF	1	2	3	4
1. Attractiveness	3.22	0.47	9.37				
2. Physical health	3.98	0.55	9.51	.94			
3. Masculinity	4.17	0.47	1.20	.13	.11		
4. Mental Health	3.86	0.47	1.32	.77	.82	−.13	
5. Latent Attractiveness	0.00	0.72	1.19	.98	.99	.12	.81

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. VIF = variance inflation factor. Latent attractiveness = the first principle component from PCA of attractiveness and perceived physical health ratings (see text).

fixed effect of actual mental health), as recommended by Bates et al. (2015). However, when reporting the significance of other beta estimates (as in Table 3), we used estimated degrees of freedom from the Satterthwaite approximation as calculated by the lmerTest package (Kuznetsova, Brockhoff, & Christensen, 2016). In practice, we compared estimates of significance from these two methods when possible, and found them to be close agreement. The analyses reported here were planned before the study, with the exception of non-linear masculinity effects, which occurred to us after initial analysis, and the causal mediation analyses, which arose from the review process. We had initially planned to run models using both attractiveness and physical health as fixed effects, but as described above, latent attractiveness proved a better choice, although we verify below this had no important effect on the outcome.

Model comparisons are given in Table 3. The crucial results relate to the effects of actual mental health. In our primary test model, the effect of actual mental health was positively weighted ($\beta = .142$, SE = .033), and the test model made significantly better predictions than the baseline model, $\chi^2(1) = 16.2$, $p < .00001$. As shown in the last row of the table, the predictions of the baseline model were significantly better than the null model, accounting for the two additional degrees of freedom, $\chi^2(2) = 60.8$, $p < .00001$. Latent attractiveness was positively weighted, $\beta = .34$, $t(75.6) = 9.54$, $p < .00001$, while masculinity was negatively weighted, $\beta = -.081$, $t(58.0) = 2.50$, $p = .015$. In summary, latent attractiveness and masculinity were visible cues to perceived mental health, but perceived mental health was explained above and beyond these ratings by the *actual* mental health reflected within the image.

We tested eight other pairs of baseline and test models, to be sure that the effect of actual on perceived mental health was not sensitive to modelling details, and in particular the use of latent attractiveness as opposed to either attractiveness, perceived physical health, or both. We found similar estimates ($\beta = .14$–$.17$) and significance values for the effect of actual mental health ($p < .00001$), whether using as baseline models: fixed effects of latent attractiveness, with and without masculinity; attractiveness only, with and without masculinity; perceived physical health only, with and without masculinity; and both attractiveness and perceived physical health, with and without masculinity.

Beyond this key result, this modelling approach allowed us to examine other issues and relationships.

3.3. Potential non-linear effects of masculinity

Our models assumed a linear effect of masculinity on mental health ratings. But there are good reasons to examine whether masculinity might be better modelled as a non-linear effect. Non-linear effects of masculinity are commonly observed on attractiveness, for example, hyper-masculinity can be regarded as more threatening and dangerous than attractive (Johnston, 2006). We tested this possibility by additionally including the squared mean of masculinity ratings to our models, to see whether there then remained any additional effect of actual mental health.

This new model added to the primary test model a squared-masculinity term as a fixed effect, and as a random slope by observer. We saw some evidence for an optimal level of masculinity, as masculinity was positively weighted and squared-masculinity was negatively weighted. However, the t-values for these effects were hardly different from unity, and the model was only marginally more successful than the primary test model after accounting for the additional degrees of freedom, $\chi^2(6) = 11.3$, $p = 0.08$. At present, the most we can say is that our data are consistent with the possibility of an optimal level of masculinity for perceived mental health. But our conclusions about actual mental health were unchanged. The "optimal masculinity" model was significantly worse if the fixed effect of actual mental health was removed, $\chi^2(1) = 13.7$, $p = .0002$. Further, the estimate for actual mental health ($\beta = .13$, SE = .034) was similar to our other results. That is, even allowing for non-linear effects of masculinity, the effect of

Table 3
Model comparison for the effects of actual on perceived mental health.

	Model			
	Baseline		Test	
Predictor	*beta*	*SE*	*beta*	*SE*
(Intercept)	−0.02	0.05	0.00	0.05
Latent attractiveness	0.34***	0.04	0.30***	0.03
Masculinity	−0.10*	0.03	−0.04	0.03
Actual mental health	–	–	0.14***	0.03
Model comparison	vs null model: $\chi^2(2) = 60.8$*** BIC = 10351		vs baseline model: $\chi^2(1) = 16.2$*** BIC = 10343	

Note. Comparison of primary models and their fixed effects, shown as standardised betas and standard error (SE). Models were identical other than inclusion of actual mental health as a fixed effect in the Test model. Model comparisons based on differences in log likelihood, BIC = Bayesian Information Criterion. p-values for fixed effects are based on Satterthwaite approximation, as described in text.

* $p < .05$.
*** $p < .0001$.

actual on perceived mental health remained essentially unchanged.

3.4. Similarities and differences between neuropsychiatric traits

We noted earlier that ANOVA on the mental health ratings suggested a Trait * Actual Health interaction, such that the effects of health were larger for schizotypy than for depression and AQ (Fig. 2). We took two approaches to better understanding this issue: a more complex single model allowing for trait interactions, and separate trait-specific models.

A "complex" model added the interaction of Trait * Actual Mental Health as fixed and random observer slope effects to the primary test model. This substantially increased the number of estimated model parameters (e.g., for each observer 8 random effects were estimated rather than 4 as in our primary test model), but did not produce a significant improvement over the primary test model, $\chi^2(30) = 18.5$, $p = .95$, and t-values for all Trait and Trait * Health estimates were below magnitude 1. The estimate for Actual Mental Health was similar to our primary test model ($\beta = .15$, $p = .005$). These results suggest that the apparent Trait * Actual Health interaction seen in our ANOVA was not due to different "amounts" of mental health information in the three sets of composites.

However, the approach of a single complex model has its drawbacks. Increasing the number of estimated parameters reduces the number of data points per estimate, and our power for detecting an effect. We wanted to be certain that the lack of significant interaction was not hiding qualitatively different patterns between traits. We therefore took another view by creating trait-specific models, to verify the effect of actual mental health status for all traits. For each trait (AQ, Depression, Schizotypy), a model was created exactly analogous to our primary test model. Fig. 3 summarizes the coefficient estimates in each case, and for comparison, from our primary analysis covering all three traits simultaneously. Confidence intervals are calculated from the degrees of freedom estimated by Satterthwaite approximation (Kuznetsova et al., 2016). The figure illustrates a general similarity between traits, and we walk through each fixed effect separately.

First, estimates for Actual Mental Health were similar and overlapping for all traits. The picture has some complications, as the 95% confidence intervals of the Actual Mental Health effect for Depression ($\beta = .074$, SE = .037, estimated $t(19.6) = 2.04$, $p = .055$) and Schizotypy ($\beta = .13$, SE = .066, estimated $t(19.1) = 2.02$, $p = .058$) touch zero, although not AQ ($\beta = .15$, SE = .061, estimated $t(21.0) = 2.41$, $p = .025$). However, the consistency of estimates is in agreement with the results from all our models, including the "complex" model above, which found a significant effect of Actual Mental Health but no Trait * Actual Mental Health interaction. We therefore believe it is a reasonable conclusion that the influence of Actual Mental Health was similar for all three traits.

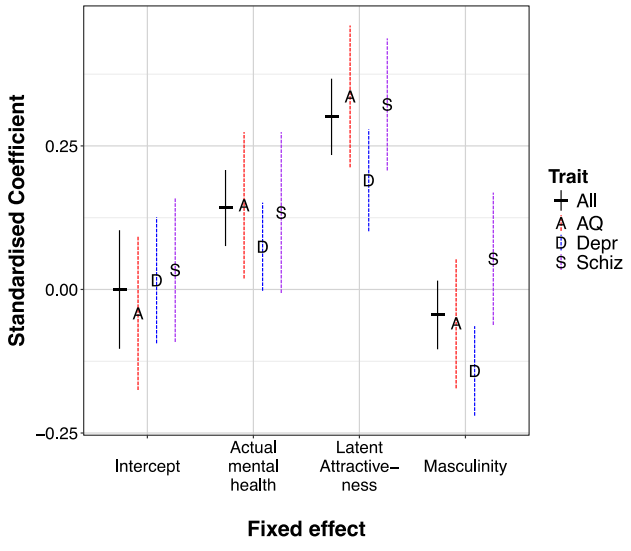

Fig. 3. Standardised fixed effects estimates, as found by modelling the data set as a whole, in our primary test model ("All"); and with models which are specific to each neuropsychiatric trait. Error bars represent estimated 95% confidence intervals.

Second, all models show a similar, positively-weighted, estimate for the effects of Latent Attractiveness. Attractiveness and perceived physical health were correlated with the perceived mental health of the target.

Finally, the estimates of Masculinity were low, little different from zero, and arguably did not show the same consistency across traits. We also checked whether masculinity might reflect sex bias in the frequency of the different disorders. That is, given that ASD and schizotypy are more frequent in men, and depression is more frequent in women, might facial dimorphism reflect these sex biases in frequency? The answer was no. Although high AQ was associated with masculinity (i.e., for this trait, masculinity was negatively weighted for mental health), high schizotypy was associated with reduced masculinity (even though it is more common in men), and depression with increased masculinity (even though it is more common in women).

3.5. Sources of accurate mental health information

With two sets of occluded stimuli, one showing only inner facial features, and the other only the eye region, we had some possibility of narrowing where within the face lay cues to mental health. The correlation between full-face and inner-face ratings was high for all judgements (correlating mean stimulus ratings: mental health, $r = .91$; attractiveness, $r = .90$; physical health, $r = .90$; masculinity, $r = .84$). Given this high correlation with full-face ratings, it is perhaps not surprising that repeating our primary analyses of baseline and test models on inner face ratings produced the same pattern of significant results as the full-face ratings. The test model, including actual mental health, was superior to the baseline model without, $\chi^2(1) = 8.52$, p = .0035; and both latent attractiveness ($\beta = .255$, SE = .038, estimated $t(87.5) = 6.83$, $p < .00001$) and actual mental health ($\beta = .115$, SE = .038, estimated $t(72.1) = 3.01$, $p = .0036$) were significant predictors.

However, the results from the eyes-only stimuli showed a different and revealing pattern. Ratings of mental health from the eyes-only were also significantly higher for high compared to low mental health composites, $t(52) = 3.0$, $p = .004$. However, this time the test model -- including actual mental health as a predictor – was *not* better than the baseline model without, $\chi^2(1) = .69$, $p = .41$. That is, the ratings of mental health from the eyes were well explained by the combination of latent attractiveness and masculinity ratings, and no additional cue specific to actual mental health was needed. Therefore, we suggest that the cues to mental health per se, evident in the full-face, are present in the inner features, but not exclusively within the eye region.

3.6. Causal mediation analysis

The analyses above demonstrate that actual mental health predicted perceptions of mental health, even when simultaneously considering the influence of attractiveness, perceived physical health, and masculinity. The mixed-effects modelling we used does not make or test assumptions about the causal relationships between these different variables. Of course, to the extent possible, it would be useful to better understand the potential causal

relationships among them. We explored this issue with causal mediation analyses (using the R packages "mediation", Tingley, Yamamoto, Hirose, Keele, & Imai, 2014; and "psych", Revelle, 2017). These analyses arose from issues raised during the review process, and were not planned at the time of the initial study. We wished to further understand the possible causal relationship between actual and perceived mental health, and specifically, to what extent the effect of actual on perceived mental health might be mediated by the other appearance variables we tested. Our modelling used the mean appearance ratings for each of the full-face composites, as well as the actual mental health value of those composites (i.e., $N = 54$). Actual mental health was the predictor, perceived mental health the outcome, and various combinations of the other appearance ratings were tested as mediators.

The most relevant model used latent attractiveness as the sole mediator. The average causal mediation effect (ACME = .21; 95% CI = [.05, .38], $p = .020$), describes how actual mental health influenced perceived mental health, through the influence of latent attractiveness. The average direct effect (ADE = .35; 95% CI = [.23, .47], $p < .0001$), is the effect of actual on perceived mental health through all other routes. Finally, the total influence of actual on perceived mental health reflects all routes, regardless of mechanism (total = .56, 95% CI = [.35, .75], $p < .0001$).

A similar analysis examining masculinity as the sole mediator, rather than latent attractiveness, did not find a mediating effect (ACME = −0.0226; 95% CI = [−0.1090, 0.04]), although the direct effect of actual on perceived mental health remained (ADE = .58; 95% CI = [.36, .80], $p < .0001$). In addition, a model with masculinity included as a second potential mediator, in addition to latent attractiveness, did not change the pattern of results from latent attractiveness as sole mediator (ACME = .25, 95% CI = [.09, .42]; ADE = .31, 95% CI = [.17, .45], $p < .0001$). We therefore conclude that masculinity is unlikely to mediate the effect of actual on perceived mental health.

Causal mediation makes some strong assumptions, including an absence of unobserved confounds between the mediator and outcome variables. For a model of latent attractiveness as mediator, this would mean unobserved confounds between latent attractiveness and perceived mental health. Sensitivity analysis to investigate the robustness of results to such unaccounted confounds is the preferred course (e.g., Cox, Kisbu-Sakarya, Miočević, & MacKinnon, 2014). In correlated residuals analysis (Imai, Keele, & Tingley, 2010), the sensitivity parameter ρ models the effect of unobserved confounds as the correlation of residuals from the mediator and outcome regressions underlying the causal mediation model (varying from −1.0 to 1.0). Sensitivity of mediation estimates to varying levels of this parameter can then be calculated. This analysis revealed that estimates for the mediated effect were not highly sensitive to unobserved confounding, and held for all $\rho < .71$, comparing favorably to other studies using this form of sensitivity analysis (Imai & Yamamoto, 2014).

A cautious interpretation of these results is that some, but not all, influence of actual on perceived mental health may be mediated by latent attractiveness. Although the mediating effect of latent attractiveness was relatively insensitive to unobserved confounds, it would be premature to rule out such confounds at this early stage. In any case, the robust direct effect of actual on perceived mental health was also consistent with the outcomes of our main analyses, which demonstrated that the effect of actual on perceived mental health was not fully explained by the other appearance variables we tested. We therefore take these insights into causality as complementary to our main analyses on the distinctive contribution of actual to perceived mental health.

4. Discussion

Observers could accurately estimate the mental health status of men in neutral facial composites: perceived mental health reflected actual mental health. Furthermore, this accuracy could not be fully explained by attractiveness, physical health, masculinity, or their combination. Instead, we found that actual mental health status had an additional influence on perceived mental health, not explained by these other, well-studied, appearance variables.

The significant effect of actual on perceived mental health was robust and did not much depend on details of the analyses. First, although ratings of attractiveness and perceived physical health were highly correlated, we found essentially the same effect of actual mental health regardless of whether we analyzed these ratings separately, together, or on the basis of "latent attractiveness" – the first component from their PCA. Second, allowing for nonlinear effects of facial masculinity did not change the estimate of actual mental health effects. Third, whether we analyzed each mental health condition separately, or all together, we found similar effects of actual mental health. The only analysis which did markedly change the actual mental health effect was when we restricted ratings to the region around the eyes. In this case, while mental health ratings were still higher for high than low composites, these ratings could be explained by attractiveness and masculinity, without additional effect of actual mental health. We suggested this is consistent with distinctive cues to actual mental health lying outside the eye region.

The significant effect of actual on perceived mental health was accompanied by an even larger effect of latent attractiveness (Fig. 3). This is perhaps not surprising, as there are many potential appearance cues to mental health which might be perceived as attractive and/or physically healthy. For example, a pleasant expression is attractive and might be perceived to indicate a state of mental well-being (Lau, 1982). Likewise, reddish skin coloration can increase perceived physical health (Stephen, Coetzee, & Perrett, 2011; Stephen, Law Smith, Stirrat, & Perrett, 2009), and this might lead observers to infer correspondingly increased mental health. A subtle cue in the composite images, perhaps similar to a micro-expression related to smiling (e.g., Ekman, 2003), could potentially drive attractiveness and mental health ratings. However, it is important to be clear that actual mental health accounted for variance in observer ratings which attractiveness and perceived physical health cues did not. Therefore, cues which are generally attractive or healthy looking cannot wholly explain the accuracy of mental health ratings. However, might they account for some of this accuracy? Perhaps the best way to address this question is through our causal mediation analysis, which suggested that latent attractiveness did mediate the effect of actual on perceived mental health. Future research to identify valid appearance cues to mental health should therefore consider possible cues which are both correlated and uncorrelated with attractiveness.

Our stimuli were created from three reasonably diverse neuropsychiatric inventories – based on depressive symptoms, autistic quotient, and schizotypy – yet we were struck more by the similarity among the trait-specific models than their differences (Fig. 3). Given this consistency, is it possible that observers were picking up on some commonality across diverse mental health conditions? We suggest there are some plausible and interesting common bases for judgement. The first two are essentially related to comorbidity among these disorders. One possibility would be trait Neuroticism. As we reviewed earlier, trait Neuroticism is associated with numerous mental health disorders (e.g., Kotov et al., 2010), including those we tested, and it is identifiable in similar facial appearance paradigms (e.g., Kramer & Ward, 2010; Jones

et al., 2012). An interesting aspect of this previous work is that, reminiscent of our findings here, discrimination accuracy for neuroticism composites was not driven by attractiveness (Kramer & Ward, 2010). An alternative is suggested by recent findings showing that identification of trait Neuroticism may be driven by the same serotonergic mechanisms implicated in fear recognition (Ward, Sreenivas, Read, Saunders, & Rogers, 2017), and so subtle expressive signals relating to fear might act as cues to Neuroticism. Another possible commonality across disorders would be the p factor (or General Psychopathology factor), which arises from factor-analytic approaches to the classification of psychopathologies, and refers to a generalized risk to all sorts of mental health disorder (Caspi et al., 2014). A generalized risk factor is consistent with other quantitative approaches to psychopathology, such as Kotov et al. (2017), which find positive correlations among different taxonomic spectra. The p factor is itself positive correlated with trait Neuroticism, and negatively correlated with Agreeableness and Conscientiousness (Caspi et al., 2014). Observers could therefore potentially pick up either on p, personality traits correlated with p, or combinations of these general factors. Of course, using p to explain our results requires that AQ be sensitive to p, and it is important to note that AQ and ASD has not yet been incorporated into quantitative psychopathology models, as far as we are aware. However, ASD is often found to be comorbid with other disorders, such as schizotypy (e.g., Barneveld et al., 2011), and this might support the relevance of a general risk factor to AQ.

A final possibility we consider is that accuracy may not be for mental health per se, but for an unrealistic self-enhancing bias. Kwan, Kenny, John, Bond, and Robins (2004) distinguished positive report about oneself from a true self-enhancing bias, by comparing measures of self-report (target reporting about themselves), self-other report (target's report about others), and others' report (what others report about the target). Using a refinement of this method, Leising, Locke, Kurzius, and Zimmermann (2016) found that in a wide range of self-report tasks, including intelligence scores and multiple personality inventories, self-enhancing bias was correlated with the social desirability of the question being asked. This finding raises an interesting question with regard to our own targets: what if the high "mental health" composites were not really high in mental health, but high in self-enhancing bias for desirable traits? That is, what if the high mental health group consisted disproportionately of men who tried to appear socially desirable, both in their questionnaires and their images? "Accurate" discrimination might then be the result of observers misattributing mental health to traits like self-esteem, or general positivity. Visual identification of self-esteem is at least plausible: For example, targets with high self-esteem are visually attractive (Mathes & Kahn, 1975; Zeigler-Hill & Besser, 2014), perhaps because attractive people may be more like to internalize socially desirable traits (Adams, 1977). Furthermore, grandiose self-esteem in the form of narcissistic personality can be detected in facial composites (Holtzman, 2011), and narcissism is itself associated with attractiveness (Dufner, Rauthmann, Czarna, & Denissen, 2013; Holtzman & Strube, 2010). Note that for a self-enhancement account of our findings to follow, enhancement must act similarly on both scores and images, and this may not hold for all potential mechanisms of self-enhancement. For example, self-enhancing bias can arise through a lack of insight into one's true states (e.g., Kruger & Dunning, 1999). While it is easy to see how lack of insight might lead to inflated mental health scores, it is less clear how lack of insight might also inflate observer ratings of appearance. If only self-ratings (or only appearance) were affected by self-enhancement, then the correlation we find between self-reported mental health and appearance would be lost. In any case, with our present dataset, we are unfortunately not able to separate self-enhancement from accurate report, because we do not have the comparison measures used by Kwan et al. and Leising et al., most importantly, other's report on the targets. However, this would be an interesting possibility for future research.

There are multiple limitations on generality at issue. First, we recognize that "mental health" covers a large range of conditions, and we have left a lot of uncharted space. For example, in the hierarchical taxonomy of Kotov et al. (2017), we have not touched on the externalizing spectra (such as alcohol or drug dependence). Broadening this space is an interesting possibility, particularly given that we may be seeing a common factor underlying mental health accuracy in the disorders we have examined. Second, our stimulus images were composite faces, not individual faces. As we discussed earlier, our reasons for using composite faces were conceptual, as we are interested in whether there are systematic cues to mental health, and our use of composites means we are more likely to see the statistical regularities of appearance correlated to mental health. However, the flip side is that we are unable to see the influence of idiosyncratic differences, such as fluctuating asymmetries: if high mental health were associated with low fluctuating asymmetry, we would not see it. Third, as we discussed in our introduction, we are measuring perceptions from very thin slices, and have done what we reasonably can to remove controllable cues. While we have found that even such minimal cues can be valid and informative to observers, this does raise interesting potential issues about whether in the wild, controllable cues might reveal or conceal mental health status. In at least some domains, including social networking sites, controllable cues can be revealing (Back et al., 2010). Fourth, although the three mental health traits we tested showed similar influences of actual on perceived mental health, as noted in our Stimulus Creation section, the sample used to make the AQ composites was not as extreme or reliable as the schizotypy and depression composites. If anything, it seems plausible that a more extreme and more homogenous sample might produce larger, rather than smaller, effects on AQ judgement, but at present this is unknown. Finally, it is worth noting that each judgement (mental health, attractiveness, physical health, and masculinity) was made by different groups of observers. This was a deliberate design choice, to keep the rating exercise simple, and to avoid possible contamination of one judgement on others. However, a clear limitation with this approach is that we are not modelling the relations between judgements happening within observers. For example, observers with different ratings of attractiveness for the same target might rate that target correspondingly differently for mental health. For this kind of within-observer modelling, another approach would be needed, in which each observer gave ratings for all judgements.

Facial cues to behavior can be considered in the context an adaptive signal system (e.g., Kramer, King, & Ward, 2012), which are evolutionary stable only if both the signal sender and receiver gain net benefit. Being able to read mental health status seems clearly beneficial for the receiver, but how might it be beneficial for the sender to display mental illness to others? We offer a speculative although falsifiable response. We might expect a general advantage to senders in signaling their true trait levels, that is, in being predictable, even when the trait being signaled is not socially desirable. This is because sending a misleading visual signal is effectively a form of deceit, and might be correspondingly punished if detected. If this reasoning is correct, the accuracy of signals sent might reflect a trade-off, from the sender's perspective, of the benefits and costs for exaggerated socially desirability. That said, such a system might be adapted for traits within "normal" operating limits, in the same way that personality variation within a wide range could be beneficial depending on context, but very extreme variation would not be (Nettle, 2005). For people whose traits fall

well outside a mentally healthy range, valid signals to mental health status might be maladaptive, by evoking highly unfavorable observer responses.

Unfavorable observer response brings us to our final point. As we have argued elsewhere (Scott et al., 2015), the fact that mental health status can be cued to observers in the absence of behavior, raises the possibility of a vicious cycle in which those at greatest risk of mental health problems may be continuously and involuntarily broadcasting a message which is producing negative reactions in observers (Coyne, 1976). For example, consider two individuals A and B, who are equally depressed at the start of the year, but the appearance of person A happens to be more easily read as depressed than the appearance of person B. One could hypothesize that a year of consistently more negative reactions from observers could put person A at greater risk of social isolation, reduced social support, and further depressive symptoms. This kind of scenario raises important issues regarding the nature of visual cues to health status and the responses to those cues. Our current findings cannot establish the significance of this risk: on the one hand, we are using composites which might amplify the "signal" of mental health from the "noise" of other facial cues; on the other hand, we are using impoverished stimuli which may not reflect the totality of appearance cues available in even minimal face-to-face interaction. At present then, our view is that there is a relatively unappreciated and unexplored potential risk factor of appearance for the development of mental health disorders.

5. Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2018.04.007.

References

Adams, G. R. (1977). Physical attractiveness, personality, and social reactions to peer pressure. *Journal of Psychology: Interdisciplinary and Applied, 96*(2), 287–296. https://doi.org/10.1080/00223980.1977.9915911.

Aleman, A., Kahn, R. S., & Selten, J.-P. (2003). Sex differences in the risk of schizophrenia. *Archives of General Psychiatry, 60*(6), 565. https://doi.org/10.1001/archpsyc.60.6.565.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274. https://doi.org/10.1037/0033-2909.111.2.256.

Axelsson, J., Sundelin, T., Ingre, M., Van Someren, E. J., Olsson, A., & Lekander, M. (2010). Beauty sleep: Experimental study on the perceived health and attractiveness of sleep deprived people. *BMJ, 341*, c6614.

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. doi: 10.1177%2F0956797609360756.

Barneveld, P. S., Pieterse, J., de Sonneville, L., van Rijn, S., Lahuis, B., van Engeland, H., & Swaab, H. (2011). Overlap of autistic and schizotypal traits in adolescents with Autism Spectrum Disorders. *Schizophrenia Research, 126*(1–3), 231–236. https://doi.org/10.1016/j.schres.2010.09.004.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Sciences, 6*(6), 248–254. https://doi.org/10.1016/S1364-6613(02)01904-6.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17. https://doi.org/10.1023/A:1005653411471.

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

Borkenau, P., & Liebler, A. (1992). Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology, 62*(4), 645–657.

Burns, G. L., & Farina, A. (1992). The role of physical attractiveness in adjustment. *Genetic, Social & General Psychology Monographs, 118*(2), 157.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., ... Moffitt, T. E. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science, 2*(2), 119–137. https://doi.org/10.1177/2167702613497473.

Colton, C. W., & Manderscheid, R. W. (2006). Congruencies in increased mortality rates, years of potential life lost, and causes of death among public mental health clients in eight states. *Preventing Chronic Disease, 3*(2), A42. doi: A42 [pii].

Cox, M. G., Kisbu-Sakarya, Y., Miočević, M., & MacKinnon, D. P. (2014). Sensitivity plots for confounder bias in the single mediator model. *Evaluation Review, 37*(5), 405–431. https://doi.org/10.1177/0193841X14524576.

Coyne, J. C. (1976). Depression and the response of others. *Journal of Abnormal Psychology, 85*(2), 186–193. https://doi.org/10.1037/0021-843X.85.2.186.

Crespi, B., & Badcock, C. (2008). Psychosis and autism as diametrical disorders of the social brain. *Behavioral and Brain Sciences, 31*(3), 241–261. https://doi.org/10.1017/S0140525X08004214.

Daros, A. R., Ruocco, A. C., & Rule, N. O. (2016). Identifying mental disorder from the faces of women with borderline personality disorder. *Journal of Nonverbal Behavior, 40*(4), 255–281. https://doi.org/10.1007/s10919-016-0237-9.

Datta Gupta, N., Etcoff, N. L., & Jaeger, M. M. (2016). Beauty in mind: The effects of physical attractiveness on psychological well-being and distress. *Journal of Happiness Studies, 17*(3), 1313–1325. https://doi.org/10.1007/s10902-015-9644-6.

DeBruine, L. M., Jones, B. C., Little, A. C., Boothroyd, L. G., Perrett, D. I., Penton-Voak, I. S., ... Tiddeman, B. P. (2006). Correlated preferences for facial masculinity and ideal or actual partner's masculinity. *Proceedings of the Royal Society B: Biological Sciences, 273*(1592), 1355–1360. https://doi.org/10.1098/rspb.2005.3445.

Dufner, M., Rauthmann, J. F., Czarna, A. Z., & Denissen, J. J. A. (2013). Are Narcissists sexy? Zeroing in on the effect of Narcissism on short-term mate appeal. *Personality and Social Psychology Bulletin, 39*(7), 870–882. https://doi.org/10.1177/0146167213483580.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*(1), 109–128. https://doi.org/10.1037/0033-2909.110.1.109.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences, 1000*(1), 205–221. https://doi.org/10.1196/annals.1280.010.

Farina, A., Fischer, E. H., Sherman, S., Smith, W. T., Groh, T., & Mermin, P. (1977). Physical attractiveness and mental illness. *Journal of Abnormal Psychology, 86*(5), 510–517. https://doi.org/10.1037/0021-843X.86.5.510.

Fölstad, I., & Karter, A. J. (1992). Parasites, bright males, and the immunocompetence handicap. *The American Naturalist, 139*(3), 603–622. https://doi.org/10.1086/285346.

Foo, Y. Z., Simmons, L. W., & Rhodes, G. (2017). Predictors of facial attractiveness and health in humans. *Scientific Reports, 7*, 39731. https://doi.org/10.1038/srep39731.

Gangestad, S. W., & Buss, D. M. (1993). Pathogen prevalence and human mate preferences. *Ethology and Sociobiology, 14*(2), 89–96. https://doi.org/10.1016/0162-3095(93)90009-7.

Hays, R. D., Bjorner, J. B., Revicki, D. A., Spritzer, K. L., & Cella, D. (2009). Development of physical and mental health summary scores from the patient-reported outcomes measurement information system (PROMIS) global items. *Quality of Life Research, 18*(7), 873–880. https://doi.org/10.1007/s11136-009-9496-9.

Holtzman, N. S. (2011). Facing a psychopath: Detecting the dark triad from emotionally-neutral faces, using prototypes from the Personality Faceaurus. *Journal of Research in Personality, 45*(6), 648–654. https://doi.org/10.1016/j.jrp.2011.09.002.

Holtzman, N. S., & Strube, M. J. (2010). Narcissism and attractiveness. *Journal of Research in Personality, 44*(1), 133–136. https://doi.org/10.1016/j.jrp.2009.10.004.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods, 15*(4), 309–334. https://doi.org/10.1037/a0020761.

Imai, K., & Yamamoto, T. (2014). Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis, 21*(12), 141–171. https://doi.org/10.1093/pan/mps040.

Johnston, V. S. (2006). Mate choice decisions: The role of facial beauty. *Trends in Cognitive Sciences, 10*(1), 9–13. https://doi.org/10.1016/j.tics.2005.11.003.

Jones, A. L., Kramer, R. S. S., & Ward, R. (2012). Signals of personality and health: The contributions of facial shape, skin texture, and viewing angle. *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1353–1361. https://doi.org/10.1037/a0027078.

Kessler, R. C. (2003). Epidemiology of women and depression. *Journal of Affective Disorders, 74*(1), 5–13. https://doi.org/10.1016/S0165-0327(02)00426-3.

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "Big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*(5), 768–821. https://doi.org/10.1037/a0020327.

Kotov, R., Waszczuk, M. A., Krueger, R. F., Forbes, M. K., Watson, D., Clark, L. A., ... Zimmerman, M. (2017). The hierarchical taxonomy of psychopathology (HiTOP): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology, 126*(4), 454–477. https://doi.org/10.1037/abn0000258.

Kramer, R. S. S., King, J. E., & Ward, R. (2012). Cues to personality and health in the facial appearance of chimpanzees (Pan troglodytes). *Evolutionary Psychology, 10*(2), 320–337. doi: 147470491201000210.

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *The Quarterly Journal of Experimental Psychology, 63*(11), 2273–2287. https://doi.org/10.1080/17470211003770912.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 121–1134.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in Linear Mixed Effects Models. R Package Version 2.0-33. Comprehensive R Archive Network (CRAN). Retrieved from https://cran.r-project.org/package=lmerTest.

Kwan, V. S. Y., Kenny, D. A., John, O. P., Bond, M. H., & Robins, R. W. (2004). Reconceptualizing individual differences in self-enhancement bias: An interpersonal approach. *Psychological Review, 111*(1), 94–110. https://doi.org/10.1037/0033-295X.111.1.94.

Lau, S. (1982). The effect of smiling on person perception. *The Journal of Social Psychology, 117*(1), 63–67. https://doi.org/10.1080/00224545.1982.9713408.

Leising, D., Locke, K. D., Kurzius, E., & Zimmermann, J. (2016). Quantifying the association of self-enhancement bias with self-ratings of personality and life satisfaction. *Assessment, 23*(5), 588–602. https://doi.org/10.1177/1073191115590852.

Little, A. C., Jones, B. C., & DeBruine, L. M. (2011). Facial attractiveness: Evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences, 366*(1571), 1638–1659. https://doi.org/10.1098/rstb.2010.0404.

Little, A. C., & Perrett, D. I. (2007). Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology, 98*(1), 111–126. https://doi.org/10.1348/000712606X109648.

Martin, P. J., Friedmeyer, M. H., & Moore, J. E. (1977). Pretty patient-healthy patient? A study of physical attractiveness and psychopathology. *Journal of Clinical Psychology, 33*(4), 990–994.

Mathes, E. W., & Kahn, A. (1975). Physical attractiveness, happiness, neuroticism, and self-esteem. *Journal of Psychology: Interdisciplinary and Applied, 90*(1), 27–30. https://doi.org/10.1080/00223980.1975.9923921.

McGovern, R. J., Neale, M. C., & Kendler, K. S. (1996). The independence of physical attractiveness and symptoms of depression in a female twin population. *Journal of Psychology: Interdisciplinary and Applied, 130*(2), 209–219. https://doi.org/10.1080/00223980.1996.9915002.

Miller, A. G. (1970). Role of physical attractiveness in impression formation. *Psychonomic Science, 19*(4), 241–243. https://doi.org/10.3758/BF03328797.

Moran, P., Stewart, R., Brugha, T., Bebbington, P., Bhugra, D., Jenkins, R., & Coid, J. W. (2007). Personality disorder and cardiovascular disease: Results from a national household survey. *Journal of Clinical Psychiatry, 68*(1), 69–74.

Napoleon, T., Chassin, L., & Young, R. D. (1980). A replication and extension of "physical attractiveness and mental illness". *Journal of Abnormal Psychology, 89*(2), 250–253. https://doi.org/10.1037/0021-843X.89.2.250.

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. https://doi.org/10.1177/0146167209346309.

Nettle, D. (2005). An evolutionary approach to the extraversion continuum. *Evolution and Human Behavior, 26*(4), 363–373. https://doi.org/10.1016/j.evolhumbehav.2004.12.004.

O'Grady, K. E. (1982). Sex, physical attractiveness, and perceived risk for mental illness. *Journal of Personality and Social Psychology, 43*(5), 1064–1671. https://doi.org/10.1037/0022-3514.43.5.1064.

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., ... Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature, 394*(6696), 884–887. https://doi.org/10.1038/29772.

Raine, A. (1991). The SPQ: A scale for the assessment of schizotypal personality based on DSM-III-R criteria. *Schizophrenia Bulletin, 17*(4), 555–564. https://doi.org/10.1093/schbul/17.4.555.

Rhodes, G., Yoshikawa, S., Palermo, R., Simmonst, L. W., Peters, M., Lee, K., ... Crawford, J. R. (2007). Perceived health contributes to the attractiveness of facial symmetry, averageness, and sexual dimorphism. *Perception, 36*(8), 1244–1252. https://doi.org/10.1068/p5712.

Rush, A. J., Gullion, C. M., Basco, M. R., Jarrett, R. B., & Trivedi, M. H. (1996). The Inventory of Depressive Symptomatology (IDS): Psychometric properties. *Psychological Medicine, 26*(3), 477–486. https://doi.org/10.1017/S0033291700035558.

Saarni, S. I., Viertio, S., Perala, J., Koskinen, S., Lonnqvist, J., & Suvisaari, J. (2010). Quality of life of people with schizophrenia, bipolar disorder and other psychotic disorders. *British Journal of Psychiatry, 197*(5), 386–394. https://doi.org/10.1192/bjp.bp.109.076489.

Scott, I. M. L., Clark, A. P., Boothroyd, L. G., & Penton-Voak, I. S. (2013). Do men's faces really signal heritable immunocompetence? *Behavioral Ecology, 24*(3), 579–589. https://doi.org/10.1093/beheco/ars092.

Scott, N. J., Jones, A. L., Kramer, R. S. S., & Ward, R. (2015). Facial dimorphism in Autistic Quotient scores. *Clinical Psychological Science, 3*(2), 230–241. https://doi.org/10.1177/2167702614534238.

Scott, N. J., Kramer, R. S. S., Jones, A. L., & Ward, R. (2013). Facial cues to depressive symptoms and their associated personality attributions. *Psychiatry Research, 208*(1), 47–53. https://doi.org/10.1016/j.psychres.2013.02.027.

Stephen, I. D., Coetzee, V., & Perrett, D. I. (2011). Carotenoid and melanin pigment coloration affect perceived human health. *Evolution and Human Behavior, 32*(3), 216–227. https://doi.org/10.1016/j.evolhumbehav.2010.09.003.

Stephen, I. D., Law Smith, M. J., Stirrat, M. R., & Perrett, D. I. (2009). Facial skin coloration affects perceived health of human faces. *International Journal of Primatology, 30*(6), 845–857. https://doi.org/10.1007/s10764-009-9380-z.

Strine, T. W., Chapman, D. P., Balluz, L. S., Moriarty, D. G., & Mokdad, A. H. (2008). The associations between life satisfaction and health-related quality of life, chronic illness, and health behaviors among U.S. community-dwelling adults. *Journal of Community Health, 33*(1), 40–50. https://doi.org/10.1007/s10900-007-9066-4.

Tan, D. W., Gilani, S. Z., Maybery, M. T., Mian, A., Hunt, A., Walters, M., & Whitehouse, A. J. O. (2017). Hypermasculinised facial morphology in boys and girls with Autism Spectrum Disorder and its association with symptomatology. *Scientific Reports, 7*(1), 9348. https://doi.org/10.1038/s41598-017-09939-y.

Tanner, J. M. (1990). *Foetus into man: Physical growth from conception to maturity*. Cambridge, MA: Harvard University Press.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R Package for causal mediation analysis Retrieved from. *Journal of Statistical Software, 59*(5), 1–38 http://www.jstatsoft.org/v59/i05/.

Ward, R., Sreenivas, S., Read, J., Saunders, K. E. A., & Rogers, R. D. (2017). The role of serotonin in personality inference: Tryptophan depletion impairs the identification of neuroticism in the face. *Psychopharmacology, 234*(14), 2139–2147. https://doi.org/10.1007/s00213-017-4619-4.

Zeigler-Hill, V., & Besser, A. (2014). Self-esteem and evaluations of targets with ostensibly different levels of self-worth. *Self and Identity, 13*(2), 146–161. https://doi.org/10.1080/15298868.2013.770194.