Judging a Book By Its Cover: Are First Impressions Accurate?

Tess Adams

Advisor: Dr. Matthew C. Keller

Department of Psychology and Neuroscience

University of Colorado Boulder

Undergraduate Honors Thesis

Committee Members:

Dr. Matthew C. Keller: Department of Psychology and Neuroscience

Dr. Vijay Mittal: Department of Psychology and Neuroscience

Dr. Douglas Duncan: Department of Astrophysical and Planetary Sciences

Abstract

First impressions are integral to human interactions, and philosophers and scientists have long discussed the idea that the face is a window into our internal traits. We make judgments of character based on appearance daily, consciously and subconsciously. Explanations for this phenomenon include the attractiveness stereotype, self-fulfilling prophecies, or "good genes" hypotheses from evolutionary psychology, but there have been mixed findings regarding the accuracy of such judgments. The current study investigates correlations between three subjectively judged "internal" traits and objective measures of Intelligence, Extraversion, and Neuroticism on 1600 subjects. We regressed these objective measures on their respective subjective ratings and controlled for several potential mediating factors. We found that Intelligence can be judged accurately even when controlling for potential mediators including attractiveness, SES, and perceived grooming, and ethnicity. Extraversion can also be judged accurately, but appears to be mediated by attractiveness, grooming, smiling, and socioeconomic status. Judgments of Neuroticism, on the other hand, could not be predicted by subjective ratings. This suggests that we can pick up on valid cues towards a person's internal traits without seeing any of their interactions.

Judging a Book By Its Cover: Are First Impressions Accurate?

People make subjective judgments about others on a regular basis, consciously and subconsciously. But how much information can actually be gleaned from a glance at a face? The idea that internal traits can be displayed externally dates back at least to Aristotle, who states, “It is possible to infer character from features” (*Prior Analytics,* 2.27). In the late 1700’s, Johann Kaspar Lavater, a Swiss pastor, published a series of essays on this ideal – known as physiognomy – which gained a great following into the 19th century. The shape of the nose, the set of the jaw, the width of the forehead – all were key to understanding whether a person would be well-suited to a particular occupation because those physical traits were directly linked to intelligence, or kindness, or perseverance. Such judgments are based on stable traits and facial characteristics, not on fleeting expressions, emotions, or interactions.

In Darwin’s time, physiognomy was accepted as fact, and he refers to it throughout his journal in relation to native peoples he met on his travels. Darwin ran into trouble himself though, when applying to be the “adventurous young man” accompanying Captain Fitz-Roy on the HMS Beagle. “Afterwards… I heard that I had run a very narrow risk of being rejected, on account of the shape of my nose! He was an ardent disciple of Lavater, and was convinced that he could judge of a man’s character by the outline of his features; and he doubted whether anyone with my nose could possess sufficient energy and determination for the voyage. But I think he was afterwards well satisfied that my nose had spoken falsely.” (Darwin, *Autobiography,* 72).

Physiognomy fell out of favor in the late 19th Century due to its association with Phrenology – the notion that one's personality could be found by reading the bumps on his or her skull, which represented certain areas of the brain being larger or smaller. Upon opening the skull, scientists discovered that the inside of the skull is smooth – so bumps could not possibly represent areas of the brain – and thus phrenology was discredited, and it's cousin physiognomy along with it.

More recent scientific studies have once again begun looking into whether subjective impressions based on facial characteristics have any validity. People do form global and specific trait impressions automatically based on facial structure (Hassin & Trope, 2000). A study by Willis and Todorov (2006) found that these first impressions are made after a mere tenth of a second exposure to a face. 42 raters answered a questionnaire on each of 70 faces (presented as standardized photographs with neutral expressions). The authors discovered that judgments made after a 100-ms exposure to a face did not differ significantly from those made with no time constraint. Their result held true for attractiveness, likeability, trustworthiness, competence, and aggressiveness (Willis, 2006). The results from Hassin and Willis suggest that we infer personality traits from facial appearance quickly and uncontrollably, which constantly affects our social interactions whether we are aware of it or not.

Social psychological studies have drawn attention to the attractiveness stereotype – a phenomenon known as the "Halo Effect". The halo effect posits that we automatically assign positive traits to more attractive people; if a person is attractive, we also deem them more likely to be nice, intelligent, successful, and outgoing. Dion et al. point out that physical appearance is the "personal characteristic most obvious and accessible to others in social interaction"(1971). The question of the self-fulfilling prophecy then arises – do personality traits affect or reflect appearance, or does appearance mold personality? The authors found that attractive people were assumed to be more likely to lead happy successful lives, in all realms from the dating world to the professional world (Dion et al, 1971).

In 2002, Zebrowitz et al. looked into the accuracy of estimating IQ from facial photos, taking into account several past studies that had mixed results. They performed a meta-analysis on seven perceived intelligence/ measured intelligence studies from the first half of the twentieth century. Raters rated the intelligence of subjects from facial photographs. The average "accuracy" (the correlation between measured intelligence and perceived intelligence) was 0.3, but ranged between 0.07 and 0.7 depending on the study. (Zebrowitz, 2002). The studies took place from 1918 to 2001, with the number of raters ranging from 10 to 1,530, and number of targets ranging from 10 to 150. The varied characteristics of these studies may account for the large range in results.

The authors cited the halo effect as a possible explanation for the successful judgments, but went on to note that evolutionary and social expectancy theories may predict that attractiveness is associated with actual intelligence (Zebrowitz, 2002). The evolutionary theory would suggest that attractiveness is a way of broadcasting "good genes", including higher intelligence – the offspring of more intelligent mates may be more likely to survive, so traits that display intelligence would be seen as attractive. Zebrowitz also discussed the potential mediating effects of grooming, nutrition, and healthcare in the relationship between attractiveness and intelligence in the context of socioeconomic status. Socioeconomic status is a good predictor of IQ (citation), and Zebrowitz found that it is also positively associated with perceptions of attractiveness and intelligence (Zebrowitz 2002). They suggest that raters used attractiveness to determine SES, and both attractiveness and SES to determine intelligence. The authors conclude that people can successfully judge intelligence, and postulate that it is due to the "valid cue" of attractiveness.

A recent study used composite images to assess accuracy in personality attribution from looking at faces (Little and Perrett 2007). Many previous studies regarding personality attribution have involved in-person interactions, with or without verbal communications (rather than photographs only). The ability to accurately assign personality characteristics without in-person interactions is known as "zero acquaintance". Little and Perrett stated that this accuracy is found cross-culturally and regardless of medium – photograph, video, or observations. They formed composite images of people who scored either high or low on self-report measures of personality because the authors thought common characteristics would be highlighted, and non-shared characteristics would disappear by being averaged out (Little and Perrett, 2007). The photographs used to create the composites were taken with strict criteria: photos included only the face against a constant background, and participants posed with neutral expressions, no glasses, hair pulled back, and clean-shaven. The authors found significant agreement between subjective and self-report personality scores for agreeableness, conscientiousness, and extraversion. These results suggest that faces hold accurate cues to personality.

Penton-Voak et al. (2006) also studied personality judgments from natural and composite faces. They found that the perceptions of traits formed from composite faces (based on scoring either high or low on a self-report personality test) were more accurate than those formed from an individual's face. Perceptions of extraversion and agreeableness had the strongest relationships with the self-report test measures, and emotional sensitivity had a relationship only in males. There was a high degree of consensus in ratings, but the authors stated that the overall validity of the judgments were "unclear and somewhat controversial" (Penton-Voak, 2006).

The current study seeks to test the validity of the link between perceived internal traits and their objectively measured counterparts. We investigated three traits: Intelligence, Extraversion, and Neuroticism. Each of these has a subjective and objective measure, the subjective measures being perceived intelligence, perceived extraversion, and perceived emotional sensitivity, and the objective measures being IQ scores and self-report personality test scores. First, we tested to see if the results of subjective impressions correlate with objective measures of the same trait. We then examined whether any relationships remained after controlling for potential mediating variables such as attractiveness, sex, grooming, ethnicity, and socioeconomic status. A positive correlation remaining after regression would suggest that there is merit in subjective judgments of traits above and beyond information gleaned from the potential mediating variables.

Method

Participants

The samples for this study were drawn from two twin databases, with 1599 total subjects. The larger set of twins (n= 1357) is from the Genetic Epidemiology department at the Queensland Institute of Medical Research (QIMR) in Brisbane, Australia. The other set is from the Longitudinal Twin Study (LTS) at the Institute for Behavioral Genetics in Boulder, Colorado (n=242).

The gender ratio within both sets is about equal, with 54.7% female and 45.3% male twins. The twin's ages range from 15 to 23 years old at the time of the photograph. Objective data for all twins was collected prior to the current study by the researchers at QIMR and LTS. To be included in our analyses, each subject needed a photograph and data regarding IQ, personality test scores, age, height, weight, and sex. Any twin sets without a full complement of data were excluded from the study.

Photographs

Photographs for LTS twins were cropped from the photographs previously obtained for the data set. Subjects were taken into a photo room and asked to remove their shoes, glasses, jackets, and other distracting apparel. Four photographs were taken against a one-inch grid: two full body and two with head and shoulders only. Participants were asked to maintain a neutral expression. Finished photos were 29.5 KB in JPG format, and cropped to include face and hair only.

Photographs of the QIMR subjects were not as tightly controlled because photographs were intended for identification purposes rather than for subjective ratings. Participants were allowed jewelry, makeup, jackets, headbands, glasses, etc. The shots were taken from the shoulder up.

In both sets we excluded photographs in which the subjects were blinking, or turning their heads. We tilted to upright any photographs in which the subjects were tilting their heads in Adobe Photoshop in order to maintain continuity between photographs for the raters. All photos were cropped by research assistants to include face only, from just below the chin to just above the hair.

Rating Procedure

Ratings for each subjective trait were carried out in the same way. A computer program displayed photos as a slide show with 50 subjects at a time. Each group was gender-specific, and groups alternated between male and female. The first slide of each group displayed instructions "In a moment, you are going to rate the following group of faces on (Trait). But first you will see a slideshow of all the faces. Use this time to get a sense of the range and variation among the faces for the trait of (Trait)." In order to obtain a standard distribution of the trait within each group, raters viewed each face for 2 seconds in a slideshow without rating. Raters were instructed to make distribution of scores among each set of fifty approximately uniform. After the slideshow, a screen with the definition of the trait was shown prior to rating to remind research assistants of what to focus on when assigning subjective ratings to faces. An example slide (shown with composite face, not one of our subjects) is shown in Figure 1. Results from each rater's subjective impressions were averaged, and the mean was used in the correlation against the actual score.

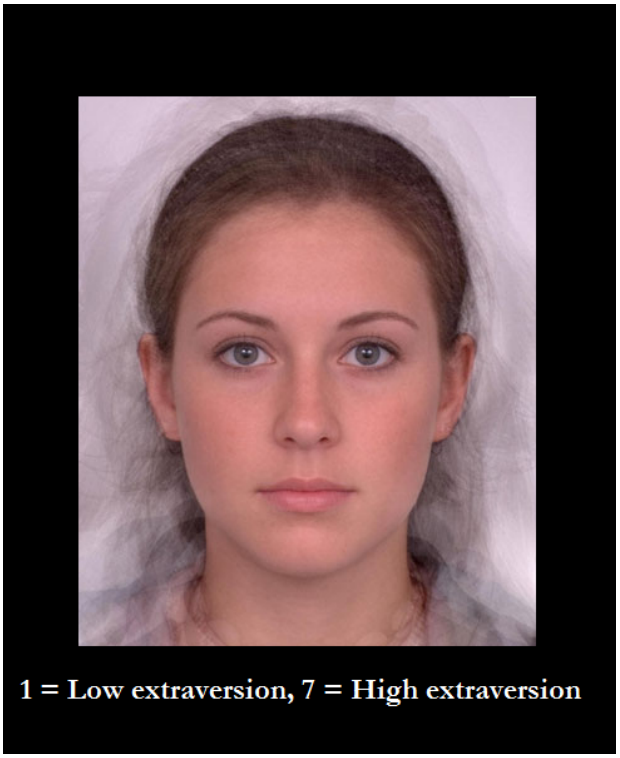
Cronbach's α was used to measure inter-rater reliability, or how consistent raters were in their ratings of each subject. It is defined as α = k $\bar{c}$ / v+ (k-1)c , where c is the average of the unique ratings covariances, v is the average of the unique variances and k is the number of raters. See table 1 for Cronbach's α for each subjectively measured trait.

$$\alpha = \frac{k\bar{c}}{\bar{v}+(k-1)\bar{c}}$$

Table 1 – Inter-rater Reliability

Trait	Cronbach's α	Number of Raters	Average Correlation between raters
Intelligence	0.60	7	r = 0.18
Extraversion	0.90	11	r = 0.47
Neuroticism	0.57	10	r = 0.13
Attractiveness	0.87	8	r = 0.45
Grooming	0.70	2	r = 0.54
Smiling	0.90	2	r = 0.82
Acne	0.77	2	r = 0.62

Figure 1



(Composite face obtained from Google Images)

Perceived Intelligence

Perceived intelligence ratings were gathered from four female and three male raters. Raters were undergraduate research assistants from the University of Colorado at Boulder. Raters were given the following instruction: “Rate this face’s intelligence on a scale from 1-7, 7 being the most intelligent”.

IQ

IQ scores were obtained from the existing QIMR and LTS twin data sets. Scores are from the Weschler Adult Intelligence Scale (WAIS).

Perceived Extraversion

Nine female and two male raters rated each face for extraversion. The instructions read: “Rate this face’s extraversion on a scale of 1-7, 7 being most extroverted.” In order to give a working definition of extraversion that was consistent across raters, raters were given the following: Extroverted people are more likely to be energetic, assertive, sociable, talkative, stimulation-seeking, action-oriented, and enthusiastic. Introverts tend to be reserved, and have a preference for quieter, less stimulating environments. Raters also had selections from the JEPQ surveys for extraversion to give them a more comprehensive working definition of extraversion.

Self-Report Extraversion

Self-report extraversion scores were obtained from the Junior Eysenck Personality Questionnaire (JEPQ) personality test, conducted by LTS and QIMR prior to the current study.

Perceived Neuroticism

Eight female and two male raters rated each face for neuroticism. Raters were asked to rate “emotional sensitivity” rather than “neuroticism” to avoid bias from the colloquial use of “neurotic” in society. Instructions read: “Rate this face’s emotional sensitivity (prone to anxiety, depression, etc.,) on a scale from 1-7, 7 being most emotionally sensitive.” The working definition of emotional sensitivity is the tendency to easily experience negative emotions. The opposite end of the spectrum would be emotional stability – people with high emotional stability are calm, less easily upset, and less likely to experience negative feelings such as anxiety, depression, self-consciousness, and vulnerability. Selections from the JEPQ test were supplied for neuroticism as well to allow raters to be more consistent.

Self-Report Neuroticism

Self-report neuroticism scores were obtained from the JEPQ personality test, conducted by LTS and QIMR prior to this study.

Control Variables

The following variables were examined as potential mediators for any correlations between subjective and objective measures of the three traits. Undergraduate raters viewed slideshows as discussed above. The following variables were collected prior to the current study.

Attractiveness

Eight undergraduate research assistants rated attractiveness on a scale from 1-7, 1 being "low attractiveness" and 7 being "high attractiveness".

Smiling

Photos were rated (n=2 raters) on a scale from one to three, one being "No Smile", two being "Partial Smile", and three being "Full Smile".

Grooming

Raters (n=2) were asked to decide how much effort the subject had put into their appearance that day. Photos were rated on a scale of 1-7, 1 being "Un-groomed", and 7 being "Well Groomed". Grooming is related to attractiveness and may contribute to the halo effect.

Acne

Photos were rated (n=2 raters) on a scale from 1-7, 1 being "No Acne" and 7 being "Heavy Acne".

Socioeconomic Status

The American Psychological Association defines socioeconomic status as "the social standing or class of an individual or group. It is often measured as a combination of education, income and occupation." (American Psychological Association, 2012). Research has shown that Socioeconomic Status and IQ are positively correlated – a higher SES predicts a higher IQ and vice-versa. We therefore controlled for SES to test whether it mediated any of the relationships between subjective and objective/self-report measures.

Genomic Principal Component Scores

Although our sample was almost exclusively Caucasian, we wanted to assess whether subtle ethnic differences might mediate any potential effects observed. To do this, we included genomic principal components in our regression analyses to control for any subtle ethnicity differences between subjects that may have mediated our results. Both QIMR and LTS twins had been previously genotyped on genome-wide platforms for unrelated studies. Such genome-wide data can be used to accurately estimate subtle ethnic differences between people using a principal components analysis conducted on the genomic relationship matrix (derived from ~ 100,000 single nucleotide polymorphisms in roughly linkage equilibrium). We included the first five principal components as covariates in our regression analyses.

Statistical Analyses

To examine the relationships between perceived and measured evaluations of intelligence, extraversion, and neuroticism, we performed correlation and regression analyses using the R statistical package. Because the subjects were all twins and siblings, statistical tests conducted on the entire sample would yield biased (too low) p-values due to the dependencies in the data. We therefore split the subjects into two groups such that only one family member was randomly selected to be in each dataset. All analyses were run twice – once with group one (n= 730) and once with group two (n= 717) – thus creating an in-study pseudo-replication. The datasets were not truly independent because the second sample contained individuals from the same family (co-twins or siblings), and twins and siblings are inherently similar (especially in the case of monozygotic twins – these subjects look identical and will most likely receive similar ratings).

Beyond the initial correlations between the rated trait and its objectively measured counterpart, we performed a multiple regression analysis to control for factors that might explain the basic correlations. These factors are age, sex, grooming, smiling, BMI, acne, socioeconomic status, and ethnicity (as defined above).

Results

Intelligence

The zero-order correlation between Perceived Intelligence and Measured IQ was r = 0.161 (*p* = 9e-6, df=758) for group one and r = .105 (*p*< 0.006, df =691) for group two. When we look at Measured IQ as predicted by Subjective Intelligence score accounting for attractiveness, age, sex, grooming, smiling, BMI, and acne, the partial correlation increased, with an r of 0.37, (p =2.28e-5, df=693) for group one and r= 0.32, (p = 1.26e-11, df =651) for group two. None of the variables we controlled for had significant effects. We residualized IQ based on Principal Components in order to leave only the portion of intelligence not related to genetic differences. Residual ratings are the degree to which the predicted rating varies from the actual rating. After taking principal components and SES into account, the correlation remains about the same, r = 0.356 (p = 5.081e-9, df = 434).

Table 2 - Bivariate correlations between subjective intelligence and potential mediating factors

	IQ	**Subj. Int**	**Groom**	**Smile**	**Acne**	**Attr.**	**SEI**
IQ	1.00	0.10	0.03	0.08	-0.01	0.02	0.27
Subj. Int.	0.10	1.00	0.13	0.36	-0.04	0.28	0.14
Groom	0.03	0.13	1.00	0.05	-0.27	0.63	0.08
Smile	0.08	0.36	0.05	1.00	0.00	0.09	0.06
Acne	-0.01	-0.04	-0.27	0.00	1.00	-0.40	-0.08
Attr.	0.02	0.28	0.63	0.09	-0.40	1.00	0.14
SEI	0.27	0.14	0.08	0.06	-0.08	0.14	1.00

Predicting IQ From Subjective Intelligence Rating

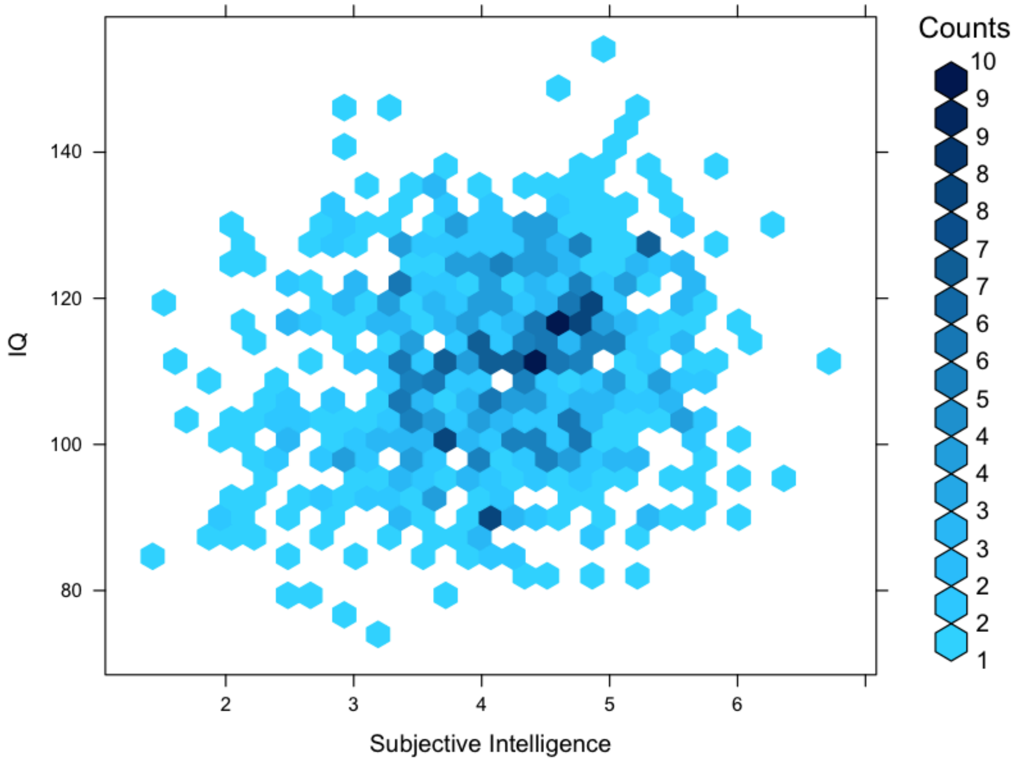


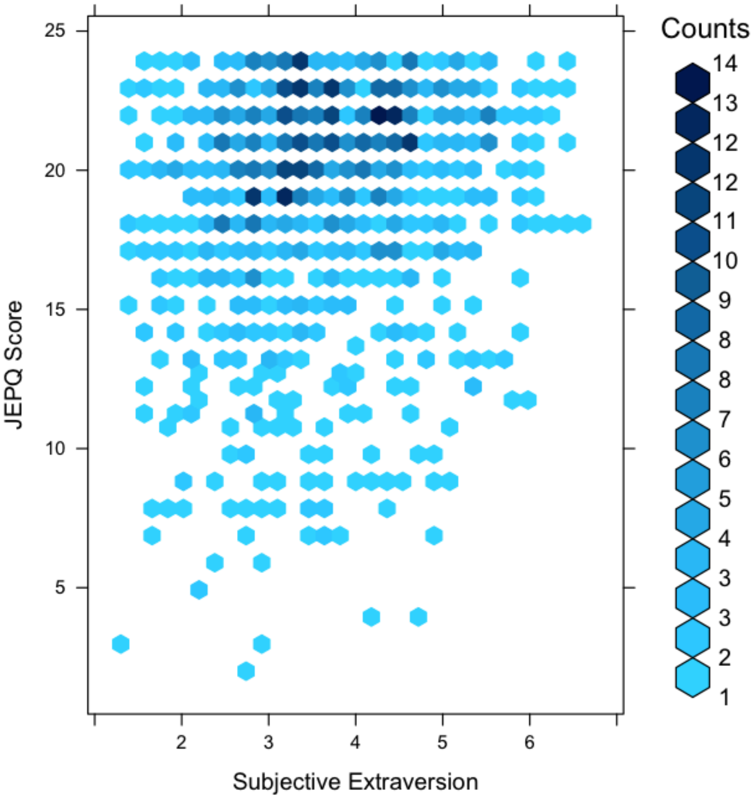
Table 3 – Bivariate correlations between subjective extraversion and potential mediating factors

	JEPQ Score	Subj. Extr.	Groom	Smile	Acne	Attr.	SEI
JEPQ Score	1.00	0.13	0.12	0.01	-0.02	0.14	-0.04
Subj. Extr.	0.13	1.00	0.40	0.64	-0.13	0.51	0.13
Groom	0.12	0.40	1.00	0.05	-0.27	0.63	0.08
Smile	0.01	0.64	0.05	1.00	0.00	0.09	0.06
Acne	-0.02	-0.13	-0.27	0.00	1.00	-0.40	-0.08
Attr.	0.14	0.51	0.63	0.09	-0.40	1.00	0.14
SEI	-0.04	0.13	0.08	0.06	-0.08	0.14	1.00

Extraversion

The zero-order correlation between Perceived Extraversion and Measured Extraversion from the Junior Eysenck Personality Questionnaire (JEPQ) is r= 0.163, (p = 0.0003004, df =488) for group one. For group two, r = 0.196 (p = 1.656e-05, df =473). When we examined Measured Extraversion as predicted by the Perceived Extraversion score accounting for grooming, smiling, and attractiveness, the correlation lost much of its significance, suggesting that those mediators were a very strong influence in the raters' perceptions of extraversion (r=0.111, df= 485, p=0.01452). When socioeconomic status was included, the correlation dropped to r = 0.08187075 (df = 369, p = 0.1154). Socioeconomic status appears influence subjective extraversion separately from grooming, smiling, and attractiveness, which can be viewed as a "self-presentation" effect.

Predicting JEPQ Scores from Subjectively Rated Extraversion



Neuroticism

We found no correlation between measured JEPQ Neuroticism scores and Subjective Emotional Sensitivity scores. The correlation was r =0.003 (df = 1063, p-value = 0.928). After controlling for BMI, sex, zygosity, age, grooming, smiling, acne, and socioeconomic status, the correlation remains insignificant: r = -0.003 (df = 369, p-value = 0.9608). The bivariate correlations suggest that subjective neuroticism is influenced negatively by grooming, smiling, and attractiveness, and positively by acne. However, none of the mediators correlate with self-report neuroticism score to a meaningful degree.

Predicting JEPQ Scores from Subjective Neuroticism

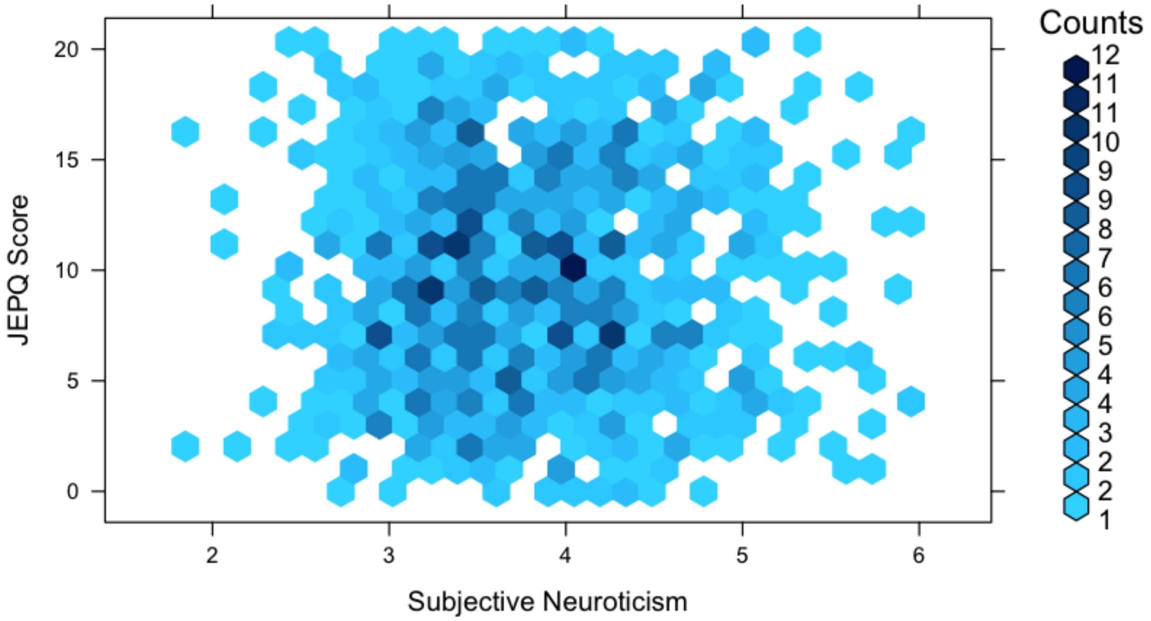


Table 4 – Bivariate correlations between subjective neuroticism and potential mediating factors

	JEPQ Score	Subj. Neur.	Groom	Smile	Acne	Attr.	SEI
JEPQ Score	1.00	-0.05	0.00	0.06	-0.08	0.03	-0.06
Subj. Neur.	-0.05	1.00	-0.35	-0.39	0.29	-0.43	-0.07
Groom	0.00	-0.35	1.00	0.05	-0.27	0.63	0.08
Smile	0.06	-0.39	0.05	1.00	0.00	0.09	0.06
Acne	-0.08	0.29	-0.27	0.00	1.00	-0.40	-0.08
Attr.	0.03	-0.43	0.63	0.09	-0.40	1.00	0.14
SEI	-0.06	-0.07	0.08	0.06	-0.08	0.14	1.00

Discussion

Our study builds on previous findings, and show that raters can judge ‘internal’ behavioral traits at levels above chance simply from brief assessments of photographs. Our results further suggest that the halo effect cannot explain these judgments because the correlation between subjective and objective measures of intelligence increased when we controlled for attractiveness. Clearly, raters can derive information from a face that is not mediated by traditional physical attractiveness.

Although the correlations between the subjective and objective measures are small, the p-values are still statistically significant, meaning these results are very unlikely to have arisen by chance. Small correlations imply that accurate information about extraversion and intelligence is available in photographs of people's faces, but there is not much of it. A large number of observations increase the power to detect small but real effects, as well as the likelihood that these results accurately estimate the relationships between people's internal traits and observers' ability to assess them in photographs.

For intelligence, the objective-subjective correlation (r = 0.316) is consistent with the result from Zebrowitz’s meta-analysis, which found that the average correlation between perceived and actual intelligence across studies is 0.3. Zebrowitz concluded that the correlation is likely due to the halo effect, meaning attractiveness can explain the accuracy in predicting IQ from subjective intelligence (2002). However, our results showed no significant effect of attractiveness in predicting IQ.

Extraversion is predictable from our subjective impressions, but seems to be partially due to observations about grooming, smiling, and attractiveness. The zero-order correlation, r = 0.188 (p=3.692e-05, df =473), was significant, and such mediators as BMI, age, and sex, do not have any significant effects. However, controlling for smiling, grooming, and attractiveness diminishes the significance substantially, suggesting that these three factors are potential mediators of the subjective-objective extraversion relationship. These factors can be thought of together as a “self-presentation” variable that raters reported using as criteria for making their ratings. Since each of these factors were related to both subjective extraversion and objective extraversion, the relationship goes down between those two after controlling for smiling, grooming, or attractiveness.

Raters reported using smiling and grooming as criteria for extraversion, but attractiveness may play a more subconscious role. Attractiveness is the "personal characteristic most obvious and accessible", according to attractiveness stereotype research (Dion, 1971). Our research therefore does supply some support for a "halo effect" of extroversion – people can indeed guess at someone's level of extraversion from their appearance, but this appears largely to be due to the fact that attractive, groomed people who smile are more likely to be extraverted. When predicting Subjective Extraversion from self-report extraversion along with other factors, the JEPQ score had a small significant effect, but most of the variance was due to grooming, attractiveness, and degree of smile. There was also a large effect from socioeconomic status (SES), which was slightly related to attractiveness (r=0.14), but not to grooming or smiling. Apparent SES therefore may be included in the halo effect: more attractive people are expected to have a higher SES and vice versa. Observing clothing, jewelry, and hairstyle may have contributed to higher ratings of both attractiveness and subjective extraversion.

Correlations between subjective and objective neuroticism measures seem to be due entirely to chance (p values were not significant). Pervious studies have found significance in judging emotional sensitivity in males but not in females (Penton-Voak, 2006). However, the present findings did not reveal any significant sex differences for neuroticism.

Cronbach's α measures for inter-rater reliability were excellent for attractiveness, smiling, and extraversion ($\alpha = 0.87$-0.9), good for grooming and acne ($\alpha = 0.7$-0.8), and decent for intelligence and neuroticism ($\alpha =0.6$). This makes sense given the raters' descriptions of their rating methodology. Smiling is more objective than subjective – "Is this subject smiling?" does not leave much room for interpretation, so $\alpha \approx 0.9$ is expected. Attractiveness ($\alpha = 0.87$) is more subjective, but previous research has shown that evaluations of facial attractiveness are highly consistent across raters, even cross-culturally (Langlois et al., 2000) or in young infants (Langlois et al., 1987; Slater et al., 1998).

Raters reported using similar criteria in making their ratings of subjective extraversion, which likely contributes to the high degree of agreement ($\alpha = 0.9$). High extraversion ratings were given for subjects with confident expressions in the eyes, genuine smiles, piercings, and wild hairstyles. Low extraversion ratings were assigned to subjects who were timid, anxious-looking, or un-groomed.

Grooming and acne were each rated by only two raters, so we can expect these α measures to be lower. Acne, the less subjective of the two variables, had an α of 0.77. Grooming ($\alpha = 0.7$) is difficult to define, and easy to confuse with attractiveness. Grooming was defined to be a measure of how much effort the person put into their appearance that morning, to try to avoid conflating grooming with attractiveness and vice-versa. Different standards of grooming were allowed in the LTS and QIMR sets, but when sample was controlled for, there we no significant differences between them.

Intelligence and Neuroticism have the lowest reliability ($\alpha = 0.6$ and $\alpha = 0.57$ respectively). Raters reported judging intelligence based on a "gut instinct". In rating high emotional sensitivity, raters paid the most attention to the expression in the eyes and the degree of smile. Low emotional sensitivity ratings were given to individuals with more defined facial features (e.g. square jaw), smaller eyes, confident body language, and genuine smile. Despite similar criteria, the α measure for Neuroticism is still low (though not unacceptable).

The true correlations between subjective measures of intelligence and neuroticism may have been higher if there had been more raters, and thus more reliable subjective impressions. Individuals are not very accurate in their subjective ratings, but once aggregated, statistically significant correlations can be observed. For example, an individual rater's subjective assessment of intelligence only correlates slightly with IQ (r= 0.), but the overall correlation once the subjective ratings are averaged between all the raters is r = 0.3. We did have a large number of subjects (n=1600), but not all data was available for all subjects, so the sample size was reduced. Splitting the data into two groups to preserve independence also decreased the sample size, and thus the power of our observations. However, the current study had a large number of target faces compared to other studies, and although some subjective measures were not as reliable as we would have liked, the large sample size provided some compensation.

The main limitation to the current study was the lack of standardization in the photographs used for ratings. Although the photos were controlled for LTS twins, the larger proportion of the sample (QIMR) had photographs for identification purposes that were not intended for use in subjective ratings. Participants were not asked to maintain neutral expressions, or refrain from grooming, and smiling and grooming had very large effects in perceived personality traits. Grooming is a difficult factor to assess regardless – some people may inherently appear more groomed if more attractive, or appear more attractive if they groom regularly.

The current study has determined that there is in fact something in the face besides attractiveness that displays internal traits, the next step is to examine what it is. Perhaps by measuring certain facial features and correlating them with our subjective impressions of traits we will discover a key to phenotypic displays of personality. People must be picking up on some facial cue in order to develop the “gut reaction” described by the raters, especially in the domain of intelligence. A large sample with controlled photos is necessary to exclude factors such as smiling. With today’s new technologies and analytic methods, people’s early fascination with judging character from features deserves a second chance.

Our study found that intelligence can be judged accurately even when controlling for potential mediators including attractiveness, SES, and perceived grooming. Extraversion can also be judged accurately, but appears to be mediated by attractiveness, grooming, and smiling. Judgments of Neuroticism, on the other hand, could not be predicted by subjective ratings. This suggests that humans can pick up on valid cues towards a person’s internal traits without observing any of their interactions. Since we make judgments about personality from facial characteristics every day, the study of personality attribution from facial features – new physiognomy – is certainly worth further study.

REFERENCES

Barlow, Nora ed. 1958. *The autobiography of Charles Darwin 1809-1882. With the original omissions restored. Edited and with appendix and notes by his grand-daughter Nora Barlow*. London: Collins

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organiza-tional Behavior and Human Decision Processes, 20*(1): 238– 252.

DeYoung, C.G., et al. (2008). Externalizing behavior and the higher order factors of the big five. *Journal of Abnormal Psychology, 117*(4): 947-953. doi:10.1037/a0013742

Hassin, R., Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology, 78*(5): 837-852. doi: 10.1037/0022-3514.78.5.837

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

McCartney, K., Harris, M.J., & Bernieri, F. (1990) Growing up and growing apart: A developmental meta-analysis of twin studies. *Psychological Bulletin*, *107*(2): 226-237. doi: 10.1037/0033-2909.107.2.226

Nisbett, R.E., Wilson, T.D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4):

250-256. doi:10.1037/0022-3514.35.4.250 Retrieved from: http://psycnet.apa.org/journals/psp/35/4/250/

Passini, F. T., Norman, W. T. (1966). A universal conception of personality structure? *Jounal of Personality and Social Psychology, 4*(1): 44-49. doi:10.1037/h0023519 Retrieved from: http://psycnet.apa.org/journals/psp/4/1/44

Penton-Voak, I., et al. (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. doi: 10.1521/soco.2006.24.5.607

Sheppard, L.D. et al. (2011). The effect of target attractiveness and rating method on the accuracy of trait ratings. *Journal of Personnel Psychology, 10*(1):24–33 doi: 10.1027/1866-5888/a000030

Thomas, J.C., Meeke, H. (2010). Rater error. *Corsini Encyclopedia of Psychology.* doi:10.1002/9780470479216.corpsy0774

Willis, J., Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science, 17*(7): 592-598. doi: 10.1111/j.1467-9280.2006.01750.x

Wright, M. (2001). Genetics of cognition: Outline of a collaborative twin study. *Twin Research, 4(*1): 58-46. DOI: http://dx.doi.org/10.1375/1369052012146

Yzerbyt, V.Y., Kervyn, N., & Judd, C. (2008). Compensation versus halo: The unique relations between the fundamental dimensions of social judgment. *Personality and Social Psychology Bulletin, 35*(8): 1110-1123.

doi: 10.1177/0146167208318602 Retrieved from: http://onlinelibrary.wiley.com/doi/10.1002/9780470479216.corpsy0774/full

Zebrowitz, L. A., Collins, M. A., & Dutta, R. (1998). The relationship between appearance and personality across the life span. *Personality and Social Psychology Bulletin, 24*(1): 736–749.

Zebrowitz, L. A., et al. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin, 28*(2): 238-249.