Cognition 158 (2017) 81–89

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Physical attraction to reliable, low variability nervous systems: Reaction time variability predicts attractiveness

Emily E. Butler [a,b], Christopher W.N. Saville [b,c], Robert Ward [b], Richard Ramsey [b,*]

[a] Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
[b] Wales Institute for Cognitive Neuroscience, School of Psychology, Bangor University, Bangor, Gwynedd, Wales, UK
[c] University Hospital for Child and Adolescent Psychiatry, University of Freiburg, Germany

ARTICLE INFO

Article history:
Received 3 December 2015
Revised 11 October 2016
Accepted 25 October 2016
Available online 2 November 2016

Keywords:
Face perception
Attractiveness
Reaction time variability
Central nervous system

ABSTRACT

The human face cues a range of important fitness information, which guides mate selection towards desirable others. Given humans' high investment in the central nervous system (CNS), cues to CNS function should be especially important in social selection. We tested if facial attractiveness preferences are sensitive to the reliability of human nervous system function. Several decades of research suggest an operational measure for CNS reliability is reaction time variability, which is measured by standard deviation of reaction times across trials. Across two experiments, we show that low reaction time variability is associated with facial attractiveness. Moreover, variability in performance made a unique contribution to attractiveness judgements above and beyond both physical health and sex-typicality judgements, which have previously been associated with perceptions of attractiveness. In a third experiment, we empirically estimated the distribution of attractiveness preferences expected by chance and show that the size and direction of our results in Experiments 1 and 2 are statistically unlikely without reference to reaction time variability. We conclude that an operating characteristic of the human nervous system, reliability of information processing, is signalled to others through facial appearance.

1. Introduction

Theories of mate selection emphasise the role of attractiveness preferences for guiding mate-choice towards high fitness partners. Specifically, traits that are associated with high fitness should be attractive to potential mates because they offer advantages to a partner as well as to future offspring (Gangestad & Scheyd, 2005; Rhodes, 2006). In humans, facial attractiveness preferences have been repeatedly shown to reflect aspects of mate quality. For example, facial symmetry is attractive and indicative of developmental stability and resilience (Simmons, Rhodes, Peters, & Koehler, 2004) and certain levels of facial colouration are attractive and denote healthy blood oxygenation (Stephen, Coetzee, Smith, & Perrett, 2009). The human face thus reflects a range of important fitness information. Given humans' high investment in the central nervous system (CNS), we would predict that cues to CNS function would be especially important in mate selection. To date, however, there is no evidence that facial appearance specifically reflects the reliability of the CNS; that is, the degree to which the nervous system functions in a consistent manner. Although consistency of behaviour would cue CNS reliability, appearance cues would offer the advantages of rapid assessment for observers as well as rapid signalling for high-fitness senders. We therefore tested if facial attractiveness preferences are sensitive to the reliability of human nervous system function.

A reliable information processor will produce relatively invariant outputs for a specified input (Shannon, 1948). One important constraint on the reliability of an information processor is the amount of endogenous noise. Endogenous noise can be defined as the amount of unpredictable fluctuation across processing operations within a system. Increasing amounts of endogenous noise eventually become an enemy of reliable information processing (Faisal, Selen, & Wolpert, 2008; Shannon, 1948). In other words, given a repeated input, low noise systems will be reliable, in the sense of producing invariant output. In contrast, high noise systems will produce more variable outputs. We hypothesised that if facial appearance reflects CNS reliability, then facial attractiveness should be correlated with the variability of behavioural outputs.

Interest in understanding variability *within* individuals is not new (Thouless, 1936; Woodrow, 1932), but it has not been widely acknowledged. Over 50 years ago, Fiske and Rice (1955) conducted

* Corresponding author.
E-mail address: r.ramsey@bangor.ac.uk (R. Ramsey).

http://dx.doi.org/10.1016/j.cognition.2016.10.012
0010-0277/

a systematic review showing that within-person variability – fluctuation in performance across trials or sessions – is not random, but stable, and provides an enduring marker of underlying function. As we describe below, the stable nature of within-person variability and its functional importance have been further supported in the following decades (MacDonald, Nyberg, & Backman, 2006), and is now most frequently assessed by standard deviation in reaction time (RT) across multiple trials (Li, Huxhold, & Schmiedek, 2004). Under this view, rather than reflecting measurement error that should be ignored, consistency of performance is predictive of psychophysiological function (Fiske & Rice, 1955; Thouless, 1936).

Using a variety of RT tasks, evidence has accumulated across cognitive, neurobiological, behavioural and health levels to demonstrate that increased RT variability is associated with reduced functional capacity of the CNS (MacDonald et al., 2006). At a cognitive level, reduced working memory, attention regulation, inhibition, and processing speed have been associated with RT variability (Kofler et al., 2013). At a neurobiological level, higher variability is associated with reduced structural and functional integrity of large-scale brain networks (MacDonald, Li, & Backman, 2009), as well as altered neurotransmitter function (Li & Rieckmann, 2014). In terms of health outcomes, RT variability predicts long-term mental and physical health with healthier individuals showing more consistent performance and those in a diseased state fluctuating more (MacDonald, Hultsch, & Dixon, 2008; MacDonald et al., 2006). Also, as human biology deteriorates in older age, performance on a range of tasks becomes more variable (Li et al., 2004). In sum, cognitive and neurobiological systems are more intact, efficient and healthy in individuals with more consistent performance and compromised in individuals with more varied performance. In addition, reaction time variability and its neural underpinnings have been shown to be heritable, using both quantitative (McLoughlin, Palmer, Rijsdijk, & Makeig, 2014) and molecular genetic approaches (Saville et al., 2014, 2015). In sum, a wealth of evidence from a range of methods supports the importance of processing reliability, as assessed by RT variability, as an important correlate to cognitive, neural, and health-related measures. Consequently, a mate-choice preference for low variability would produce indirect benefits through connection to a high-fitness partner.

Although the cognitive, neural and health correlates of RT variability are becoming clearer, the relationship between RT variability and social signalling remains unknown. Given the large investment of the human species in CNS operation, perceptible cues to CNS function would be expected to be identified and exploited. In particular, we predicted that any visual correlates to CNS reliability should be perceived as attractive. To test whether CNS reliability is visible and attractive, we created composite images from a dataset of 230 individuals who had a headshot photo taken and completed an RT task, which involved raising one of two fingers in response to numerical cues (Fig. 1A). In our first experiment, composite images were made of the 15 individuals from the dataset with the most variable (highest standard deviation of reaction time, SDRT) and least variable (lowest SDRT) latency distributions, for men and women separately. These composite images were shown to a new set of observers who were asked to pick which was more attractive and give an attractiveness rating for each face (Fig. 1B). We measured how frequently low SDRT faces were chosen as more attractive than high SDRT faces, as well as the difference in attractiveness ratings between low and high SDRT faces. If nervous system reliability is signalled through the face, then attractiveness judgments should be associated with low RT variability.

2. Experiment 1

2.1. Method

2.1.1. Participants

58 participants (29 female, M_{age} = 20.3 years, SD = 3.2) took part in the experiment. All participants had normal or corrected-to-normal vision and provided written informed consent prior to data collection. The data reported here were obtained under approval from the Research Ethics and Governance Committee of the School of Psychology at Bangor University. One participant completed the discrimination task but not the ratings task (see below for task details). Thus, 57 participants were included in the analysis of ratings data (28 female, M_{age} = 20.3 years, SD = 3.3).

2.1.2. Stimuli

In total, four composite images of faces were used (see Fig. 1). Based on prior research (Kramer & Ward, 2010; Penton-Voak, Pound, Little, & Perrett, 2006), 15 individual face images were "averaged" using a software package that enables multiple individual faces to be combined into one average face (JPyschomorph; Tiddeman, Burt, & Perrett, 2001). Separately for males and females, the composite images comprised face images from 15 individuals with the highest SDRT and lowest SDRT. SDRT was measured from a sample of 230 participants performing a cognitive control task (for full details, see Butler, Ward, & Ramsey, 2015).

The cognitive control task was developed by Brass et al. (2000) and requires participants to hold down two keys on a computer keyboard and lift one finger in response to a number cue, as quickly and accurately as possible. Simultaneously participants observe a congruent or an incongruent finger movement. Differences between congruent and incongruent conditions were not relevant to the current study and are reported elsewhere (Butler et al., 2015). As such, SDRT is calculated across all 60 trials. Importantly, reaction time variability is a relatively stable construct, which has been shown to have good test-retest and odd-even reliability metrics (Saville et al., 2011). In addition, factor analytic approaches have shown that single factor solutions have normally been adequate to summarise reaction time variability across a number of tasks (Saville et al., 2012; Schmiedek, Oberauer, Wilhelm, Süß, & Wittmann, 2007). Thus, it is likely that the task used here would yield comparable measures of reaction time variability to most other conventional reaction time tasks.

To calculate SDRT scores in order to make the stimuli for the current study, we first excluded participants who were <50% accurate for either condition. We then excluded trials where RT was <100 ms, or >1500 ms, as Ratcliff (1993) showed that this improved power to detect changes in the tau component of the ex-Gaussian distribution, which has been shown to be highly correlated with SDRT (Saville et al., 2011). This led to the exclusion of less than half a percent of the overall number of RTs. Finally, we computed standard deviations for congruent trials and incongruent trials separately and took a mean of these so that each participant had one average SDRT score. Individuals were then ranked according to SDRT. Separate rank orders were produced for males and females in order to generate separate male and female composite images. Face images of the 15 individuals with the lowest SDRTs were then combined into a composite image (low SDRT). The same process was followed using the 15 individuals with the highest SDRTs (high SDRT). The age range of included individuals was narrow for female (low SDRT: 18–27; high SDRT: 18–26) and male composites (low SDRT: 18–26; high SDRT: 18–22) and did not differ between low and high composites (female mean

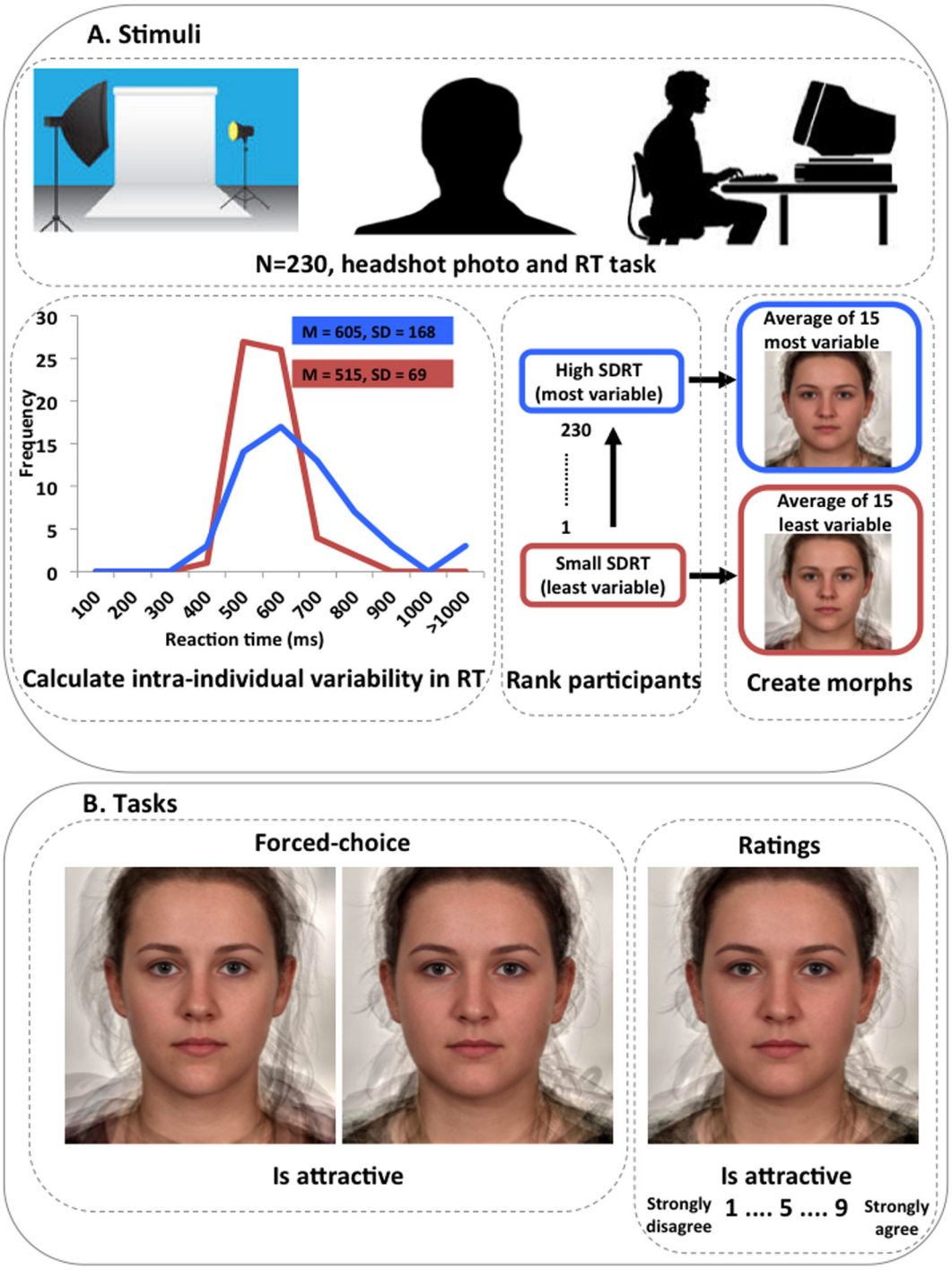

Fig. 1. Stimuli and tasks. (A) Stimuli were generated from data collected from 230 individuals, each of whom had a photo taken (headshot, hair pinned back, makeup and jewellery removed), before completing a computer-based reaction time task. For each individual, intra-individual (within-person) variability in RTs was calculated using standard deviation in RT across trials (SDRT). Participants were ranked from least variable to most variable. Photographs of the 15 individuals with the biggest SDRT were morphed into one composite image (high SDRT). The same procedure was carried out with photographs from the 15 individual's with the smallest SDRT across trials (low SDRT). These composite images were then used in subsequent tasks. (B) Judgements of composite images were measured using two different tasks. A forced-choice discrimination task asked participants to choose which of two images matched a statement best. By contrast, a ratings task showed one composite image per trial and asked participants to what extent they agreed on a 1–9 scale with the statement.

difference 0.78 years [−1.20, 2.76],[1] male mean difference 0.93 years [−0.62, 2.49]).

[1] Consistent with the American Psychological Association's *Publication Manual* (6th edition), square brackets denote lower and upper bounds of 95% Confidence Intervals.

2.1.3. Procedure and judgment tasks

To measure perceptions of attractiveness, a discrimination task and a ratings task were used (see below for task details). In addition, because more attractive faces are typically perceived to be physically healthier (Cunningham, 1986; Grammer & Thornhill, 1994) and more sex-typical, at least for women (Perrett et al.,

1998; Rhodes, Hickford, & Jeffery, 2000), we also assessed judgments of physical health and sex-typicality with the same two tasks. For full reporting and transparency (Simmons, Nelson, & Simonsohn, 2011), participants performed further ratings on these faces as part of a different line of enquiry.[2] All participants first completed the rating task and then the discrimination task. Within each task, each trial was shown once in a random order.

The discrimination task involved a two-alternative forced-choice task, where participants were presented with high and low SDRT composite faces (male and female versions were presented across different trials). Underneath the pair of composite images, participants were presented with a statement. The task was to choose which of the faces best represented this statement. Participants were instructed that they were under no time constraint to answer but that they should try to use their "gut instinct". A single item "Is attractive" was used to assess attractiveness judgements. For physical health judgements, four items were used from the Short-Form 12-Item Health Survey, which assesses physical health (Ware, Kosinski, & Keller, 1996). An example physical health statement is "Finds it easy to climb the stairs". For sex-typicality judgments, participants responded to a single item "Is sex-typical - looks more masculine if a man, and more feminine if a woman". Therefore, there were 6 statements of interest (one for attractiveness, one for sex-typicality and four for physical health statements). Each statement was presented with the male and female face pairs for a total of 12 trials.

On each trial, a fixation cross was shown for 500 ms followed by presentation of a face pair and statement, which remained on screen until the participants made their response. Participants responded by pressing the 'n' key for the left face and the 'm' key of the right face. The high and low SDRT faces were randomly presented on the left and the right of the screen, and statement order was randomised for each participant.

For the ratings task, on each trial, participants saw a fixation cross for 500 ms followed by a single face image in the centre of the screen with a statement and the rating scale underneath. Participants were asked to rate how well the statement described the face, and were again told that there was no time constraint but that they should try to use their "gut instinct". For ratings of physical health and attractiveness, the scale was from 1 = strongly disagree to 9 = strongly agree, and for the sex-typicality ratings, the scale was from 1 = masculine to 9 = feminine. For sex-typicality judgments of male faces, ratings were reverse-scored so that higher scores reflect greater sex-typicality. The statements used were identical to those used in the discrimination task. Each face was presented with each statement once, which means there were 24 trials in total. The face, statement, and rating scale remained on screen until participants made their response.

2.1.4. Data analysis

For the discrimination task, the percentage that the low SDRT face was picked as being more attractive, sex-typical or physically healthy was calculated. For each measure, a group mean for the sample was calculated. Deviation from chance performance (50%), would suggest that high and the low SDRT faces are perceived differently. Values greater than 50% would suggest that the low SDRT face is perceived as more attractive, physically healthy and sex-typical, whereas values less than 50% would suggest the opposite. For the ratings data, ratings of high SDRT faces were subtracted from ratings of low SDRT faces. Thus, a positive number would suggest that low SDRT faces were rated as more attractive, physically healthy or sex-typical, depending on the question.

For both discrimination and ratings data, effects were estimated using 95% confidence intervals (CIs) and measures of effect size (Cumming, 2014). For the discrimination task, if the 95% CIs overlap with chance performance (50%), it will show that participants do not perceive the high or the low SDRT faces as reliably different. If the 95% CIs do not overlap with 50% it will show that the SDRT faces are being perceived differently. For the ratings task data, a difference to zero would show that ratings differ between high and low SDRT faces. Cohen's d_z will be used to measure effect size for group differences, which is calculated by dividing the average difference by the standard deviation of the difference (Cohen, 1992; Lakens, 2013). Sample size was determined by the following rule, which was to test at least 50 participants and stop data collection at the end of the semester. In a paired design, where participants complete both conditions, setting a two-tailed alpha level of 0.01 and a correlation between repeat measurements of 0.7, a sample size of 50 would provide 96% power to detect an effect a Cohen's d of 0.5 (calculated in ESCI; Cumming, 2012), which is conventionally considered a medium effect size (Cohen, 1992).

Linear mixed effects models, as implemented in the lme4 package (http://CRAN.R-project.org/package=lme4) within R (http://www.R-project.org/), were fit to the ratings data to determine whether the predictive power of SDRT on attractiveness could be accounted for by other likely candidate variables. In the baseline models, attractiveness was the dependent variable and fixed effects were fitted for physical health ratings, sex-typicality ratings, stimulus sex (0 = male, 1 = female), participant sex (0 = male, 1 = female), and the sex-typicality by stimulus sex interaction term. This was compared to a full model that also included SDRT. Both models had identical random effects structures, with a random intercept for each rater, and in line with the "keep it maximal" approach (Barr, Levy, Scheepers, & Tily, 2013), a random slope of physical health rating, sex-typicality rating, stimulus sex, typicality * stimulus sex, and SDRT for each rater. All variables were demeaned and scaled and models were fit by maximum likelihood. To evaluate model fit we examined Bayesian information criteria (BIC) for both models and also present the results of a χ^2 test for goodness of fit between the two models.

2.2. Results

2.2.1. Attractiveness judgements

Consistent with our prediction, low SDRT faces were chosen above chance-level (50%; Fig. 2), both for female composites 91.38% [84.09, 98.67] Cohen's $d_z = 1.46$[3] and male composites, 63.79% [51.32, 76.27] Cohen's $d_z = 0.28$. In addition, low SDRT faces were rated as more attractive than high SDRT faces, both for female composites 1.67 [1.11, 2.22] Cohen's $d_z = 0.78$ and male composites 1.04 [0.5, 1.57] Cohen's $d_z = 0.5$ (Fig. 2).

2.2.2. Physical health and sex-typicality judgments

For judgements of physical health, low SDRT faces were chosen above chance-level, both for female composites 76.29% [69.43, 83.16] Cohen's $d_z = 0.99$ and male composites, 66.81% [59.44, 74.18] Cohen's $d_z = 0.59$ (Supplementary Fig. 1A). In addition, low SDRT faces were rated as more physically healthy than high SDRT faces, both for female composites 0.59 [0.27, 0.92] Cohen's $d_z = 0.47$ and male composites 0.44 [0.18, 0.69] Cohen's $d_z = 0.45$ (Supplementary Fig. 1C).

[2] Additional statements comprised twenty items from the mini International Personality Item Pool (Donnellan, Oswald, Baird, & Lucas, 2006) assessing the Big-5 dimensions of personality and a further four items assessing the likelihood that the composite would imitate during social interactions. Four additional faces, which investigated a separate research question, were also assessed on all statements.

[3] Cohen's d is a measure of effect size with values of 0.2, 0.5 and 0.8 generally considered small, medium and large effects, respectively (Cohen, 1992).

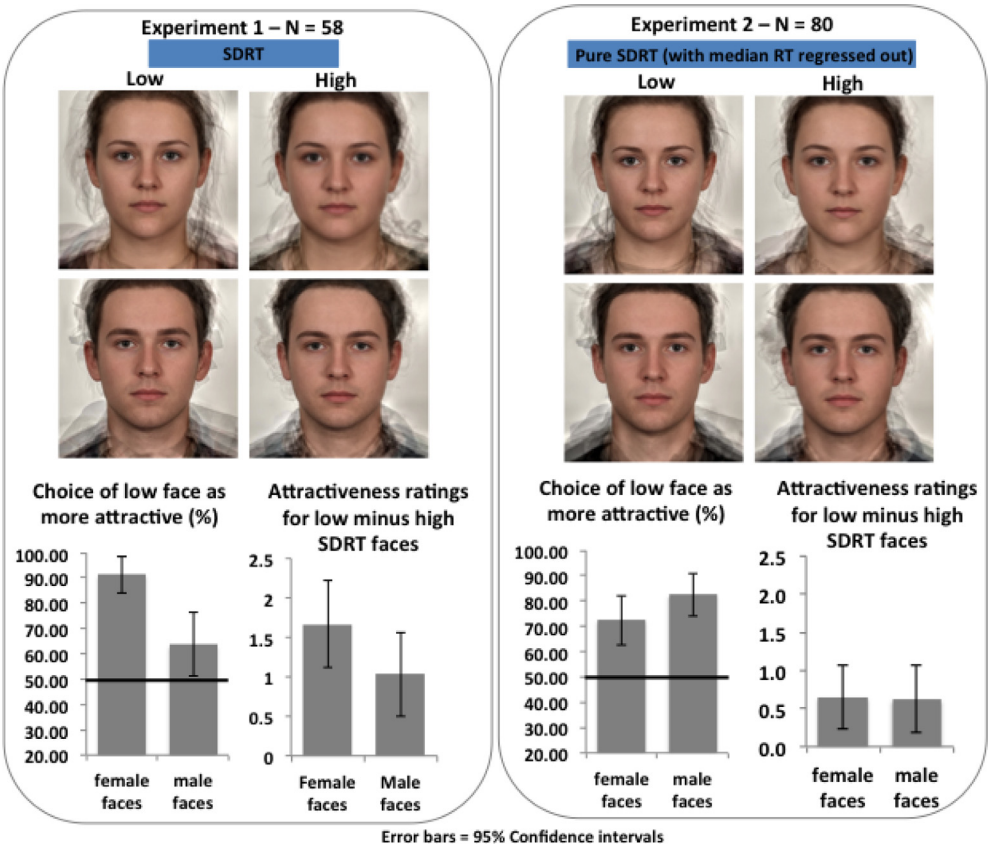

Fig. 2. Attractiveness judgements across Experiments 1 and 2. For each experiment, the top panel shows stimuli that were used (low and high SDRT/Pure SDRT). Underneath results from the forced-choice task and ratings task are displayed. For the forced-choice data, the percentage of times that the low SDRT composite was chosen is displayed. A score higher than chance performance (50%) indicates a preference for low SDRT faces when judging attractiveness. For the ratings data, a difference score is presented (high SDRT ratings subtracted from the low SDRT ratings). Thus, a positive score represents a higher rating for low than high SDRT faces.

For sex-typicality judgments, low SDRT faces were chosen above chance-level for female composites 84.48% [75.08, 93.88] Cohen's $d_z = 0.94$, but not male composites 51.72% [38.75, 64.70] Cohen's $d_z = 0.03$ (Supplementary Fig. 1A). In addition, low SDRT faces were rated as more sex-typical than high SDRT faces for female composites 1.32 [0.73, 1.90] Cohen's $d_z = 0.58$, but not male composites 0.35 [−0.21, 0.92] Cohen's $d_z = 0.16$ (Supplementary Fig. 1C).

2.2.3. Linear modelling

To investigate the features of attractiveness that may be driving these effects we used linear mixed effects modelling. Model fit statistics show that the model including SDRT outperformed the baseline model, which did not include SDRT ($BIC_{Baseline} = 640.31$, $BIC_{Full} = 626.25$; $p < 0.001$). Parameters for the full model can be found in Table 1. The negative weighting for SDRT reflects the attractiveness advantage for low variability over high.

2.3. Discussion

These findings provide the first evidence that reaction time variability is visible to others and attractive. Moreover, variability in reaction time made a unique contribution to perceptions of attractiveness above and beyond the contribution made from perceptions of physical health and sex-typicality, which have previously been associated with attractiveness judgments (Grammer & Thornhill, 1994; Perrett et al., 1998). Thus, neither perceptions of physical health nor sex-typicality fully capture the influence that nervous system variability has on facial attractiveness.

From these findings, however, we cannot infer that *variability* per se predicts attractiveness. Because RT is bounded with a

Table 1
Linear modelling results based on ratings data from Experiments 1 and 2.

	Experiment 1			Experiment 2		
Term	β	σ_β	t	β	σ_β	t
Intercept	−0 21	0.10	−2.05	−0.09	0.10	−0.95
Physical health	0. 18	0.07	2.35	0.18	0.08	2.43
Sex-typicality	0. 06	0.07	1.83	0.08	0.06	1.45
Participant sex	0.27	0.14	1.92	0.25	0.14	1.81
Stimulus sex	0. 27	0.06	4.07	0.06	0.06	1.00
Sex-typicality * stimulus sex	0.16	0.07	−2.16	−0.13	0.06	−2.34
SDRT (Exp.1), Pure SDRT (Exp.2)	−0.24	0.05	−4.78	−0.16	0.03	−4.73

Note: The negative weighting for SDRT and Pure SDRT reflects the attractiveness advantage for low variability over high variability.

minimum but effectively no maximum, individuals with more variable RTs will also have a higher ratio of slower responses (Jensen, 1992; Klein, Wendling, Huettner, Ruder, & Peper, 2006). As such, SDRT correlates with measures of speed, such as mean and median RT, although these components of RT are dissociable. For instance, median RT and SDRT predict individual difference outcomes in distinct manners (Hultsch, MacDonald, & Dixon, 2002; Kirkeby & Robinson, 2005). Hence, there is evidence to suggest speed and variability could have partially distinct relationships with nervous system function and biological signalling, which we investigate further in Experiment 2.

3. Experiment 2

3.1. Introduction

In this experiment we created new stimuli that dissociated speed from variability. New composite face images were generated by first regressing median RT from SDRT for each member of the database. A new variable - pure SDRT - was created that indexed variability with the effects of speed partialled out. We then ranked individuals within the database on pure SDRT and made composite morph images based on the highest and lowest individuals. A new set of observers then performed the identical tasks as Experiment 1. If variability, specifically, is signalled through the face, then attractiveness judgments should be associated with our new index of low RT variability, which is independent to the influence of general speed.

3.2. Method

3.2.1. Participants

Eighty participants who did not complete Experiment 1 (40 female, M_{age} = 19.9 years, SD = 2.7) had normal or corrected-to-normal vision and provided written informed consent prior to data collection.

3.2.2. Stimuli

Stimuli were produced using the same averaging procedure as in Experiment 1. However, individual images were chosen based on SDRT with median RT regressed out. To calculate SDRT scores that were independent of median RT we followed the same steps as in Experiment 1 in order to compute median RT for each participant. Using the statistical program R, we then fit a regression model to predict individual differences in SDRT using individual differences in median RT. In Experiment 2, the measure of SDRT was the residuals from this model, and thus is a measure of SDRT that is independent of median RT. We call this measure pure SDRT to denote that it is unrelated to general speed. We then ranked the faces by residual SDRT and chose the top 15 and bottom 15 following the same procedure as in Experiment 1. Thus, individuals with high pure SDRT were not necessarily those who were also slower overall. This said there was partial overlap between individuals used to create composites of SDRT and pure SDRT. Of the 15 faces in each composite, five individuals were in both composites for females with low SDRT, seven for males with low SDRT, and eleven for both the female and male high SDRT composites. The age range of included individuals was narrow for female (low SDRT: 18–27; high SDRT: 18–25) and male composites (low SDRT: 18–25; high SDRT: 18–22) and did not differ between low and high composites (female mean difference −0.40 years [−2.40, 1.60], male mean difference 1.00 years [−0.46, 2.46]).

3.2.3. Procedure and data analysis

The procedure and data analysis was identical to Experiment 1, with only one change to the rating scale presented with statements of sex-typicality. In Experiment 2, to improve clarity the rating scale for judgments of sex-typicality was from "1 = not very sex-typical i.e., feminine if a man and masculine if a woman" to "9 = very sex-typical i.e., masculine if a man and feminine if a woman".

3.3. Results

3.3.1. Attractiveness judgements

Consistent with our predictions, low pure SDRT composites were chosen as more attractive than high (Fig. 2), both for female faces 72.50% [62.65, 82.35], Cohen's d_z = 0.5, and male faces 82.50% [74.12, 90.88], Cohen's d_z = 0.85. In addition, low pure SDRT faces were rated as more attractive than high SDRT faces, both for female composites 0.65 [0.23, 1.07] Cohen's d_z = 0.34 and male composites 0.63 [0.19, 1.06] Cohen's d_z = 0.31 (Fig. 2).

3.3.2. Physical health and sex-typicality judgments

For judgements of physical health, low pure SDRT faces were not chosen above chance-level for female composites 54.69% [47.93, 61.44] Cohen's d_z = 0.15 but they were chosen above chance-level for male composites, 67.50% [61.19, 73.81] Cohen's d_z = 0.61 (Supplementary Fig. 1B). In terms of ratings data, low pure SDRT faces were not rated as more physically healthy than high pure SDRT faces, both for female composites 0.09 [−0.14, 0.32] Cohen's d_z = 0.09 and male composites 0.13 [−0.19, 0.44] Cohen's d_z = 0.09 (Supplementary Fig. 1D).

For sex-typicality judgments, low pure SDRT faces were chosen above chance-level both for female composites 63.75% [53.15, 74.35] Cohen's d_z = 0.28, and male composites 61.25% [50.51, 71.99] Cohen's d_z = 0.23 (Supplementary Fig. 1B). In addition, low pure SDRT faces were rated as more sex-typical than high pure SDRT faces for female composites 0.65 [0.17, 1.13] Cohen's d_z = 0.30, but not male composites 0.21 [−0.19, 0.62] Cohen's d_z = 0.12 (Supplementary Fig. 1D).

3.3.3. Linear modelling

As Experiment 1, we used linear mixed effects modelling to investigate the influence of RT variability on attractiveness judgments in comparison to other factors such as health and sex-typicality judgments. Model fit statistics indicated that the model including pure SDRT outperformed the baseline model ($BIC_{Baseline}$ = 893.52, BIC_{Full} = 879.16; $p < 0.001$). Parameters for the full model can be found in Table 1. As Experiment 1, the negative weighting for pure SDRT reflects the attractiveness advantage for low variability over high.

3.4. Discussion

The results of the second experiment demonstrate that even without a contribution from general speed, variability in RT predicts facial attractiveness. Further, pure SDRT made a unique contribution to perceptions of attractiveness above and beyond the contribution made from perceptions of physical health and sex-typicality. This shows that neither perceptions of physical health nor sex-typicality fully capture the influence that nervous system variability has on facial attractiveness.

The first two experiments show that there is a consensus across participants in judgements of attractiveness as a function of SDRT. However, these judgements were based on two pairs of stimuli per experiment (a male pair and a female pair). As such, although it is unlikely, it is conceivable that all four pairs of stimuli differed significantly in attractiveness because of chance variation in attractiveness, rather than due to SDRT. Experiment 3 addressed

this issue by measuring how frequently our method of stimulus generation would produce stimuli that differ in attractiveness preferences by chance.

4. Experiment 3

4.1. Introduction

In Experiment 3, we sought to estimate how frequently our method of stimulus generation would produce stimuli that differ in attractiveness preferences by chance. To do so, we generated 100 new random pairs of stimuli for male faces and 100 new random pairs of stimuli for female faces. Each time we created a new face pair, we randomly ordered our face database (without reference to SDRT) and created a new composite using the top 15 individuals and a new composite using the bottom 15 individuals. The two composite images became a new pair of stimuli. We then showed these new stimuli to a new set of participants and recorded attractiveness preferences in a similar manner to Experiments 1 and 2. This design, therefore, uses stimuli as targets of analysis, rather than participants. Across 200 pairs of stimuli (100 pairs per sex), we will be able to establish a baseline distribution of preferences. By referencing this baseline distribution, we will then be able to calculate the likelihood of obtaining similar results to our effects in Experiment 1 and 2 by chance alone. If our results from Experiments 1 and 2 are likely by chance, we should expect them to be close to the middle of the distribution. If our results are relatively unlikely by chance, we should expect them to be towards the tail of the distribution.

4.2. Method

4.2.1. Participants

Twenty-six participants who did not complete Experiment 1 or 2 (15 female, M_{age} = 21.3 years, SD = 2.2) had normal or corrected-to-normal vision and provided written informed consent prior to data collection.

4.2.2. Stimuli

Stimuli were produced using the identical averaging procedure as in Experiments 1 and 2. However, the individual images within each composite were chosen randomly, rather than according to SDRT. To do so, for male and female faces separately, the individual faces were first ranked in a random order. Then, the top 15 faces were averaged to form one composite image, and the bottom 15 faces were averaged to form a second composite image. The resultant two composite images became a face pair, which would later be used as a stimulus pair in the experiment. We then repeated this process to create 100 face pairs per sex.

4.2.3. Procedure and data analysis

We used the same discrimination task that was used in Experiments 1 and 2 and only assessed judgments of attractiveness. Participants were shown each face pair twice in a random order with each face image shown once on the left and once on the right. Therefore, participants completed 400 trials in total (100 face pairs per sex, each shown twice).

For each face pair, we arbitrarily labelled one of the images as ‘a’ and the other image as ‘b’. We then calculated the percentage of times that image ‘b’ was selected as more attractive than ‘a’, as well as the inverse preference (the percentage of times that ‘a’ was chosen as more attractive than ‘b’). We calculated both directions of the preference in order to avoid bias from arbitrary labelling the images. If, for example, by chance, more faces labelled ‘a’ were perceived as more attractive than ‘b’, then this would introduce unwanted bias and skew the distribution towards ‘a’ more than ‘b’. By including the inverse judgment, we perfectly balance any unwanted bias and still preserve the variance in judgements across stimuli, which is the key parameter that we want to estimate. In this way, we compiled a distribution of preference scores (200 values in total for each face sex). This distribution represents a baseline distribution by showing how many stimulus pairs, by chance, would produce a range of preference values.

In order to compare the effects from Experiment 1 and 2 to our baseline distribution, we calculated where each of our effects from Experiments 1 and 2 would be placed in terms of a percentile of the baseline distribution. In addition, we calculated a probability statement, which indexed the likelihood of generating a stimulus pair with a preference score equivalent to the effects observed in Experiments 1 and 2. As our original hypothesis was one-tailed (i.e., low SDRT faces would be judged as more attractive than high SDRT faces), we applied the same one-tailed logic to our new distribution. As such, we calculated the probability of obtaining a score towards one-tail of our distribution. To assign a probability value to each effect from our experiments, we converted mean difference scores into z-scores by dividing by the mean difference by the standard deviation of difference scores across all stimuli. We then associated each z-score with a corresponding probability statement (p value).

4.3. Results

The frequency distributions of preference scores across randomly-generated male and female composite stimuli are

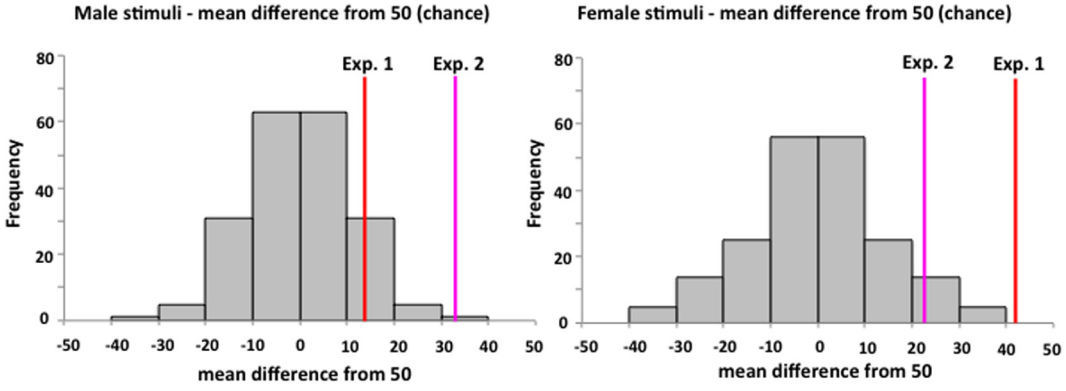

Fig. 3. Distribution of attractiveness judgements in Experiment 3. Separately for male and female stimuli, the distribution of attractiveness judgements in Experiment 3 are plotted. The effect plotted along the x axis is the difference score from chance (50%). Also plotted are the effects obtained in Experiments 1 and 2.

plotted in Fig. 3. For comparison, the equivalent mean difference scores from Experiments 1 and 2 are superimposed. In Experiment 1, the mean difference for female stimuli (41.37%) was equivalent to the 100th percentile, and translated into a z score of 2.82 and a probability of p = 0.002. The mean difference score for male stimuli (13.39%) was equivalent to the 90th percentile, which translated into a z score of 1.22 and a probability of p = 0.11.

In Experiment 2, the mean difference for female stimuli (22.5%) was equivalent to the 92nd percentile, which translated into a z score of 1.53 and p = 0.06. The mean difference for male stimuli (32.5%) was equivalent to the 99th percentile, which translated into a z score of 2.86 and p = 0.002.

In each experiment, male and female stimuli were selected from entirely independent samples. Therefore, we can calculate the probability of creating two stimuli that differed in attractiveness judgements per experiment as the compound probability across male and female stimuli (male probability ∗ female probability). In Experiment 1 the compound probability is p = 0.00022 and in Experiment 2 the compound probability is 0.00012.

4.4. Discussion

The results of the third experiment demonstrate that it is unlikely that our results in Experiments 1 and 2 were due to the chance construction of stimuli that coincidentally differed in attractiveness. Indeed, the four stimuli from Experiments 1 and 2 were in the 90th, 92nd, 99th or 100th percentile when referenced to our baseline distribution of preferences. Furthermore, we had strong *a priori* evidence to make a one-tailed directional hypothesis based on the effects of SDRT. Therefore, it is statistically unlikely that by chance alone we could have created stimuli with the magnitude of attractiveness differences we found, and in the direction we hypothesised.

5. General discussion

In sum, we show that the more reliable an individual's nervous system is, as evidenced by consistency of response time performance, the more attractive they appear to others. In evolutionary terms, the human species has invested heavily in the central nervous system. Here we have shown that at least some aspects of this investment are visible: visual facial traits are correlated with the reliability of information processing, and these traits are perceived as attractive. To our knowledge, this study is the first to demonstrate links between visual social cues and nervous system reliability in humans.

Attractiveness preferences have been repeatedly argued to be adaptively significant and to guide mate-selection towards desirable, high-fitness others (Gangestad & Scheyd, 2005; Rhodes, 2006). For example, a variety of cues to health and developmental stability are attractive (e.g., Simmons et al., 2004; Stephen et al., 2009). Given the importance of CNS reliability and its relationship with myriad cognitive, neural and health outcomes (MacDonald et al., 2006, 2009), as well as its heritability (McLoughlin et al., 2014; Saville et al., 2014), we suggest that visual cues to nervous system reliability could also bias mate-choice towards high-fitness partners. That is, a mate-choice based on attractiveness could, in part, guide one towards potential partners with more effective and efficient information processing systems and more favourable health outcomes. Social selection in this manner would produce benefits through connection to high-fitness others.

Importantly, reliability of information processing made a unique contribution to attractiveness judgements, which was above and beyond other factors that could guide mate-selection. For instance, although reliability is correlated with physical health and even mortality outcomes (MacDonald et al., 2008), we found that effects of reliability on attractiveness were dissociable from ratings of physical health. The attractiveness associated with reliability is therefore distinguishable from the attractiveness attributable to physical health ratings. Similarly, although sexual dimorphism has also been associated with physical health (Rhodes, 2006), the attractiveness associated with reliability was not a simple proxy for judgements of sex-typicality. Further, low levels of fluctuating asymmetry (deviation from perfect symmetry in bilateral features) has been frequently identified both as a feature of attractive faces, and as a potential consequence of healthy development (Rhodes, 2006). However, all our stimuli were composites of fifteen faces and are low in fluctuating asymmetry. Based on our current findings, therefore, the observed relationship between nervous system reliability and attractiveness preferences cannot be explained by physical health, sexual dimorphism or fluctuating asymmetry. One further factor to consider as a potential explanatory variable is psychometric g, which has been shown to be negatively correlated with reaction time variability (Larson & Alderton, 1990; Schmiedek et al., 2007; but see Saville et al., 2016 for a counterexample). Given the necessarily imperfect correlation between reaction time variability and IQ, the effect of g on facial appearance would have to be very large to account for the whole of our effects. However, recent studies do not show strong support for the claim that facial appearance provides valid cues to IQ (Mitchem et al., 2015; Talamas, Mavor & Perrett, 2016). At present, therefore, psychometric g is an unlikely explanation of our effects. In sum, future research will be required to further delineate the nature of the relationship between nervous system reliability and facial cues, which is likely to be complex.

Finally, our results draw further attention to the potential importance of studying the reliability of the human nervous system through reaction time variability. Over fifty years ago, reaction time variability was proposed to be a stable marker of psychophysiological function, rather than noise that should be ignored (Fiske & Rice, 1955). More recently, a widespread range of effects, some of which have profound consequences for health, have been associated with reaction time variability (MacDonald et al., 2006, 2009). To this we add that facial attractiveness is a cue to the reliability of the underlying nervous system, and that facial attractiveness therefore reflects an important operating characteristic of the human nervous system.

Author contributions

EEB, CWNS, RW and RR designed the experiments. EEB collected the data. EEB, CWNS, RW and RR analysed the data. RR and RW drafted the manuscript. EEB and CWNS made edits to the manuscript.

Acknowledgements

This work was supported by the Economic and Social Research Council (Grant Number: ES/K001884/1 to R.R.).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2016.10.012.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001.

Brass, M., Bekkering, H., Wohlschlager, A., & Prinz, W. (2000). Compatibility between observed and executed finger movements: Comparing symbolic, spatial, and imitative cues. *Brain and Cognition, 44*(2), 124–143.

Butler, E. E., Ward, R., & Ramsey, R. (2015). Investigating the relationship between stable personality characteristics and automatic imitation. *PLoS One, 10*(6). doi: ARTN e012965110.1371/journal.pone.0129651.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. http://dx.doi.org/10.1037/0033-2909.112.1.155.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. http://dx.doi.org/10.1177/0956797613504966.

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness – Quasi-experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology, 50*(5), 925–935. http://dx.doi.org/10.1037/0022-3514.50.5.925.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The Mini-IPIP scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment, 18*(2), 192–203. http://dx.doi.org/10.1037/1040-3590.18.2.192.

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience, 9*(4), 292–303.

Fiske, D. W., & Rice, L. (1955). Intra-individual response variability. *Psychological Bulletin, 52*(3), 217–250. http://dx.doi.org/10.1037/h0045276.

Gangestad, S. W., & Scheyd, G. J. (2005). The evolution of human physical attractiveness. *Annual Review of Anthropology, 34*(1), 523–548. http://dx.doi.org/10.1146/annurev.anthro.33.070203.143733.

Grammer, K., & Thornhill, R. (1994). Human (homo-sapiens) facial attractiveness and sexual selection – The role of symmetry and averageness. *Journal of Comparative Psychology, 108*(3), 233–242. http://dx.doi.org/10.1037/0735-7036.108.3.233.

Hultsch, D. F., MacDonald, S. W. S., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 57*(2), P101–P115. http://dx.doi.org/10.1093/geronb/57.2.P101.

Jensen, A. R. (1992). The importance of intraindividual variation in reaction time. *Personality and Individual Differences, 13*(8), 869–881. http://dx.doi.org/10.1016/0191-8869(92)90004-9.

Kirkeby, B. S., & Robinson, M. D. (2005). Impulsive behavior and stimulus–response variability in choice reaction time. *Journal of Research in Personality, 39*(2), 263–277. http://dx.doi.org/10.1016/j.jrp.2004.04.001.

Klein, C., Wendling, K., Huettner, P., Ruder, H., & Peper, M. (2006). Intra-subject variability in attention-deficit hyperactivity disorder. *Biological Psychiatry, 60*(10), 1088–1097. http://dx.doi.org/10.1016/j.biopsych.2006.04.003.

Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. (2013). Reaction time variability in ADHD: A meta-analytic review of 319 studies. *Clinical Psychology Review, 33*(6), 795–811. http://dx.doi.org/10.1016/j.cpr.2013.06.001.

Kramer, R. S. S., & Ward, R. (2010). Internal facial features are signals of personality and health. *Quarterly Journal of Experimental Psychology, 63*(11), 2273–2287. doi: Pii 922304861l0.1080/17470211003770912.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863.

Larson, G. E., & Alderton, D. L. (1990). Reaction time variability and intelligence: A "worst performance" analysis of individual differences. *Intelligence, 14*(3), 309–325.

Li, S.-C., Huxhold, O., & Schmiedek, F. (2004). Aging and attenuated processing robustness. *Gerontology, 50*(1), 28–34.

Li, S.-C., Lindenberger, U., Hommel, B., Aschersleben, G., Prinz, W., & Baltes, P. B. (2004). Transformations in the couplings among intellectual abilities and constituent cognitive processes across the life span. *Psychological Science, 15*(3), 155–163. http://dx.doi.org/10.1111/j.0956-7976.2004.01503003.x.

Li, S.-C., & Rieckmann, A. (2014). Neuromodulation and aging: Implications of aging neuronal gain control on cognition. *Current Opinion in Neurobiology, 29*, 148–158. http://dx.doi.org/10.1016/j.conb.2014.07.009.

MacDonald, S. W. S., Hultsch, D. F., & Dixon, R. A. (2008). Predicting impending death: Inconsistency in speed is a selective and early marker. *Psychology and Aging, 23*(3), 595–607. http://dx.doi.org/10.1037/0882-7974.23.3.595.

MacDonald, S. W. S., Li, S.-C., & Backman, L. (2009). Neural underpinnings of within-person variability in cognitive functioning. *Psychology and Aging, 24*(4), 792–808. http://dx.doi.org/10.1037/a0017798.

MacDonald, S. W. S., Nyberg, L., & Backman, L. (2006). Intra-individual variability in behavior: Links to brain structure, neurotransmission and neuronal activity. *Trends in Neurosciences, 29*(8), 474–480. http://dx.doi.org/10.1016/j.tins.2006.06.011.

McLoughlin, G., Palmer, J. A., Rijsdijk, F., & Makeig, S. (2014). Genetic overlap between evoked frontocentral theta-band phase variability, reaction time variability, and attention-deficit/hyperactivity disorder symptoms in a twin study. *Biological Psychiatry, 75*(3), 238–247. http://dx.doi.org/10.1016/j.biopsych.2013.07.020.

Mitchem, D. G., Zietsch, B. P., Wright, M. J., Martin, N. G., Hewitt, J. K., & Keller, M. C. (2015). No relationship between intelligence and facial attractiveness in a large, genetically informative sample. *Evolution and Human Behavior, 36*(3), 240–247.

Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from natural and composite facial images: More evidence for a "kernel of truth" in social perception. *Social Cognition, 24*(5), 607–640. http://dx.doi.org/10.1521/soco.2006.24.5.607.

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.

Ratcliff, R. (1993). Methods for dealing with reaction-time outliers. *Psychological Bulletin, 114*(3), 510–532. http://dx.doi.org/10.1037/0033-2909.114.3.510.

Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology, 57*(1), 199–226. http://dx.doi.org/10.1146/annurev.psych.57.102904.190208.

Rhodes, G., Hickford, C., & Jeffery, L. (2000). Sex-typicality and attractiveness: Are supermale and superfemale faces super-attractive. *British Journal of Psychology, 91*, 125–140. http://dx.doi.org/10.1348/000712600161718.

Saville, C. W., Beckles, K. D., MacLeod, C. A., Feige, B., Biscaldi, M., Beauducel, A., & Klein, C. (2016). A neural analogue of the worst performance rule: Insights from single-trial event-related potentials. *Intelligence, 55*, 95–103.

Saville, C. W. N., Lancaster, T. M., Davies, T. J., Toumaian, M., Pappa, E., Fish, S., ... Klein, C. (2015). Elevated P3b latency variability in carriers of ZNF804A risk allele for psychosis. *NeuroImage, 116*, 207–213. http://dx.doi.org/10.1016/j.neuroimage.2015.04.024.

Saville, C. W. N., Lancaster, T. M., Stefanou, M. E., Salunkhe, G., Lourmpa, I., Nadkarni, A., ... Klein, C. (2014). COMT Val158Met genotype is associated with fluctuations in working memory performance: Converging evidence from behavioural and single-trial P3b measures. *NeuroImage, 100*, 489–497. http://dx.doi.org/10.1016/j.neuroimage.2014.06.006.

Saville, C. W. N., Pawling, R., Trullinger, M., Daley, D., Intriligator, J., & Klein, C. (2011). On the stability of instability: Optimising the reliability of intra-subject variability of reaction times. *Personality and Individual Differences, 51*(2), 148–153. http://dx.doi.org/10.1016/j.paid.2011.03.034.

Saville, C. W., Shikhare, S., Iyengar, S., Daley, D., Intriligator, J., Boehm, S. G., ... Klein, C. (2012). Is reaction time variability consistent across sensory modalities? Insights from latent variable analysis of single-trial P3b latencies. *Biological Psychology, 91*(2), 275–282.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H. M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General, 136*(3), 414.

Shannon, C. (1948). The mathematical theory of communication. *Bell Systems Technical Journal, 27*, 379–423.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. http://dx.doi.org/10.1177/0956797611417632.

Simmons, L. W., Rhodes, G., Peters, M., & Koehler, N. (2004). Are human preferences for facial symmetry focused on signals of developmental instability? *Behavioral Ecology, 15*(5), 864–871. http://dx.doi.org/10.1093/beheco/arh099.

Stephen, I. D., Coetzee, V., Smith, M. L., & Perrett, D. I. (2009). Skin blood perfusion and oxygenation colour affect perceived human health. *PLoS One, 4*(4). doi: ARTN e508310.1371/journal.pone.0005083.

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

Thouless, R. H. (1936). Test unreliability and function fluctuation. *British Journal of Psychology. General Section, 26*(4), 325–343. http://dx.doi.org/10.1111/j.2044-8295.1936.tb00802.x.

Tiddeman, B., Burt, M., & Perrett, D. (2001). Prototyping and transforming facial textures for perception research. *IEEE Computer Graphics and Applications, 21*(5), 42–50. http://dx.doi.org/10.1109/38.946630.

Ware, J. E., Kosinski, M., & Keller, S. D. (1996). A 12-item short-form health survey – Construction of scales and preliminary tests of reliability and validity. *Medical Care, 34*(3), 220–233. http://dx.doi.org/10.1097/00005650-199603000-00003.

Woodrow, H. (1932). *Quotidian variability* (Vol. 39 (pp. 245–256). US: Psychological Review Company.