This paper was accepted for publication in ***Psychological Science***. This is a non-final and non-copy-edited version of the paper.

Personality across world regions predicts variability in the structure of face impressions

DongWon Oh, Jared D. Martin, and Jonathan B. Freeman

New York University

Corresponding author:
DongWon Oh
6 Washington Place
New York, NY 10003
orcid.org/0000-0002-2105-3756
Email: dongwon.oh@nyu.edu

Abstract

Research on face impressions has often focused on a fixed and universal architecture, treating regional variability as noise. Here, we demonstrate a crucial yet neglected role of cultural learning processes in forming face impressions. In Study 1, we found that variability in the structure of perceivers' face impressions across 42 world regions (n=287,178) could be explained by variability in the actual personality structure of people living in those regions. In Study 2, data from 232 world regions (n=307,136) revealed that perceivers use the actual personality structure learned from their local environment to form lay beliefs about personality, which in turn scaffold the structure of perceivers' face impressions. Together, these results suggest that people form face impressions based on a conceptual understanding of personality structure that they have come to learn from their regional environment. The findings call for greater attention on the regional and cultural specificity of face impressions.

Keywords: person perception, face processing, social cognition, semantic memory, cultural psychology

Abstract word count: 149

Statement of Relevance

Research on inferring personality traits from facial appearance (i.e., face impressions) has tended to focus on a universal cognitive architecture, often treating any regional variability as noise. Here, we propose a critical role of regional and cultural learning, whereby people learn how to form face impressions based in part on the structure of other people's personality traits that they encounter in their local environments. In two studies, we demonstrate meaningful variability across world regions in the structure of perceivers' face impressions, which was related to variability in the actual personality structure of people living in those regions. Further, we provide evidence that perceivers use this actual personality structure learned from their local environment to form lay theories about personality, and these lay theories in turn scaffold the structure of perceivers' face impressions. These findings suggest the crucial importance of regional and cultural learning in forming face impressions.

Statement of Relevance word count: 147

Personality across world regions predicts variability in the structure of face impressions

Although warned not to judge a book by its cover, people infer about any number of personality traits based on others' facial appearance. These trait judgments, or face impressions, are made with less than 100 ms of exposure and tend to be consistent across different perceivers (for review, see Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015). Judgments of specific traits (e.g., friendliness) are highly correlated with one another; thus, the structure of face impressions can be summarized by only a few dimensions, such as trustworthiness and dominance (Lin, Keles, & Adolphs, 2021; Oosterhof & Todorov, 2008; Sutherland et al., 2013). These core dimensions have often been interpreted through the lens of universal, evolutionarily adaptive processes, such as tracking other people's intentions (e.g., trustworthiness) and the ability to enact those intensions (e.g., dominance) (e.g., Fiske, Cuddy, & Glick, 2007; Oosterhof & Todorov, 2008). Indeed, recent studies found that the structure of face impressions was largely consistent across world regions (Jones et al., 2021; Lin et al., 2021), lending some support for a universal structure, although between-region variability was also observed.

Regional variability has been documented in various domains of face perception, including face impressions (Birkás, Dzhelyova, Lábadi, Bereczkei, & Perrett, 2014; Zhang et al., 2019), emotion perception (Elfenbein, Mandal, Ambady, Harizuka, & Kumar, 2002; Jack, Garrod, Yu, Caldara, & Schyns, 2012; Soto & Levenson, 2009), and more basic face perception processes (Caldara, 2017). However, to our knowledge, regional variability in the structure of face impressions has yet to be systematically demonstrated. Increasingly, research has documented meaningful differences across individual perceivers in their face impressions (Holzleitner & Perrett, 2017; Hönekopp, 2006; Martinez, Funk, & Todorov, 2020; Oh, Grant-Villegas, & Todorov, 2020; Sutherland et al., 2020; Xie, Flake, & Hehman, 2018) and across

target social categories (Collova, Sutherland, & Rhodes, 2019; Hehman, Sutherland, Flake, & Slepian, 2017; Oh, Dotsch, Porter, & Todorov, 2020; Sutherland, Young, Mootz, & Oldmeadow, 2015). These findings suggest that, while the notion of a universal structure of face impressions may successfully explain average trends, key variability may have gone relatively ignored. One source of this variability is idiosyncratic differences in perceivers' conceptual beliefs about traits and their covariation (e.g., to what extent being 'aggressive' relates to being 'intelligent') (Stolier, Hehman, Keller, Walker, & Freeman, 2018). Notably, these conceptual trait relations could be acquired through statistical learning processes as perceivers observe their social environment, including how other people's personality traits covary (Stolier, Hehman, & Freeman, 2020) – a premise consistent with classic research on how we implicitly infer others' personalities (Schneider, 1973) and more recent research on the role of environmental factors (Sutherland et al., 2020) and statistical learning processes (Dotsch, Hassin, & Todorov, 2016) in face impressions.

Much like conceptual relations across traits in people's minds, people's actual personality traits tend to be correlated along a small set of dimensions (e.g., Big Five) (Costa & McCrae, 1992). As actual personality traits are highly correlated, a simple strategy for perceivers to optimize trait inference would be to learn this correlation structure and make predictions accordingly. If perceivers learn the actual structure of personality, traits that are more similar in actual human personality would become conceptually believed to be more similar. If true, this possibility suggests that the conceptual structure of personality traits (and in turn, the structure of trait judgments of faces) would approximate the structure of actual personality traits perceivers observe in their social environments.

One important social environment may be the world region in which perceivers reside. Although the structure of personality is theorized to be universal (e.g., McCrae & Costa, 1997), reliable regional and cultural differences are often observed (McCrae, 2001, 2002). In particular, the dimensions of extraversion, agreeableness, and openness to experience have been found to vary across world regions (McCrae & Terracciano, 2005; Rolland, 2002). Thus, despite a degree of universality in the structure of personality, meaningful variability in human personality structure across world regions may shape the structure of perceivers' conceptual understanding of personality, which in turn may drive regional differences in the structure of how personality is judged from faces.

Here, we hypothesize that the average personality in perceivers' world regions will explain the conceptual understanding of personality in perceivers' minds, and in turn, explain how they infer personality in others' faces. For example, if a perceiver grows up in a world region wherein aggressive individuals tend to be intelligent, then the perceiver will tend to believe that aggressiveness and intelligence are conceptually related. As a result, the same perceiver will use similar facial features to judge whether targets are aggressive or intelligent and, thus, show positive correlation in their face judgments of these two traits. On the other hand, a perceiver who grows up in a region wherein aggressiveness and intelligence have little relationship would not develop this conceptual association and, in turn, would not show such a correlation in face judgments. As this learning process occurs for all pairs of personality traits, the structure of personality in one's world region would become the structure of one's conceptual beliefs about personality, which in turn would drive their face-based personality judgments. Indeed, environmental factors and statistical learning processes play a key role in face-based

personality judgments (Dotsch et al., 2016; FeldmanHall et al., 2018; Stolier et al., 2020; Sutherland et al., 2020; Verosky & Todorov, 2010).

We test our overall hypothesis across two studies. In Study 1, using personality data from individuals across 42 world regions, together with and an independent dataset of face-based trait judgments from the same regions, we test whether the regional structure of personality traits predicts the regional structure of face-based trait judgments. In Study 2, we examine the intermediary role that perceivers' learned conceptual understanding about personality traits plays. Together, two studies will suggest that the structure of people's actual personalities in a given world region shapes the conceptual understanding of personality in that region, which in turn affects how trait impressions of faces are formed in the region.

STUDY 1

Using international datasets of self-reported personality inventories and face-based trait judgments across 42 world regions, we tested whether people's self-reported personalities were related to how individuals in those regions judge others' personalities from faces.

Method

Participants

For the personality data, we used self-reported personality ratings from multiple world regions (Johnson, 2014; data available at https://osf.io/wxvth). Online participants living in 232 different world regions (n=307,313) participated in a personality survey. For face impressions data, 13,671 total participants living in 43 world regions participated in a laboratory setting (Jones et al., 2021; data available at https://osf.io/f7v3n). We only used the subset of face impressions data that corresponded to the same regions as those of the personality data. This resulted in final samples of self-reported personality data from 287,178 participants in 42 regions

(mean age=25.22, SD age=10.05; male 39.24%, female 60.76%) and face judgment data from 13,671 participants in those same 42 regions (mean age=22.63, SD age=7.00; male 29.12%, female 69.60%, declined to report gender/other gender 0.73%). The final 42 world regions were geographically and culturally diverse (see **Supplementary Figure 1a** for complete list and locations of the regions). All participant samples were of convenience. We used all available participants' data, and we did not predetermine the sample size.

NEOPI Personality Data

Participants answered the 300 items of the International Personality Item Pool Representation of the Revised NEO Personality Inventory (NEOPI), a well-known personality inventory administered online (IPIP-NEOPI) (Johnson, 2014). Each of these 300 items described a person's affective, behavioral, and/or cognitive tendency, with each item contributing to one of the 6 facets that compose each of the Big Five factors (agreeableness: *morality, altruism, cooperation, modesty, sympathy, trust;* conscientiousness: *self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, cautiousness;* extraversion: *friendliness, gregariousness, assertiveness, activity level, excitement-seeking, cheerfulness;* neuroticism: *anxiety, anger, depression, self-consciousness, immoderation, vulnerability;* openness to experience: *imagination, artistic interests, emotionality, adventurousness, intellect, liberalism*). For example, "Worry about things" measures the *anxiety* facet of the neuroticism factor; "Often feel blue" measures the *depression* facet of the neuroticism factor, "Prefer variety to routine" measures the *adventurousness* facet of the openness to experience factor, and "Like to get lost in thought" measures the *imagination* facet of the openness to experience factor. Each participant was given a 5-point scale to rate how accurately each item described themselves (1 Very inaccurate–5 Very accurate). Details of the 300-item international NEOPI personality survey procedure are

described in Johnson (2014). Mean NEOPI scales averaged across participants within specific world regions have been found to convey meaningful region-specific information (Allik et al., 2017).

Full details on data exclusion procedures are provided in Johnson (2005). Participants were instructed not to skip multiple responses, consecutively use the same response multiple times, or respond randomly. Randomness of responses was determined by within-participant reliability (correlation of non-overlapping subsets of a participant's responses that corresponded to one other in meaning). Participants were excluded if they did not follow the instructions. These exclusions aimed to remove participants who did not understand the questions or were not paying attention. The IPIP-NEOPI was administered in English across all regions, and all participants indicated understanding test instructions and the purpose of the test. Thus, participants with poor English comprehension were excluded.

Face Impressions Data

For face-based trait judgments, participants judged 120 faces on 13 personality traits. The faces were standardized photos from the well-validated Chicago Face Database (Ma, Correll, & Wittenbrink, 2015), including 30 Asian, 30 Black, 30 Hispanic, and 30 White faces (half male and half female within each race). Each participant was asked to rate the 120 target faces, one at a time, on one of the 13 personality traits: *aggressiveness, attractiveness, caringness, confidence, dominance, emotional stability, unhappiness, intelligence, meanness, responsibility, sociability, trustworthiness, weirdness* (taken from Oosterhof & Todorov, 2008). On each trial, a 9-point scale with a prompt was presented below the face (e.g., "How [aggressive] is this person? 1 Not at all [aggressive]–9 Very [aggressive]"). In each region, 25 or more raters were recruited to rate faces on each of the 13 traits for a sufficient level of interrater reliability. The task in each data-

collecting laboratory used the official language of their region (e.g., Farsi in Iran) or the most widely used language in the region (e.g., English in the US) to allow all raters to complete the task in their native language. In each region, the data-collecting teams translated the trait terms and the task instructions from an initial English version with the help of English-language dictionary definitions denoting the intended meaning of each of the trait words used. This approach had been used in prior studies testing for cultural differences in face processing (Han et al., 2018). Full details of the face-based trait rating procedure can be found in Jones et al. (2021).

Language Covariates

Previous studies have consistently found that administering a NEOPI personality inventory in two different languages (e.g., English and a non-English native language) on the same group of multilingual individuals produce highly similar individual NEOPI scores and regional NEOPI structure (e.g., Church & Katigbak, 2002; Gülgöz, 2002; McCrae, 2001; McCrae, Yik, Trapnell, Bond, & Paulhus, 1998; Piedmont, Bain, McCrae, & Costa, 2002; Piedmont & Chae, 1997; Simakhodskaya, 2000). For instance, when participants across multiple world regions completed the Revised NEO Personality Inventory (NEO-PI-R) in two different languages about a week apart, a high test-retest reliability between the languages was observed (median r=.86, McCrae, 2001). These findings suggest that, although English language was used for participants across all regions in the NEOPI personality data from Johnson (2014), it is unlikely to have introduced any meaningful biases in results (McCrae, 2001). The concern is further alleviated by the fact that, as described earlier, participants were excluded if their responses showed evidence of inconsistency or a lack of comprehension.

Nevertheless, it remains a possibility that regional differences in how well non-native English speakers comprehended the English-language NEOPI inventory could have confounded

the results. For example, if most participants in a region did not understand specific questions due to their limited facility with English, or misunderstood the questions, this could result in biases in responses. It is also possible that similarity between any two regions' primary language could confound effects of NEOPI similarity on face-judgment similarity. For instance, rather than similarity in the face-judgment structure of Germany and France being attributable to corresponding similarity in the two regions' personality structure, it is possible that these effects may be better attributed to the similarity of the German and French languages (e.g., suggesting that similarity in the face-judgment data reflects similar linguistic processes in interpreting the trait words and task, rather than any genuine effect of the two regions' NEOPI structure). To eliminate these possibilities, we repeated all regression analyses so as to include several language-related covariates.

In the case of analyses conducted at the level of traits (trait-level analysis; see *Analytic Approach* below), we included regions' primary language and level of English proficiency (using two complementary measures) in regression models. In the case of analyses conducted at the level of regions (region-level analysis), we included a measure of language dissimilarity between the primary language spoken in each pair of regions, as well as dissimilarity in the level of English proficiency in each pair of regions (using the two measures).

Pairwise language distance measures were derived from the Automated Similarity Judgment Program (ASJP) (Holman et al., 2008b). The ASJP database contains a set of common words across >7,000 languages throughout the globe (Søren, Holman, & Brown, 2020). Language relatedness data, such as the ASJP language distance, have been found to be capable of reconstructing the evolution of human language and culture (Atkinson, Meade, Venditti, Greenhill, & Pagel, 2008; Pagel, Atkinson, & Meade, 2007; Pompei, Loreto, & Tria, 2011) and

are associated with the geography of regions in which languages are spoken (e.g., distance from water) (Bentz, Dediu, Verkerk, & Jager, 2018). Using ASJP data for all available words with respect to the primary language spoken in each region, we calculated the Levenshtein distance for each pair of regions. Levenshtein distance is the standard method for calculating dissimilarity of languages (Holman et al., 2008a) and is based on the distance between pairs of words that have identical meanings. Specifically, it quantifies the difference between two strings, as defined by the minimum number of edited letters (i.e., insertion/deletion/substitution) needed to transform one string to the other (e.g., *blood* and *sangre*). As is common practice, after calculating the Levenshtein distance for all available words in the ASJP database based on the primary language for each pair of regions, we corrected them for word length and generated a normalized Levenshtein distance measure (as longer words would lead to an unwarrantedly larger dissimilarity value) (Holman et al., 2008a). All 42 regions had their primary languages in the ASJP database (24 languages in total), which allowed us to calculate dissimilarity values between all 276 language-pairs.

We also included two complementary measures of regions' English proficiency in regression models: the regional average TOEFL (Test of English as a Foreign Language) score (Educational Testing Service, 2021) and the regional average EPI (English Proficiency Index) (Education First, 2020) score. For region-level analyses, we included the dissimilarity (i.e., difference score) in the TOEFL score and in the EPI score between each pair of regions. Both scores are based on large number of test takers (over 1 million each) administered across the globe, allowing us to approximate region's average level of facility with English. The most recent TOEFL and EPI reports provided the measures on 165 and 100 regions, respectively.

Among the 42 target regions considered in Study 1, all 42 regions had TOEFL scores (100% of all regions) and 36 regions had EPIs (85.71%) available.

Ethnic Diversity Covariate

We also considered the ethnic diversity of the population in each region. Exposure to varying levels of ethnic diversity in each region could, in theory, affect perceivers' face judgments (e.g., Birkás et al., 2014; Hills & Pake, 2013; Xie et al., 2018; Zhang et al., 2019), particularly faces that varied in ethnicity as in the data from Jones et al. (2021). We used the Herfindahl-Hirschman Index (HHI) (Hirschman, 1945), which is a measure of homogeneity in a given group, derived from a regional ethnic fractionalization index (a probability of two randomly picked individuals belonging to two different ethnic groups) (Alesina, Devleeschauwer, Easterly, Kurlat, & Wacziarg, 2003). The ethnic-fractionalization HHI is correlated with regional differences in face-related variables, such as emotional expressivity (e.g., Rychlowska et al., 2015). The HHI was available for all 42 regions.

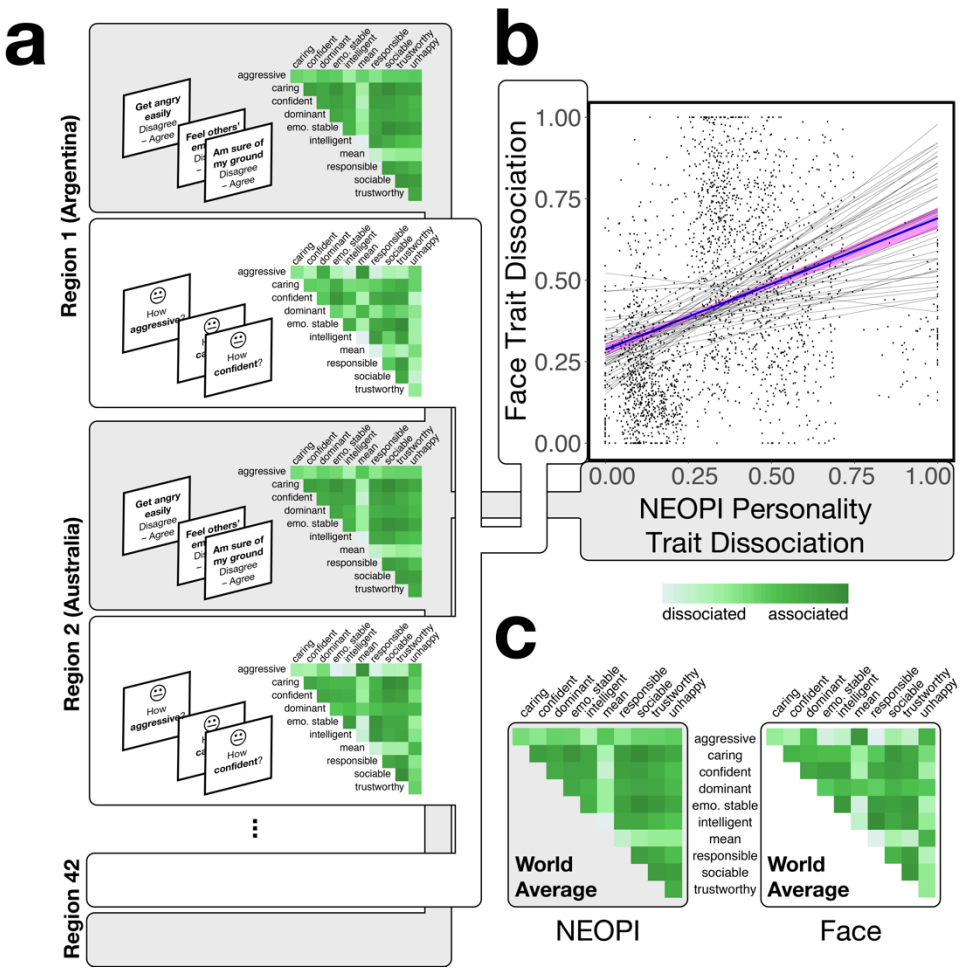

Figure 1. The analytic approach and results of Study 1's trait-level RSA. Subjects living in 42 regions (n=287,178) answered questions about their personality ("Lose my temper."; data from Johnson, 2014). An independent group of subjects residing in the same 42 regions (n=13,671) judged a set of 120 faces on personality traits (e.g., "How aggressive is this person?"). **(a)** Dissimilarity matrices (DMs) comprising all pairwise trait combinations in NEOPI self-reported personality and face-based ratings were generated for each region, wherein the euclidean distance between each pair of traits served as a measure of dissimilarity. Unique dissimilarity values in the DMs were vectorized and submitted to multilevel models predicting the structure of face-based trait impressions from the structure of NEOPI personality in the same regions. **(b)** A significant positive relationship was observed between dissimilarity values of NEOPI trait pairs and face-judgment trait pairs. Dots indicate individual trait pairs, e.g., aggressive–intelligent; thinner lines indicate slopes for individual regions; thicker line indicates average linear fit across regions; the shaded area represents the 95% confidence interval for the average linear fit (shown for illustrative purposes only; actual analyses were run using GEE multilevel regression). **(c)** The mean distances in personality traits and mean distances in face-based trait ratings, averaged across regions, are shown. In the panels a and c, only the upper triangles of the matrices are displayed to avoid redundancy.
Note: RSA=representational similarity analysis, NEOPI=Neuroticism-Extraversion-Openness Personality Inventory, emo. stable=emotionally stable.

Analytic Approach

To test whether the structure of personality traits of people in different regions predicts the structure of personality traits judged from faces in those regions, we took a representational similarity analysis (RSA) approach that tests the correspondence between regions' self-reported NEOPI personality trait space and face-judgment trait space. The NEOPI space was represented by a NEOPI trait dissimilarity matrix (DM) comprising all pairwise dissimilarities between self-reported personality traits in each world region (i.e., how similarly/dissimilarly people rate themselves in terms of their personality). The face-judgment space was represented by a face-based trait DM comprising all pairwise dissimilarities between evaluated trait dimensions of the set of 120 faces for each region (i.e., how similarly/dissimilarly people rate others' faces).

To map the NEOPI and face-judgment space via RSA, we analyzed only those traits common to both spaces. Two out of the 13 traits used in the face judgment task were excluded. Attractiveness was excluded because unlike the other traits, it refers to physical characteristics rather than inferred personality, and weirdness was excluded because it is not captured well by a high or low score on any single NEOPI personality trait. For each region, we averaged face ratings across all participants on each of the 11 remaining traits. We then mapped the NEOPI self-reported personality items (300 items) to the traits used for face judgments (11 traits).

Mappings were created using two converging approaches. In the first approach, the second author and four research assistants served as five coders, who for each NEOPI item marked which (if any) of the 11 face traits best described the item and in which direction it was related to the item (positive/negative). "None of these options" was included as the last option, to avoid any imprecise mapping. For each NEOPI item, when a majority (≥3) agreed that an item

corresponded to one of the 11 traits (and in the same coding direction), we considered the NEOPI-face coding of that particular NEOPI item as conclusive. A total of 124 final NEOPI items reached such agreement. For instance, "Like to solve complex problems." was coded as '*intelligent*' in the positive direction and "Avoid difficult reading material." was coded as '*intelligent*' in the negative direction.

To seek converging evidence, in the second approach, we recruited independent raters from Amazon Mechanical Turk living in the US (n=49). Raters were asked to indicate which (if any) of the 11 face traits best described each of the 300 NEOPI items (including a "None of these options" response). Although the coders of the first approach additionally rated positive vs. negative coding direction, we excluded these additional ratings for the independent raters. Including coding direction would have required an additional 300 responses per rater, which was infeasible given the time constraints of an online Mechanical Turk study. If a majority (>50%) of raters agreed that an item corresponded to one of the 11 traits, we considered that NEOPI-face mapping to be conclusive. Coding direction (positive/negative) for conclusive items was taken from the in-lab coder data and were self-evident (e.g., "Lose my temper." clearly corresponds to *aggressive* rather than not *aggressive*).

We only included data from independent raters who followed instructions and passed all attention check trials (18 trials randomly interspersed across 318 total trials, e.g., "Select the option comprised of four words.", "Select the second option from the bottom."). Because our aim was to extract reliable mappings, we adopted a high criterion of 100% accuracy in attention checks for data inclusion. This procedure left us with 24 raters (mean age=40.88, SD age=11.71, male 58.33%, female 41.67%; Black 8.33%, Hispanic 12.50%, Native American 4.17%, White

75.00%). Among these raters, a majority (i.e., ≥12) reached consensus that 103 NEOPI items reliably corresponded to one of the 11 traits.

The coders and independent raters showed substantial agreement in their mappings. Among 218/300 items (72.76%), the two groups agreed that an item corresponded to the same trait (or did not correspond to any trait). There was only one item for which the two groups mapped differently on a trait: 'Get angry easily' was judged by the coders as best corresponding to 'aggressive' but by the independent raters as 'emotionally unstable'. See **Supplementary Table 1** for the complete NEOPI–face coding scheme.

We averaged, for each region, all individual NEOPI responses across participants. We prepared for each of the 42 regions DMs that represented a NEOPI space and a face-judgment space. In the 42 region-specific NEOPI DMs, we calculated the euclidean distance for every pair of traits using average trait scores of all respondents on all 11 personality traits. As a result, each cell corresponded to the extent to which on average a trait pair co-occurred in individuals' self-reported personality in that particular region (e.g., co-occurrence of aggressiveness and intelligence, **Figure 1a**). In the 42 region-specific face-judgment DMs, we calculated the euclidean distance between the 120-face trait ratings (averaged across all participants in the region) for every pair of traits. As a result, each cell corresponded to the extent to which people's face judgments of the two traits (e.g., judgments of aggressiveness and intelligence) tended to covary (**Figure 1a**). Both DMs, for each region, were an 11×11 matrix, in which cells represented all pairs of 11 total traits (**Figure 1a** and **1c**). A larger value in any matrix indicated a stronger dissimilarity (i.e., greater euclidean distance). As is customary in RSA, similarity values were rank-ordered prior to regression models (or in correlation analyses, tested using Spearman rank-ordered correlations) so as to not assume linear relationships between variables

(Kriegeskorte, Mur, & Bandettini, 2008). The 55 unique trait pairs – the unique values under the diagonal in the NEOPI DMs and face-judgment DMs – were vectorized and submitted to regression analyses testing the relationship between face-judgment dissimilarity values and NEOPI personality dissimilarity values. To appropriately account for the multilevel nature of the data (55 trait-pairs nested in each of 42 regions), we conducted multilevel regressions using generalized estimating equations (GEE) (Liang & Zeger, 1986). In all GEE models, we report unstandardized regression coefficients (*B*) and Wald *Z* as a measure of effect size. For ease of interpretation, prior to analyses, all variables were rescaled to vary between [0,1] such that 0 corresponded to the smallest distance (maximum similarity between regions) and 1 corresponded to the largest distance (minimum similarity between regions).

We also conducted complementary RSA at the level of regions. We again mapped across a NEOPI personality trait space and a face-based trait space, but this time with 42×42 DMs with each cell representing the dissimilarity between any given pair of regions (**Figure 2**). NEOPI and face-judgment dissimilarity values between pairs of regions were calculated as the euclidean distance between the two regions' aggregated values for the 11 personality traits (i.e., the 13 traits of the face-judgment data after excluding attractiveness and weirdness, and using the coded mappings of those 11 traits to the NEOPI items, described above). However, this region-level RSA permitted greater flexibility in testing multiple indices of NEOPI dissimilarity and face-judgment dissimilarity, because correspondence did not need to be evaluated at the level of individual traits (only at the level of regions). Thus, to evaluate the robustness of the effects, the NEOPI 42×42 DM was also calculated using dissimilarity between pairs of regions in a) the 5 NEOPI factors, b) the 30 NEOPI facets, and c) the full 300 NEOPI items. The face-judgment 42×42 DM was also calculated using dissimilarity between pairs of regions in the full 13 traits

(reincluding attractiveness and weirdness). Notably, the additional analyses using all 300 available NEOPI items (a–c) and the additional analysis using all 13 face traits ensure that the effects of interest do not depend on any specific personality-face mappings applied to our data (i.e., by the in-lab coders or independent raters).

To test the relationship between the 42×42 NEOPI DM and 42×42 face-judgment DM in each of these cases, the 861 unique values under the diagonal of the DMs were vectorized and submitted to a Spearman correlation. Unlike with the trait-level analyses, the data are not multilevel and thus do not require GEE regression; however, for direct statistical comparison, we complemented Spearman correlations with GEE regressions for the region-level RSA.

All regression models were repeated after including the additional language use and ethnic diversity covariates described earlier.

Results

Trait-Level Analyses

We conducted a series of complementary multilevel regression analyses to provide evidence that people's unique personality structure in different world regions is reflected in how people form trait judgments of faces in those regions. First, we regressed regions' NEOPI dissimilarity values for the 55 trait-pairs onto their face-judgment dissimilarity values using GEE regression (trait-pairs nested within regions). There was a strong positive relationship, regardless of whether using mappings derived from coders (B=0.12, SE=0.02, 95% CI [0.08,0.16], Z=5.59, p<.001) or independent raters (B=0.06, SE=0.03, 95% CI [0.01,0.11], Z=2.21, p=.027), showing that the structure of people's personalities in a region was reflected in the structure of that region's face-based trait judgments (**Figure 1**). For example, if aggressiveness and intelligence

tend to co-occur more in the personalities of people in a given region, then people also tend to evaluate aggressiveness and intelligence more similarly in others' faces in that region.

To more directly assess unique and idiosyncratic differences in NEOPI and face-judgment structure across region, we conducted an additional multilevel regression analysis that clustered the data by trait-pair instead of region. This analysis thereby aims to show that, within a given trait-pair (e.g., aggressiveness and intelligence), regions with higher NEOPI dissimilarity values tend to also be the regions with higher face-judgment dissimilarity values for that specific trait-pair. This analysis therefore serves as a more stringent test of unique inter-regional differences in NEOPI structure that may be reflected in regions' face-judgment structure. NEOPI and face-judgment dissimilarity values were *z*-normalized within each region, thereby removing any differences in magnitude or scale in these variables (i.e., the possibility that some regions have higher/lower dissimilarity values overall, or more/less dispersion, across all 55 trait-pairs). Using GEE regression (regions nested within trait-pairs), we regressed trait-pairs' NEOPI dissimilarity values for the 42 regions onto their face-judgment dissimilarity values, which revealed a strong positive relationship regardless of whether mappings were derived from coders (B=0.42, SE=0.02, 95% CI [0.38,0.45], Z=23.75, p<.001) or independent raters (B=0.43, SE=0.02, 95% CI [0.39,0.48], Z=18.88, p<.001; **Supplementary Figure 2**).

We reconducted our analyses, this time including four covariates accounting for regions' language use and ethnic diversity: primary language, EPI, TOEFL, and ethnic-fractionalization HHI. Inclusion of these covariates did not meaningfully change the relationship between personality and face impressions. Specifically, the effects of NEOPI personality structure on face-impressions structure remained strongly significant, regardless of whether using mappings derived by coders (clustered by region: B=0.22, SE=0.02, 95% CI [0.18,0.27], Z=9.17, p<.001;

clustered by trait-pair: B=0.42, SE=0.02, 95% CI [0.38,0.46], Z=21.52, p<.001) or derived by independent raters (clustered by region: B=0.18, SE=0.03, CI 95% [0.12,0.24], Z=5.66, p <.001; clustered by trait-pair: B=0.42, SE=0.03, CI 95% [0.37, 0.47], Z=16.71, p<.001). See **Supplementary Table 2** for full statistics.

These results show, for example, that if aggressiveness and intelligence tend to co-occur in people's personalities more so in Australia than in Iran, then people in Australia also tend to evaluate the aggressiveness and intelligence of faces more similarly than do people in Iran. These complementary analyses therefore provide strong evidence that unique differences in human personality across world regions are reflected in corresponding differences in how people in those regions judge personality traits in others' faces.

Region-Level Analyses

As a corroborating analysis, we conducted region-level RSA, mapping NEOPI personality trait space and a face-based trait space by region rather than individual traits using 42×42 DMs, with each cell representing the dissimilarity between any given pair of regions on the basis of the 11 personality traits (**Figure 2**). Vectorizing the 861 unique values in the DMs, we observed a strong positive relationship between the NEOPI DM and face-judgment DM, regardless of whether using mappings derived from coders (B=0.26, SE=0.03, 95% CI [0.20,0.32], Z=8.04, p<.001; Spearman ρ=.26, 95% CI [.19,.32], p<.001) or independent raters (B=0.24, SE=0.03, 95% CI [0.18,0.30], Z=7.55, p<.001; Spearman ρ=.24, 95% CI [.18,.30], p<.001). When including the four covariates capturing dissimilarity in linguistic and ethnic diversity between region pairs – AJSP language distance, EPI difference, TOEFL difference, and HHI ethnic diversity difference – the results did not meaningfully change. Specifically, the relationship between NEOPI structure and face-impressions structure remained strong and

significant when the four covariates were included, whether using mappings derived by coders (*B*=0.19, SE=0.04, 95% CI [0.12,0.27], *Z*=4.92, *p*<.001) or independent raters (*B*=0.21, *SE*=0.04, 95% CI [0.13,0.28], *Z*=5.32, *p*<.001). See **Supplementary Table 3** for full statistics.

As this region-level RSA does not require trait-level correspondence across NEOPI and face-judgment space, this permitted greater flexibility to demonstrate the robustness of this relationship in a manner that does not require any personality-face mappings whatsoever. The strong positive relationship persisted regardless of whether the NEOPI DM was calculated using pairwise regional dissimilarity on the basis of the 5 NEOPI factors (*B*=0.13, *SE*=0.03, 95% CI [0.06,0.19], *Z*=3.76, *p*<.001; Spearman ρ=.20, 95% CI [.14,.27], *p*<.001), the 30 NEOPI facets (*B*=0.18, *SE*=0.03, 95% CI [0.11,0.24], *Z*=5.40, *p*<.001; Spearman ρ=.18, 95% CI [.11,.24], *p*<.001), or the full 300 NEOPI items (*B*=0.24, *SE*=0.03, 95% CI [0.17,0.30], *Z*=7.33, *p*<.001; Spearman ρ=.23, 95% CI [.17,.30], *p*<.001), or when the face-judgment DM was calculated on the basis of the full 13 traits (after reincluding attractiveness and weirdness) using mappings by coders (*B*=0.23, *SE*=0.03, 95% CI [0.17,0.29], *Z*=7.16, *p*<.001; Spearman ρ=.23, 95% CI [.17,.29], *p*<.001) or independent raters (*B*=0.23, *SE*=0.03, 95% CI [0.17,0.29], *Z*=7.10, *p*<.001, Spearman ρ=.23, 95% CI [.16,.29], *p*<.001). Thus, the region-level RSA demonstrated a highly robust relationship between NEOPI structure and face-impressions structure across regions. The extent to which any two regions' (e.g., Australia and Iran) personality structure was more similar predicted a corresponding similarity in perceivers' face-trait structure in those two regions.

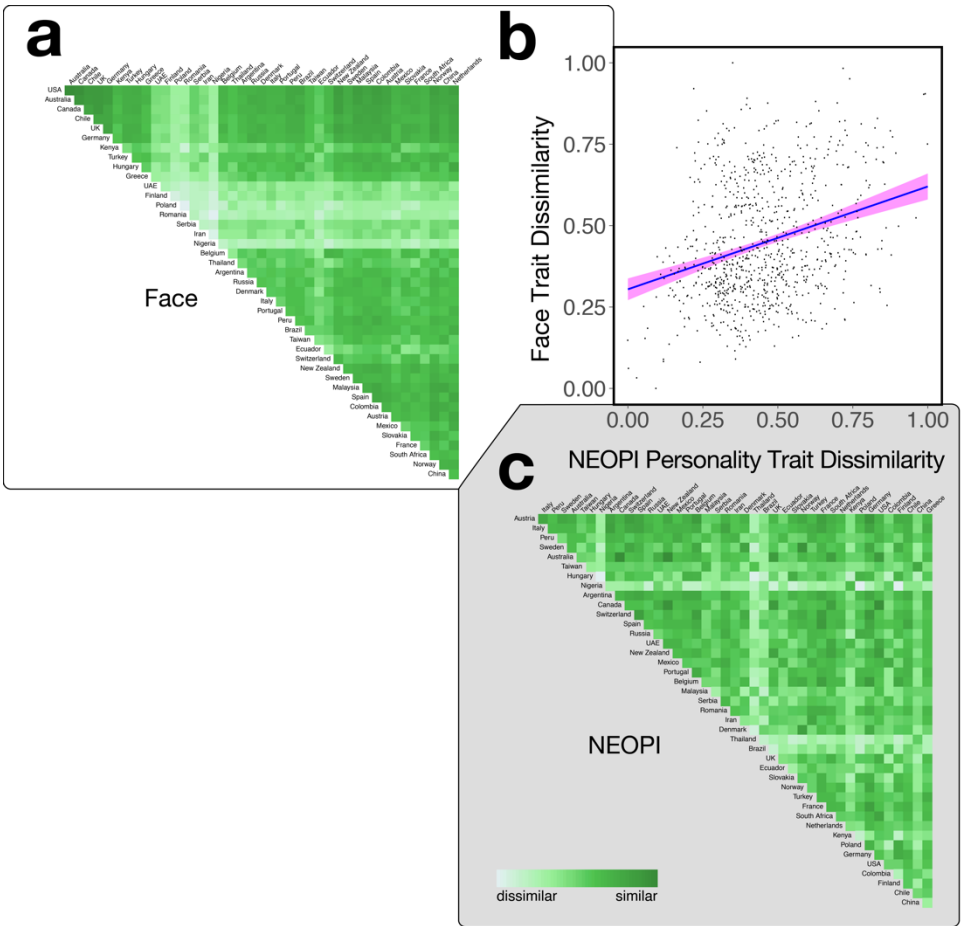

Figure 2. The results of Study 1's region-level RSA. Two independent groups of subjects residing in 42 regions answered questions about their personality using the NEOPI (N=287,178) and made trait judgments of a set of 120 faces (N=13,671). We calculated for all pairs of regions the euclidean distance between the two regions' NEOPI personality structure and face-based trait judgment structure (a). The 861 unique values under the diagonal were vectorized and submitted to Spearman correlation to test their correspondence, which revealed a positive relationship. The dots indicate individual region pairs (e.g., Australia–Iran); the line indicates the average linear fit across all region pairs; the shaded area represents the 95% confidence interval for the average linear fit (b). The distance between regions in NEOPI and face-traits, respectively, are displayed (a, c). In the panel a and c, only the upper triangles of the matrices are displayed to avoid redundancy.

Note: RSA=representational similarity analysis, NEOPI=Neuroticism-Extraversion-Openness Personality Inventory, USA=United States of America; UK=United Kingdom, UAE=United Arab Emirates.

Discussion

Across two types of RSA conducted at multiple levels of analysis (trait- and region-level), the results show that unique variability in the structure of human personality across world regions is reflected in the structure of how people in those regions form trait impressions of others' faces. Moreover, these effects hold even when accounting for regional variability in language use and ethnic diversity.

STUDY 2

We have hypothesized that perceivers' conceptual understanding of personality traits may explain the relationship between world regions' personality structure and the structure of those face impressions in those regions. For instance, if a perceiver observes that aggressive people tend be intelligent, they will conceptually associate those traits as cooccurring; in turn, that perceiver may use similar facial appearance to judge whether targets are aggressive or intelligent. In Study 2, we test the possibility that regional variability in personality structure is reflected in the structure of face impressions, which may be partly explained by regional perceivers' conceptual trait structure.

Method

Previous research has shown with US samples that US conceptual trait associations predict the structure of US face-based trait judgments (Stolier et al., 2020; Stolier et al., 2018). We used these previous US data on conceptual trait associations and face-based trait judgments in tandem with the publicly available dataset of NEOPI personality across world regions used in Study 1 (which includes the US). Using RSA, this allowed us to test whether the similarity in a given world region's personality structure to that of the US can predict how similarly that region's personality structure also resembles the structure of US conceptual beliefs and US face-based judgments. For instance, if Australia's personality structure is more similar to US

personality structure than is Syria's, we would expect Australia's personality structure to also more closely resemble the structure of US conceptual trait beliefs and US face-based judgments than Syria's personality structure. Thus, although widespread cross-regional data like that of NEOPI data and face-judgment data (used in Study 1) are not available for conceptual trait associations, using cross-regional NEOPI and face-judgment data in tandem with full data from the US (NEOPI, face-judgment, and conceptual-trait data) provides a valuable opportunity to test our hypothesis regarding the intermediary role of conceptual trait associations.

Participants

For the conceptual trait associations, we used published data from 115 US participants (mean age=35.38, SD age=10.47, male 47.83%, female 50.43%, declined to report gender/other gender 1.74%) (Stolier et al., 2020, Study 1). For the face-based trait judgments, we used data from 482 US participants (mean age=35.51, SD age=12.30, male 41.29%, female 58.30%, declined to report gender/other gender 0.42%). As an additional replication, we used an additional sample of face-based trait judgments from 496 participants (mean age=30.31, SD age=6.74, male 47.78%, female 51.81%, declined to report gender/other gender 0.40%) (Stolier et al., 2020, Studies 1 and 2). All data were taken from Stolier et al. (2020).

For the NEOPI personality data, we used the same personality data as in Study 1 of participants from different world regions (Johnson, 2014). As here we are not constrained by the subset of regions also available in the face-judgment dataset used in Study 1, for the present study what remained was a sample of 307,136 personality respondents across 232 regions, including the US (mean age=25.19, SD age=10.00; male 39.74%, female 60.26%). See **Supplementary Figure 1b** for the complete list and geographic locations of the regions. All

participant samples were of convenience. We used all available participants' data, and we did not predetermine the sample size.

NEOPI Personality Data

For the stimuli and procedure used for the personality data collection, see Study 1.

Conceptual Trait Association Data

For the trait association rating task, participants were asked to provide conceptual similarity ratings for all pairwise combinations of 15 personality traits: *adventurous, angry, anxious, assertive, cautious, cheerful, cooperative, depressed, dutiful, emotional, friendly, intellectual, self-disciplined, sympathetic, trustworthy*. The 15 traits represented 15 NEOPI facets (3 facets representing each of the 5 NEOPI factors). These 15 representative facets of the total 30 were found to be able to explain various domains of social perception, including representations of social groups and face impressions of strangers (Stolier et al., 2020). For example, for the pair of 'adventurous' and 'assertive', they were asked "How likely is an [adventurous] person to be [assertive]?" on a 7-point scale (1 Not at all likely–7 Very likely). Participants evaluated faces on the 15 same personality traits. Each participant rated the degree of association across all 105 unique trait-pairs, and all traits were presented twice to capture the association bidirectionally (e.g., how likely an *adventurous* person is *assertive*, and how likely an *assertive* person is *adventurous*). Details of the personality trait association rating procedure can be found in Stolier et al. (2020).

Face Impressions Data

For face-based personality trait judgments, participants judged 90 target faces on the same 15 personality traits (15 NEOPI facets) used in the conceptual trait association task. All images were of an identical race and gender (White male) and taken from the Chicago Face

Database (Ma et al., 2015). Independent groups of participants (each group n=25–30) were assigned to each of the 15 traits, and thus participants only rated faces on a single trait. Participants provided 7-point ratings (e.g., 1 Not at all [adventurous]–7 Very [adventurous]). Details of the face-judgment procedure can be found in Stolier et al. (2020).

Language Covariates

As in Study 1, we repeated all analyses while including covariates related to language use. Here, all language distance metrics captured the distance between language use of any given region and the US. The official language of each region was considered its primary language; if English was one of a region's multiple official languages, English was considered the primary language. The ASJP language distance was used to index language dissimilarity between a region's primary language and English, and a TOEFL difference score between a region's TOEFL score and the US' TOEFL score was used to index the difference in a region's facility with English relative to the US' facility with English. An EPI difference score was not included in the models because EPI is not measured in regions in which English is widely spoken as a first language, including the US, which prevents us from calculating an EPI difference score between any region and the US. Among the 232 world regions considered, all 232 regions had their primary languages (69 languages in total) in the ASJP database available (100%), and 172 regions had regional TOEFL scores available (74.46%).

Ethnic Diversity Covariate

To consider regional differences in ethnic diversity, as in Study 1, we included an ethnic-fractionalization HHI difference score (between a region's HHI and the US' HHI) as a covariate. The HHI was available for 166 regions (71.55% of all regions).

Analytic Approach

Because participants in the conceptual-trait and face-judgment tasks evaluated the identical 15 personality traits as the 15 NEOPI facets, data could be linked across NEOPI personality data, conceptual-trait data, and face-judgment data at the level of the same 15 personality traits. Unlike Study 1's trait-level RSA (but similar to Study 1's region-level RSA), analyses in Study 2 did not require a trait-level correspondence between the different data sources. Thus, personality-face mappings were not necessary for Study 2. For each of the 231 non-US regions, we calculated three measures, each a correlation between the non-US region's personality structure and either the (1) US' personality structure (US-to-region personality correlation), (2) US' conceptual-trait structure (US-to-region conceptual-trait correlation), and (3) US' face-judgment structure (US-to-region face-judgment correlation). For each of the three measures, there were 231 final values corresponding to the 231 world regions.

The first measure, the US-to-region personality correlation, was calculated as the Pearson correlation between the 300-item NEOPI trait scores of the US and the 300-item NEOPI trait scores of the non-US region. This measure thereby represents how similar the personality structure was between the US and any given world region.

The second measure, the US-to-region conceptual-trait correlation, was computed using RSA. We first created two 15×15 DMs, one for the US' conceptual-trait data and one for the non-US region's NEOPI data, with cells reflecting the dissimilarity (euclidean distance) between all pairwise combinations of the 15 personality traits. Cells of the US conceptual-trait 15×15 DM reflected US participants' conceptual beliefs that any given pair of traits tend to cooccur in other people; cells of the non-US region's NEOPI 15×15 DM reflected the extent to which that same pair of traits tends to actually cooccur in other people's personality in the region. The 105 unique values in the two DMs were vectorized and submitted to a Pearson correlation. This US-to-

region conceptual-trait correlation thereby represents the correspondence between a given non-US region's NEOPI personality structure and the US' conceptual-trait structure.

The third measure, the US-to-region face-judgment correlation, was also computed as RSA. Two 15x15 DMs were created, one for the US' face-judgment data and one for the non-US region's NEOPI data, with cells reflecting the dissimilarity (euclidean distance) between all pairwise combinations of the 15 personality traits. Cells of the US face-judgment 15×15 DM reflected US participants' tendencies to judge two personality traits similarly in response to the same faces; cells of the non-US region's NEOPI 15×15 DM reflected the extent to which that same pair of traits tends to actually cooccur in other people's personality in the region. The 105 unique values in the two DMs were vectorized and submitted to a Pearson correlation. This US-to-region face-judgment correlation thereby represents the correspondence between a given non-US region's NEOPI personality structure and the US' face impressions structure.

Mediation analyses were used to test the intermediary role (i.e., indirect effect) of conceptual trait associations. As in Study 1, we also reran all analyses after including covariates related to language use and ethnic diversity. We also conducted an additional corroborating analysis; the larger number of regions analyzed in Study 2 allowed us to conduct multi-level GEE regression analyses that clustered by primary language. If the relationships of interest persist even within clusters of regions with the same primary language (e.g., within the 75 English-speaking regions, within the 26 French-speaking regions, within the 20 Arabic-speaking regions, and within the 19 Spanish-speaking regions), this helps cement the evidence that the effects of NEOPI structure via conceptual-trait structure are not confounded by language.

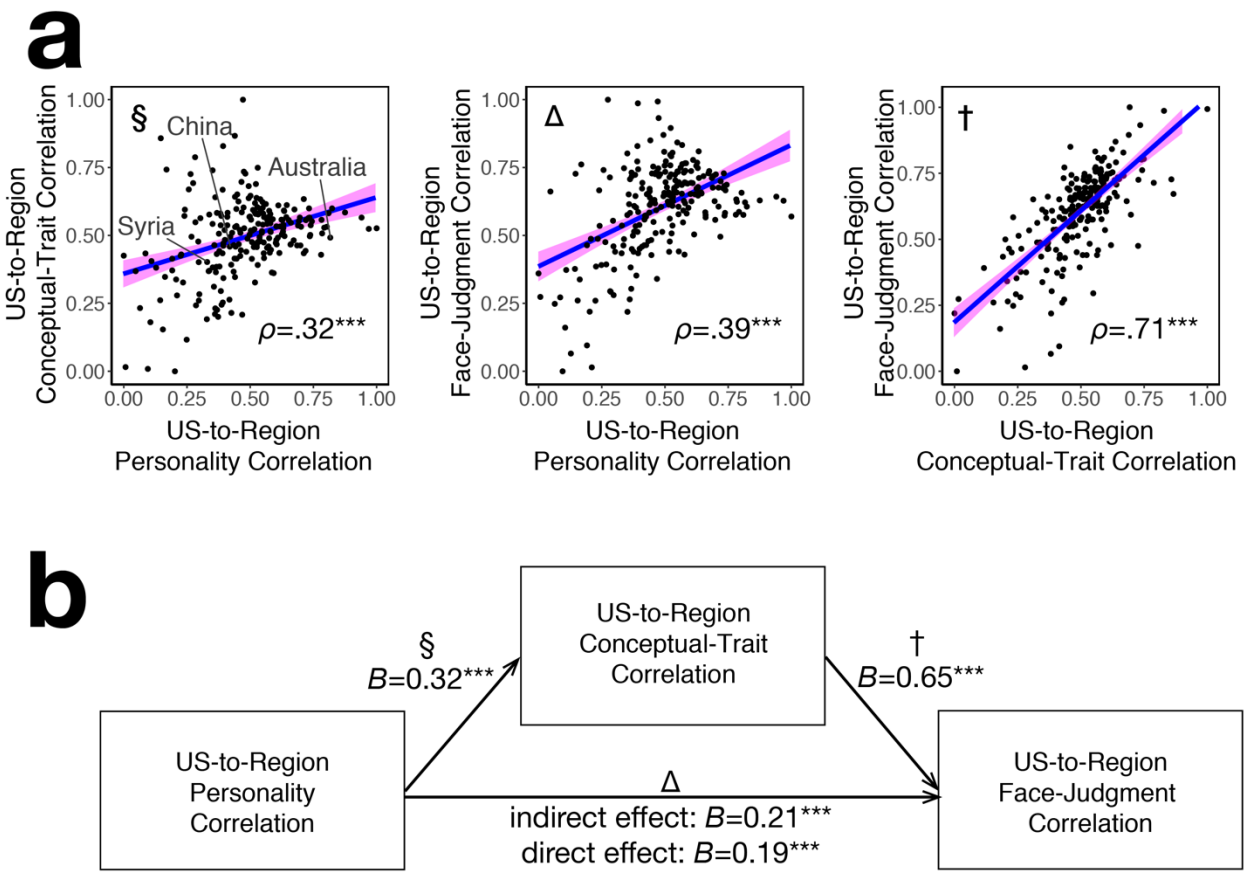

Figure 3. The results of Study 2. The similarity in residents' personalities between a given non-US world region and the US (§) predicted how strongly that region's personality was correlated with the conceptual personality trait associations in the US (∆), and in turn, with the face-based trait judgments in the US (†). Scatterplots representing the correlations across the three values are displayed with the blue line indicating the linear fit and dots indicating each of the 231 non-US regions. For illustrative purposes, Spearman correlations are displayed using unranked *r* coefficients (a). The US-to-region similarity in actual personality (§) predicted the US-to-region correlation in face impressions (†), which was partly explained by the US-to-region correlation in conceptual-trait associations (∆) (b).
Note: ***=p<.001.

Results

Using RSA, previous reports of the US data have shown that the structure of US NEOPI personality predicts the structure of US conceptual-trait associations (Spearman ρ=.77, 95% CI [.68,.84], p<.001) (Study 7, Stolier et al., 2020) which in turn predicts the structure of US face impressions (Spearman ρ=.80, 95% CI [.71,.86], p<.001) (Study 1, Stolier et al., 2020). This

previous result suggests that actual personality in the US environment may shape the structure of US perceivers' conceptual understanding of personality, which in turn sets the stage for their judgments of others' faces. Study 1 provided evidence that regional variability in personality structure relates to regional variability in the structure of face impressions. Our analyses here focused on helping explain this regional association described in Study 1 by way of conceptual-trait associations, using international data across 232 world regions (including the US) in terms of their NEOPI personality structure (as in Study 1) together with US-only data on face impressions and conceptual-trait structures. To the extent that any given world region is more similar to the US in terms of NEOPI personality structure, that region's NEOPI structure should be able to more strongly predict the US' conceptual-trait structure, and in turn, face-judgment structure (relative to other regions less similar to the US in terms of personality structure).

The three variables of interest were: (1) US-to-region personality correlation (correlation between regional NEOPI structure and US NEOPI structure); (2) US-to-region conceptual-trait correlation (correlation between regional NEOPI structure and US conceptual-trait structure); and (3) US-to-region face-judgment correlation (correlation between regional NEOPI structure and US face-judgment structure). The three variables (*r* coefficients) for the 231 world regions were submitted to Spearman correlation analyses, which revealed they were all positively correlated (Spearman ρs=.32–.71, *p*s<.001). Thus, if a given region (e.g., Australia) was more similar in personality structure to the US, then that region's personality structure was better able to predict the US' conceptual-trait structure and US' face-impressions structure.

To test the possibility that US-to-region conceptual-trait correlations (mediator) may be partly explaining the relationship between US-to-region personality correlations (IV) and US-to-region face-judgment correlations (DV), we conducted a mediation analysis. As expected given

the correlational analyses above, the IV was strongly related to both the DV (*B*=0.39, *SE*=0.06, 95% CI [0.28,0.51], *t*=6.49, *p*<.001) and the mediator (*B*=0.32, *SE*=0.06, 95% CI [0.20,0.44], *t*=5.15, *p*<.001). Further, the relationship between the mediator and the DV remained significant even after statistically controlling for the IV (*B*=0.65, *SE*=0.05, 95% CI [0.55,0.74], *t*=13.56, *p*<.001). Most importantly, bootstrapping analyses demonstrated a significant indirect effect, whereby US-to-region conceptual-trait correlations (mediator) partly explained the relationship between US-to-region personality correlations (IV) and US-to-region face-judgment correlations (DV), *B*=0.21, 95% CI [0.13,0.29], *p*<.001 (**Figure 3b** and **Supplementary Table 6a**).

To demonstrate robustness and generalizability of the effects, we reran analyses used a complementary DV. Rather than being asked to judge faces based on 15 trait adjectives directly (e.g., "How likely is this person to be [*adventurous*]?"), a separate group of US participants were asked to judge faces using phrase descriptions as stand-ins for the 15 traits (e.g., "How likely is this person to [*enjoy visiting new places*]?"). A full list of phrase descriptions associated with traits is available in **Supplementary Table 4** and data are taken from Stolier et al. (2020). We again observed strong positive relationships between the IV and the DV (*B*=0.30, *SE*=0.06, 95% CI [0.18,0.43], *t*=4.79, *p*<.001), the IV and the mediator (*B*=0.32, *SE*=0.06, 95% CI [0.20,0.44], *t*=5.15, *p*<.001), and the mediator and the DV (B=0.75, SE=0.04, 95% CI [0.66,0.84], *t*=16.86, *p*<.001), as well as a significant indirect effect (*B*=0.24, 95% CI [0.14,0.34], *p*<.001; **Supplementary Figure 3** and **Supplementary Table 6b**). Note that the IV-mediator relationship does not involve the DV; thus, this result is identical regardless of whether trait words vs. phrases were used for the DV.

Controlling for Language Use and Ethnic Diversity

The results were also robust to the inclusion of the language use and ethnic diversity covariates, again observing positive relationships between the IV and the DV (*B*=0.32, *SE*=0.09, 95% CI [0.14,0.51], *t*=3.42, *p*=.001); the IV and the mediator (*B*=0.32, *SE*=0.09, 95% CI [0.14,0.51], *t*=3.47, *p*=.001); and critically, the mediator and the DV when controlling for the IV (*B*=0.69, *SE*=0.06, 95% CI [0.57,0.81], *t*=11.37, *p*<.001) as well as a significant indirect effect (*B*=0.22, CI 95% [0.10,0.35], *p*<.001). We reran the same analysis using the complementary DV (i.e., face-trait ratings derived from trait phrases) and results were unchanged, again observing a significant indirect effect (*B*=0.24, CI 95% [0.11,0.39], *p*<.001). See **Supplementary Tables 5 & 6** for full statistics.

The larger number of regions available in Study 2 afforded an additional corroborating analysis that clustered regions by primary language using multi-level GEE regression. We only considered sets of same-language regions with sufficient size (≥10 regions per language). This resulted in clusters of 75 English-, 26 French-, 20 Arabic-, 19 Spanish-speaking regions (140 regions in total, 61% of all 231 non-US regions). If the relationships of interest persist even within clusters of regions with the same primary language, that would represent strong evidence that the effects of NEOPI structure via conceptual-trait structure are not confounded by language. To further control for potential confounding effects of language use and ethnic diversity within the same-language groups, we included the same language use and ethnic diversity covariates.

Clustering by language across the 140 regions revealed virtually identical results. Whether using face-trait ratings derived from trait words or trait phrases, US-to-region personality correlations predicted US-to-region face-judgment correlations (trait words: *B*=0.28, *SE*=0.05, 95% CI [0.18,0.38], *Z*=5.40, *p*<.001; trait phrases: *B*=0.39, *SE*=0.09, 95% CI [0.21,0.58], *Z*=4.25, *p*<.001) and predicted US-to-region conceptual-trait correlations (*B*=0.34,

SE=0.13, 95% CI [0.01,0.59], *Z*=2.72, *p*=.007). US-to-region conceptual-trait correlations also predicted US-to-region face-judgment correlations (trait words: *B*=0.69, *SE*=0.03, 95% CI [0.62,0.75], Z=20.66, *p*<.001; trait phrases: *B*=0.71, *SE*=0.11, 95% CI [0.49,0.91], *Z*=6.50, *p*<.001). See **Supplementary Table 7** for full statistics. Mediation analysis and estimates of the indirect effect are not possible with multi-level GEE regression.

In sum, these results suggest that when a region's personality structure was similar to that of the US (e.g., Australia), that region's personality structure could more strongly predict the structure of face-based trait judgments in the US (better than could regions whose personality structure was dissimilar to that of the US, e.g., Syria). Importantly, this relationship was partly explained by how well that region's personality structure could predict conceptual-trait associations in the US, even when controlling for language use and ethnic diversity. The results held regardless of whether participants were asked to judge trait adjectives (e.g., *adventurous*) or to judge phrase descriptions (e.g., *enjoy visiting new places*), which alleviates concerns that correspondence between face impressions, trait concepts, and actual personality structure may be solely due to semantic confounds (i.e., using the same adjectives in all tasks).

Discussion

The present results replicate those of Study 1, showing that a region's face impressions reflect its personality structure. Furthermore, the findings implicate conceptual-trait associations as playing an intermediary role in the relationship between regional variability in personality structure and regional variability in the face impressions structure.

GENERAL DISCUSSION

In two studies, we found that the actual personalities of people in a region are related to how individuals in that region judge others' traits from faces. For example, in a region where

people were more likely to be simultaneously aggressive and intelligent, people in that region were more likely to judge a person with a face appearing more aggressive as more intelligent (vs. other regions) (Study 1). Moreover, the personality structure of regions that were more similar to the US in personality better predicted US conceptual-trait structure and US face-impressions structure (vs. regions less similar to the US in personality) (Study 2). These effects generalized across different ways of assessing face impressions (adjectives or phrases), alleviating the concern of semantic confounds. Together, the findings suggest that people form face-based inferences of others' personalities based on a conceptual understanding of personality that they learn from their regional environment.

The fact that people's face impressions vary depending on the social environment is consistent with evidence for the role of learning in face impressions (Dotsch et al., 2016; Stolier et al., 2020; Sutherland et al., 2020). The role of conceptual associations in guiding face impressions extends previous studies (Stolier et al., 2020; Stolier et al., 2018) by implicating these associations as a mechanism by which region-specific social experience can affect face impressions. Because personality structure (McCrae & Costa, 1997) and face impressions structure show a general consistency across cultures (Todorov & Oh, 2021), variations in these structures have often been overlooked, describing departures from a universal dimensional structure as statistical noise. Our findings bridge variability in personality and variability in face impressions, demonstrating that this 'noise' may contain information about person perception. Regional differences in personality have been suggested to result from various regional factors, such as culture (McCrae & Terracciano, 2005) and socioecological complexity (Lukaszewski, Gurven, von Rueden, & Schmitt, 2017). Future research could examine how these factors affect

not only the actual personalities of local residents, but also how those residents think about personality and judge personality in others.

Our approach was correlational, which afforded a comprehensive assessment across a large number of regions, but limits the ability to make causal claims. While we propose that conceptual associations in the form of lay theories of personality serve as a causal mechanism linking personality in the environment to face impressions, this possibility was not directly tested. The potential roles of other intermediary factors, such as cultural differences in basic face processing (e.g., Caldara, 2017; Hills & Pake, 2013) could be examined in future research. Another limitation is that English-only questionnaires were used to obtain the personality data. Prior work has shown that personality structure is highly similar when the NEOPI questionnaire is administered in English vs. a respondent's native language (McCrae, 2001), and any respondent in our datasets whose data suggested poor English comprehension or confusion was excluded (Johnson, 2005, 2014). Nevertheless, we comprehensively controlled for the potential confounding role of language. Using multiple measures of regions' English proficiency and the linguistic similarity between any given regions' primary languages and corroborating analyses that tested our effects within regions of the same language, we found no evidence that language confounded the results. However, future research could collect personality data in participants' native languages to further understand the potential interplay of regional language and personality in shaping face impressions structure.

It is important to recognize that the present results cannot directly speak to questions on the accuracy of face impressions. Our findings can only speak to how traits are judged from faces, not how they may manifest or be expressed on people's faces. Even if people learned an accurate trait structure, it could only help them accurately infer a person's personality traits when

they already possess accurate information about another trait (which covaries with the trait in question). Thus, accurately learning the structure of personality traits in the social environment need not imply that perceivers can accurately intuit specific traits in others. Future research could explore these questions directly.

In sum, the current results suggest that perceivers use the actual personality structure learned from their social environment to form lay theories about personality, which in turn scaffold the structure of perceivers' face impressions. The findings call for a greater focus on the regional and cultural specificity of face impressions and the role of social experience in how we infer personality from facial appearance.

Author Contribution

All authors developed the study concept and contributed to the study design. DO and JDM performed data analysis and interpretation under the supervision of JBF. DO drafted the manuscript, and all authors provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgements

We thank Rick Dale, Karina Tachihara, and Seongyong Lee for their insight regarding linguistic measures. This work was supported in part by research grant BCS-1654731 (JBF).

Reference

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (2003). Fractionalization. *Journal of Economic growth, 8*(2), 155-194.

Allik, J., Church, A. T., Ortiz, F. A., Rossier, J., Hřebíčková, M., De Fruyt, F., . . . McCrae, R. R. (2017). Mean profiles of the NEO personality inventory. *Journal of Cross-Cultural Psychology, 48*(3), 402-420.

Atkinson, Q. D., Meade, A., Venditti, C., Greenhill, S. J., & Pagel, M. (2008). Languages evolve in punctuational bursts. *Science, 319*(5863), 588. doi:10.1126/science.1149683

Bentz, C., Dediu, D., Verkerk, A., & Jager, G. (2018). The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour, 2*(11), 816-821. doi:10.1038/s41562-018-0457-6

Birkás, B., Dzhelyova, M., Lábadi, B., Bereczkei, T., & Perrett, D. I. (2014). Cross-cultural perception of trustworthiness: The effect of ethnicity features on evaluation of faces' observed trustworthiness across four samples. *Personality and Individual Differences, 69*, 56-61. doi:10.1016/j.paid.2014.05.012

Caldara, R. (2017). Culture reveals a flexible system for face processing. *Current Directions in Psychological Science, 26*(3), 249-255.

Church, A. T., & Katigbak, M. S. (2002). The five-factor model in the Philippines. In *The Five-Factor Model of personality across cultures* (pp. 129-154): Springer.

Collova, J. R., Sutherland, C. A., & Rhodes, G. (2019). Testing the functional basis of first impressions: Dimensions for children's faces are not the same as for adults' faces. *Journal of Personality and Social Psychology, 117*(5), 900-924.

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO Personality Inventory Revised–(NEO-PIR) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Dotsch, R., Hassin, R. R., & Todorov, A. (2016). Statistical learning shapes face evaluation. *Nature Human Behaviour, 1*(1), 0001. doi:10.1038/s41562-016-0001

Education First. (2020). EF English Proficiency Index: Global ranking of countries and regions. Retrieved from https://www.ef.com/wwen/epi

Educational Testing Service. (2021). *TOEFL iBT: Test and score data summary 2020*.

Elfenbein, H. A., Mandal, M. K., Ambady, N., Harizuka, S., & Kumar, S. (2002). Cross-cultural patterns in emotion recognition: Highlighting design and analytical techniques. *Emotion, 2*(1), 75-84.

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A. T., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences, 115*(7), E1690–E1697.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83. doi:10.1016/j.tics.2006.11.005

Gülgöz, S. (2002). Five-factor model and NEO-PI-R in Turkey. In *The five-factor model of personality across cultures* (pp. 175-196): Springer.

Han, C., Wang, H., Hahn, A. C., Fisher, C. I., Kandrik, M., Fasolt, V., . . . Jones, B. C. (2018). Cultural differences in preferences for facial coloration. *Evolution and Human Behavior, 39*(2), 154-159. doi:10.1016/j.evolhumbehav.2017.11.005

Hehman, E., Sutherland, C. A. M., Flake, J. K., & Slepian, M. L. (2017). The unique contributions of perceiver and target characteristics in person perception. *Journal of Personality and Social Psychology, 113*(4), 513-529. doi:10.1037/pspa0000090

Hills, P. J., & Pake, J. M. (2013). Eye-tracking the own-race bias in face recognition: revealing the perceptual and socio-cognitive mechanisms. *Cognition, 129*(3), 586-597. doi:10.1016/j.cognition.2013.08.012

Hirschman, A. O. (1945). *National power and the structure of foreign trade*. Berkeley & Los Angeles, CA: Univeristy of California Press.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008a). *Advances in automated language classification*. Paper presented at the Quantitative Investigations In Theoretical Linguistics (QITL3), Helsinki.

Holman, E. W., Wichmann, S., Brown, C. H., Velupillai, V., Müller, A., & Bakker, D. (2008b). Explorations in automated language classification. *Folia Linguist, 42*(2), 331-354.

Holzleitner, I. J., & Perrett, D. I. (2017). Women's preferences for men's facial masculinity: Trade-off accounts revisited. *Adaptive Human Behavior and Physiology, 274*(3), 1–17.

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 199-209. doi:10.1037/0096-1523.32.2.199

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences, 109*(19), 7241-7244. doi:10.1073/pnas.1200155109

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*(1), 103-129. doi:10.1016/j.jrp.2004.09.009

Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78-89.

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., . . . Coles, N. A. (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour, 5*, 159-169. doi:10.1038/s41562-020-01007-2

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Front Syst Neurosci, 2*(4), 1–28. doi:10.3389/neuro.06.004.2008

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13–22.

Lin, C., Keles, U., & Adolphs, R. (2021). Four dimensions characterizing trait attributions from faces. *Nature Communications, 12*, 5168. doi:10.1038/s41467-021-25500-y

Lukaszewski, A. W., Gurven, M., von Rueden, C. R., & Schmitt, D. P. (2017). What explains personality covariation? A test of the socioecological complexity hypothesis. *Social Psychological and Personality Science, 8*(8), 943-952.

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*(4), 1122-1135.

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to stimulus evaluations. *Behavior Research Methods, 52*, 1428–1444.

McCrae, R. R. (2001). Trait psychology and culture: Exploring intercultural comparisons. *Journal of Personality, 69*(6), 819-846.

McCrae, R. R. (2002). NEO-PI-R data from 36 cultures. In R. R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 105-125): Springer Science+Business Media New York.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52*(5), 509-516.

McCrae, R. R., & Terracciano, A. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*(3), 407.

McCrae, R. R., Yik, M. S., Trapnell, P. D., Bond, M. H., & Paulhus, D. L. (1998). Interpreting personality profiles across cultures: Bilingual, acculturation, and peer rating studies of Chinese undergraduates. *Journal of Personality and Social Psychology, 74*(4), 1041.

Oh, D., Dotsch, R., Porter, J., & Todorov, A. (2020). Gender biases in impressions from faces: Empirical studies and computational models. *Journal of Experimental Psychology: General, 149*(2), 323-342. doi:10.1037/xge0000638

Oh, D., Grant-Villegas, N., & Todorov, A. (2020). The eye wants what the heart wants: Females' preference in male faces are related to partner personality preference. *Journal of Experimental Psychology: Human Perception and Performance, 46*(11), 1328–1343. doi:10.1037/xhp0000858

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087-11092. doi:10.1073/pnas.0805664105

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature, 449*(7163), 717-720. doi:10.1038/nature06176

Piedmont, R. L., Bain, E., McCrae, R. R., & Costa, P. T. (2002). The applicability of the Five-Factor Model in a sub-Saharan culture. In *The five-factor model of personality across cultures* (pp. 155-173): Springer.

Piedmont, R. L., & Chae, J.-H. (1997). Cross-cultural generalizability of the five-factor model of personality: Development and validation of the NEO PI-R for Koreans. *Journal of Cross-Cultural Psychology, 28*(2), 131-155.

Pompei, S., Loreto, V., & Tria, F. (2011). On the accuracy of language trees. *PLoS ONE, 6*(6), e20109. doi:10.1371/journal.pone.0020109

Rolland, J.-P. (2002). The cross-cultural generalizability of the Five-Factor model of personality. In R. R. McCrae & J. Allik (Eds.), *The five-factor model of personality across cultures* (pp. 7–28): Springer Science+Business Media New York.

Rychlowska, M., Miyamoto, Y., Matsumoto, D., Hess, U., Gilboa-Schechtman, E., Kamble, S., . . . Niedenthal, P. M. (2015). Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles. *Proceedings of the National Academy of Sciences, 112*(19), E2429-E2436. doi:10.1073/pnas.1413661112

Schneider, D. J. (1973). Implicit personality theory: A review. *Psychological Bulletin, 79*(5), 294-309.

Simakhodskaya, Z. (2000). *Russian Revised NEO-PI-R: Concordant validity and relationship to acculturation.* Paper presented at the 108th Convention of the American Psychological Association, Washington, DC.

Søren, W., Holman, E. W., & Brown, C. H. (2020). *The ASJP Database (version 19.1).* Retrieved from: https://asjp.clld.org

Soto, J. A., & Levenson, R. W. (2009). Emotion recognition across cultures: the influence of ethnicity on empathic accuracy and physiological linkage. *Emotion, 9*(6), 874.

Stolier, R. M., Hehman, E., & Freeman, J. B. (2020). Trait knowledge forms a common structure across social cognition. *Nature Human Behaviour, 4*, 361–371. doi:10.1038/s41562-019-0800-6

Stolier, R. M., Hehman, E., Keller, M. D., Walker, M., & Freeman, J. B. (2018). The conceptual structure of face impressions. *Proceedings of the National Academy of Sciences, 115*(37), 9210-9215. doi:10.1073/pnas.1807222115

Sutherland, C. A. M., Burton, N. S., Wilmer, J. B., Blokland, G. A. M., Germine, L., Palermo, R., . . . Rhodes, G. (2020). Individual differences in trust evaluations are shaped mostly by environments, not genes. *Proceedings of the National Academy of Sciences, 117*(19), 10218-10224. doi:10.1073/pnas.1920131117

Sutherland, C. A. M., Oldmeadow, J. A., Santos, I. M., Towler, J., Michael Burt, D., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition, 127*(1), 105-118. doi:10.1016/j.cognition.2012.12.001

Sutherland, C. A. M., Young, A. W., Mootz, C. A., & Oldmeadow, J. A. (2015). Face gender and stereotypicality influence facial trait evaluation: Counter-stereotypical female faces are negatively evaluated. *British Journal of Psychology, 106*(2), 186–208.

Todorov, A., & Oh, D. (2021). The structure and perceptual basis of social judgments from faces. In B. Gawronski (Ed.), *Advances in Experimental Social Psychology* (Vol. 63, pp. 189-245). Cambridge, MA: Academic Press.

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

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science, 21*(6), 779-785. doi:10.1177/0956797610371965

Xie, S. Y., Flake, J. K., & Hehman, E. (2018). Perceiver and target characteristics contribute to impression formation differently across race and gender. *Journal of Personality and Social Psychology, 117*(2), 364–385. doi:10.1037/pspi0000160

Zhang, L., Holzleitner, I. J., Lee, A. J., Wang, H., Han, C., Fasolt, V., . . . Jones, B. C. (2019). A data-driven test for cross-cultural differences in face preferences. *Perception, 48*(6), 487-499. doi:10.1177/0301006619849382