Quantifying idiosyncratic and shared contributions to judgment

Joel E. Martinez^{1,2} · Friederike Funk³ · Alexander Todorov¹

Published online: 2 January 2020

© The Psychonomic Society, Inc. 2019

Abstract

Identifying relative idiosyncratic and shared contributions to judgments is a fundamental challenge to the study of human behavior, yet there is no established method for estimating these contributions. Using edge cases of stimuli varying in intrarater reliability and interrater agreement—faces (high on both), objects (high on the former, low on the latter), and complex patterns (low on both)—we showed that variance component analyses (VCAs) accurately captured the psychometric properties of the data (Study 1). Simulations showed that the VCA generalizes to any arbitrary continuous rating and that both sample and stimulus set size affect estimate precision (Study 2). Generally, a minimum of 60 raters and 30 stimuli provided reasonable estimates within our simulations. Furthermore, VCA estimates stabilized given more than two repeated measures, consistent with the finding that both intrarater reliability and interrater agreement increased nonlinearly with repeated measures (Study 3). The VCA provides a rigorous examination of where variance lies in data, can be implemented using mixed models with crossed random effects, and is general enough to be useful in any judgment domain in which agreement and disagreement are important to quantify and in which multiple raters independently rate multiple stimuli.

Keywords Judgment · Multilevel modeling · Variance component analysis · Interrater agreement · Measurement error

A fundamental challenge in the study of human behavior is identifying the shared and idiosyncratic contributions to judgments. Every judgment made in life contains some level of agreement with other people. In many psychology studies, this agreement or consensus is often equated with meaningful variance in the data. However, for any judgment, stable idiosyncratic differences might explain a larger proportion of the variance than is explained by consensus. Hence, it is essential to estimate the shared (consensus) and idiosyncratic contributions to judgments. Different methodological approaches to estimating these contributions can lead to radically different theoretical implications, and essentially can change how we think about the nature of human preferences.

As such, estimating the shared and idiosyncratic contributions to judgments is particularly relevant to the domains of basic science and replicability, and it has important practical policy implications in a variety of domains, ranging from legal sentencing decisions to medical diagnosis. The relative influences of people (i.e., idiosyncrasy) or stimuli (e.g., sharedness) on judgments are key to theoretical debates on finding general principles of, for example, morality (Heiphetz & Young, 2017), intergroup relations (Carter & Murphy, 2017; Xie, Flake, & Hehman, 2019), face preferences (Cunningham, Roberts, Barbee, & Druen, 1995; Grammer & Thornhill, 1994; Hehman, Sutherland, Flake, & Slepian, 2017; Hönekopp, 2006; Langlois et al., 2000; Rhodes, 2006), and object or art preferences (Kurosu & Todorov, 2017; Leder, Goller, Rigotti, & Forster, 2016; Schepman, Rodway, & Pullen, 2015; Vessel, 2010; Vessel, Maurer, Denker, & Starr, 2018). These estimates can also contribute practical knowledge to current discussions of replicability. Often replications are focused on recapturing the effect sizes of different manipulations. Another criterion by which to judge replications is the extent to which similar influences of participants or stimuli (Gantman et al., 2018), or even unexplained variance (Doherty, Shemberg, Anderson, & Tweney, 2013), are recaptured across studies. The interpretation and possibility of replicating or not replicating an effect size changes if the relative influences of people, stimuli, and

✉ Joel E. Martinez
joelem@princeton.edu

✉ Alexander Todorov
atodorov@princeton.edu

¹ Department of Psychology, Princeton University, Princeton, NJ, USA

² Woodrow Wilson School of Public and International Affairs, Princeton University, Princeton, NJ, USA

³ Department of Psychology, University of Cologne, Köln, Germany

unexplained variance underlying the effect change or stay the same. Finally, practical concerns in a variety of domains rely on these estimates. Quantifying how judgments vary by judges or cases can shape understandings of whether the judicial system may be unjustly punitive through inconsistent sentencing (Austin & Williams III, 1977; Forst & Wellford, 1981; Hofer, Blackwell, & Ruback, 1999) and the effectiveness of medical procedures in reliably identifying diagnoses (Shoukri, 2011).

One reason for varied conclusions about the uniqueness or sharedness of judgments is that there is currently no agreed-upon optimal estimation method. This lack of standard has led to the use of many methods and divergent conclusions about the nature of judgmental agreement in many fields. For example, the decision to use a differences in means or a variance decomposition approach can lead to opposite theoretical interpretations of whether judgments of racism are shared within and across racial groups, which has implications for how antiprejudice endeavors should be structured (Martinez & Paluck, 2019). Consequently, the aim of the present study is to identify a general method for disentangling sources of agreement in judgments across a variety of stimuli with different psychometric properties.

All estimation methods require repeated judgments so as to extract the two critical measurements on which idiosyncratic and shared judgments are based: *intrarater reliability* and *interrater agreement*. Unfortunately, very few psychology studies include repeated measures, which are essential for measuring not only intrarater reliability across time, but also the general data quality and potential replicability of findings. In the extreme case in which intrarater reliability is indistinguishable from zero, computing interrater agreement is meaningless, and any results will be spurious.

Standard practices for estimating rater agreement

Judgment studies often use several techniques to estimate rater agreement. The most common measure of this agreement is *Cronbach's alpha*. However, one can obtain very high estimates of alpha with a sufficiently large sample of raters even when the agreement between individual raters is extremely low (Hönekopp, 2006; Kramer et al., 2018). A better measure of rater agreement is the *interrater correlation*, where higher correlations indicate greater consensus. This correlation can be computed between individuals or groups (Kurosu & Todorov, 2017; Ma, Xu, & Luo, 2016; Zebrowitz, Montepare, & Lee, 1993) or between individuals and the group's averaged ratings (Engell, Haxby, & Todorov, 2007; Germine et al., 2015; Zebrowitz, Franklin, Hillman, & Boc, 2013). How these correlations are computed can lead to very different interpretations of interrater agreement. To the extent that the average agreement between individual raters is greater

than zero, correlations between the aggregated ratings across raters are bound to inflate interrater agreement. Such correlations do not necessarily mean high consensus among raters, since the average correlation between any two individual raters could be much lower.

The same issues arise for correlations between participant ratings and measured or computed stimulus attributes. For example, the correlation between face ratings and the physical size of different facial features might be used to estimate the size of the relationship between physiognomic characteristics and social judgments. Whereas some studies correlate stimulus attributes with the group's average ratings (Cunningham et al., 1995), it is also possible to examine how these attributes correlate with individual ratings (Hönekopp, 2006; Jacobsen, Schubotz, Höfel, & Cramon, 2006; Kurosu & Todorov, 2017). Again, the latter correlations are bound to be lower than the former.

The fact that individual-level correlations are lower than aggregated-level correlations suggests much larger idiosyncratic contributions to judgment. However, it is also possible that these low correlations result from measurement error or unreliability of the individual raters. The only way to rule out the latter explanation is to compute intrarater correlations: a measure of how consistent a rater is across two or more time points. Unfortunately, even when repeated ratings are collected, intra- and interrater correlations are typically examined separately. But we cannot assume that these correlations are independent. When the intrarater correlations are positive, we would expect that averaging within raters across repeated measurements would increase their reliability and, consequently, result in higher estimates of the idiosyncratic contributions to judgments. But it is also possible that more reliable individual ratings would result in higher interrater agreement and, consequently, in higher estimates of the shared contributions to judgments. If so, one might derive very different estimates of idiosyncratic and shared contributions, depending on the number of repeated measurements.

Estimating shared versus idiosyncratic contributions

A general method for estimating shared versus idiosyncratic contributions to judgments is to compare the variance components from different clusters in the data (Hehman et al., 2017; Hönekopp, 2006; Kenny, 1996; Leder et al., 2016). Clusters are components of a study that are similar across measurements, such as raters or stimuli or occasions, and are treated as if sampled from a random population (Judd, Westfall, & Kenny, 2012; Westfall, Kenny, & Judd, 2014). The goal of a variance component analysis (VCA) is to quantify and attribute systematic variance to specific clusters by estimating their variance components. Comparing the size of the variance

components can provide information about the importance of different clusters to the ratings. The more variance in a cluster, as measured by variance portioning coefficients (Goldstein, Browne, & Rasbash, 2002), the more systematic differences between instances of that cluster (Shavelson, Webb, & Rowley, 1989).

Within the VCA framework, estimating the shared versus idiosyncratic contributions to judgments is straightforward. Shared contributions are estimated by the stimulus cluster variance. There will be greater stimulus cluster variance if everyone agrees on the judgment of each stimulus and if the stimuli are distinct on the judged dimension (e.g., if everyone rates Stimulus 1 as beautiful and Stimulus 2 as not). Idiosyncratic contributions are estimated by the variance attributed to the Rater $\times$ Stimulus cluster. The latter cluster represents idiosyncratic taste, as it measures differences in ranking preferences. For example, this variance will be large if Participant A prefers Stimulus 1 more than Stimulus 2, as compared to Participant B, who has stronger but opposite preferences.

The role of the rater cluster is controversial. Although it represents individual differences (e.g., personality, mood, or even subjective construal of the experimental materials; Paluck & Shafir, 2017), it is unclear whether it should count as a source of idiosyncratic contribution for judgment (Hönekopp, 2006). For example, if Participants A and B rate Stimuli 1, 3, 2, in the same preference order, but Participant A's ratings are higher than Participant B's, this mean difference does not necessarily count as idiosyncratic taste, because the participants would still share preference rankings. On the other hand, the participant with higher mean ratings may genuinely like the stimuli more, and this greater liking can lead to different behaviors.

Once the variance components are computed, one can create variance ratios that represent the shared and idiosyncratic contributions to judgments. Hönekopp (2006) called these ratios *beholder indices*. Following Hönekopp, we calculated two beholder indices, to take into account the ambiguity of the interpretation of the rater cluster. The first index, b_1 , ignores the role of the rater cluster, and is simply the ratio of the Rater $\times$ Stimulus variance and the sum of the latter and the stimulus variance (see Formula 2 below). The second index, b_2 , includes the rater variance (see Formula 3 below). Ratios higher than .50 indicate stronger idiosyncratic than shared contributions to judgments.

Recently, Germine et al. (2015) introduced an intuitive measure of shared versus idiosyncratic contributions to judgments. The measure is a correlation index that directly combines inter and intrarater correlations and is similar to the beholder index. The correlation index partitions out the amount of interrater agreement from the variance accounted for by repeated judgments, the intrarater correlation, thereby treating the intrarater correlation as the ceiling of meaningful variance and any further variance as noise. This index

suggests that there are more shared contributions if both the interrater correlation and the intrarater correlations are high and more idiosyncratic contributions if interrater correlation is low but the intrarater correlation is high.

Here we tested whether the VCA provides sensible estimates across varied stimulus contexts and how sensitive it is to measurement error. We also compared estimates of shared and idiosyncratic contributions, using beholder and correlation indices. In sum, we sought to find the kind of analysis that best quantifies shared and idiosyncratic contributions to judgments, in order to develop guidelines on best practices.

Overview of studies

In Study 1, we used three sets of stimuli rated on their beauty, a domain in which the psychometric properties of different stimuli are well characterized. The stimuli provided extreme test cases for measuring idiosyncratic and shared judgments: randomly generated complex color patterns (close to zero intrarater reliability and zero interrater agreement), a set of novel objects (high intrarater reliability and close to zero interrater agreement; see Kurosu & Todorov, 2017), and faces (high intrarater reliability and high interrater agreement). The first two cases are particularly important, because they allowed us to distinguish between proper and improper estimation (e.g., estimating meaningful shared components of judgments in cases in which the interrater agreement is zero). We found that the VCA is able to capture the psychometric properties of each stimulus set.

In Study 2, we moved beyond aesthetic ratings, to report a simulation study designed to (1) show the generality of these methods and (2) examine the relative advantages of increasing either sample or stimulus size, or both, to estimations. Currently, a lot of attention has been directed at larger sample sizes as being indicative of better estimate precision, power, and replicability (Maxwell, Kelley, & Rausch, 2008). The number of stimuli has also been given recent consideration as another important design choice (Judd et al., 2012; Westfall, Judd, & Kenny, 2015; Westfall et al., 2014), particularly when using too few stimuli can obscure whether results are due to the specific stimuli chosen or the sample (Wells & Windschitl, 1999). Here we found that the number of stimuli can often (but not always) have a larger impact than the number of raters on the precision of correlations and VCA estimates.

In Study 3, we further explored how multiple repeated measures of aesthetic judgments of objects and faces affect estimates of idiosyncratic and shared judgments. With more repeated measures, the intra- and interrater correlations both increased nonlinearly at different rates and were not independent from each other. Any index created from these

correlations would therefore change with more repeated measures. The VCA estimates, however, stabilized with more measures.

Study 1

The aim of Study 1 was to test whether the VCA provides sensible estimates about the shared and idiosyncratic contributions to beauty judgments of stimuli with varying psychometric properties: faces (high interrater agreement and intrarater reliability), objects (low interrater agreement, high intrarater reliability), and patterns (low interrater agreement and intrarater reliability). First, we validated that the intra- and interrater correlations reflected the hypothesized stimulus properties. Then we checked whether the VCA arrived at appropriate conclusions. Finally, we qualitatively compared the beholder index estimates with the correlation index estimates.

Method

Participants

One hundred twenty-four participants were recruited using Amazon Mechanical Turk, in accordance with the Princeton University Institutional Review Board (Protocol: 0000007301), to evaluate either faces ($N = 40$), objects ($N = 40$), or patterns ($N = 44$) (Fig. 1). For sample size, we used past research on the reliability of face (Oosterhof & Todorov, 2008) and object (Kurosu & Todorov, 2017) judgments, in which the minimum sample needed was about 18–25 participants to obtain high reliability ($r = .9$ for faces, $.5$ for objects). For the patterns, for which little previous research had focused on reliability, we decided a priori on 40 participants, which was larger than the number needed for faces or objects.

Fig. 1 Examples of the stimuli: Faces, objects, and patterns

Stimuli

We randomly selected 50 faces (half female) from the Karolinska Directed Emotional Faces database (KDEF; Lundqvist, Flykt, & Öhman, 1998). The pictures depicted white men and women with neutral facial expressions. We randomly selected 50 3-D objects, created with Grasshopper algorithms (<http://www.grasshopper3d.com>), from the set developed by Kurosu and Todorov (2017). We generated 50 color matrices in Python. Specifically, each matrix consisted of 100×100 square blocks, in which the color of each block was randomly generated using the RGB color system (see Fig. 1).

Procedure

Participants completed a self-timed rating task in which they judged “how beautiful is this {face, object, image},” using a scale from 1 (*not at all*) to 7 (*extremely*). Participants rated each stimulus twice, once in two successive blocks. The stimulus order was randomized for each participant and within each block.

Descriptive analysis

The intrarater correlation, or test–retest reliability, examines the consistency of raters’ ratings with themselves across repeated measurements (Gwet, 2014). We computed Pearson correlations between the two time points for the same individual. The interrater correlation measures the consensus between raters. For descriptive purposes, here we compute this value with pairwise Pearson correlations between raters as well as correlations between a rater and the group’s average ratings (RtG).

Variance component analysis

Implementation A recent approach for estimating variance components is through maximum likelihood (ML) and restricted maximum likelihood (REML) estimation in a mixed regression model (Nakagawa & Schielzeth, 2010; Searle, Casella, & McCulloch, 2006). For this analysis, we used the Linear Mixed Effect 4 (lme4) package, version 1.1.11 in R (Bates, Mächler, Bolker, & Walker, 2015), to estimate the VCA. Since the stimuli were rated twice by multiple participants, our models were cross-classified (Schielzeth & Nakagawa, 2013). The random effects included the rater, stimulus, and Rater $\times$ Stimulus, as well as the block, Block $\times$ Rater, and Block $\times$ Stimulus, clusters, and were estimated using REML, which is comparable to a random-effects analysis of variance (ANOVA; Corbeil & Searle, 1976; Searle et al., 2006). To minimize convergence issues, the models

were optimized using the “bobyqa” option in lme4 and were run for a maximum of 500,000 iterations.

To test an additional analytic option, we also estimated the VCA using random-effects ANOVA from the (“VCA”) package, version 1.3.2 in R (Schuetzenmeister, 2016), which provides sum-of-squares (SS)-derived variance components. The VCA estimates were similar across (restricted) maximum likelihood and ANOVA implementations (Supplementary Fig. 1), suggesting that the choice of VCA implementation may be informed by the desire for further functionality, such as bootstrapping, in the relevant software packages, by the size of the data as variance estimates are partially pooled in mixed models (Gelman & Hill, 2007), or by issues with convergence (see Appendix C).

Variance partitioning coefficient (VPC) The equation for the VPC is

$$VPC = \frac{\sigma_{\text{cluster}}^2}{\sigma_{\text{cluster}}^2 + \sigma_{\text{residual}}^2} \quad (1)$$

which is the ratio of a cluster’s variance to the total variance (Goldstein et al., 2002). The $\sigma_{\text{cluster}}^2$ represents the between-instances-of-a-cluster variance, whereas the $\sigma_{\text{residual}}^2$ represents within-cluster variance. When the VPC is closer to one, the between-cluster variance will explain most of the variance and the within-cluster variance will be low. When the VPC is closer to zero, there is no between-cluster variance, and thus most of the variance is within-cluster. The analyses in this study contained random intercepts only, suggesting that the VPCs in these analyses can be interpreted as intraclass correlations (ICCs; Goldstein et al., 2002; Nakagawa & Schielzeth, 2010).

Beholder indices Following Hönckopp (2006) and using the relevant VPCs, we computed two beholder indices. The first index, b_1 , ignores the role of the rater cluster variance (Formula 2), whereas the second, b_2 , includes the rater variance as an idiosyncratic taste (Formula 3).

$$b_1 = \frac{\sigma_{\text{Rater} \times \text{Stimulus}}^2}{\sigma_{\text{Rater} \times \text{Stimulus}}^2 + \sigma_{\text{stimulus}}^2} \quad (2)$$

$$b_2 = \frac{\sigma_{\text{rater}}^2 + \sigma_{\text{Rater} \times \text{Stimulus}}^2}{\sigma_{\text{rater}}^2 + \sigma_{\text{Rater} \times \text{Stimulus}}^2 + \sigma_{\text{stimulus}}^2} \quad (3)$$

Typically, beholder indices are reported as is; however, we subtracted them from 1 for the purpose of standardizing value interpretation and comparison with the correlation index (see below). After this transformation, a beholder index greater than .5 would be interpreted as evidence that shared taste contributes more to judgments than does idiosyncratic taste, whereas a beholder index smaller than .5 would indicate a greater role for idiosyncratic taste.

Correlation index The correlation index (Germine et al., 2015) assumes a zero-sum relationship between the intra- and interrater correlations. Germine and colleagues calculated pairwise interrater correlations, squared them, and then averaged them to obtain “the proportion of ratings across faces that overlap between two typical individuals.” They also calculated the mean of the squared pairwise intrarater correlations. The formula used to obtain an estimate of *shared preferences* was

$$\text{Shared} = \frac{E[r_b^2]}{E[r_w^2]} \quad (4)$$

where r_b represents the interrater correlations, r_w the intrarater correlations, and E is the expected value or average. The formula for *idiosyncratic preferences* is simply $1 - \text{Shared}$. These formulas as written can accommodate only data sets in which pairwise interrater correlations are uniformly positive—that is, data in which everyone agrees to some extent (which was the case for the preselected stimuli Germine and colleagues used). For data in which disagreement between raters exists, this formula would estimate shared contributions when there were none (Supplementary Fig. 1). This is due to the squaring procedure: squaring negative correlations to transform them into variance-explained metrics (r^2) would impute agreement from disagreement between individuals. We therefore made one slight modification in order to account for disagreement, so that the modified *shared preferences* are now

$$\text{Shared}_{\text{modified}} = \frac{E\left[r_b^2 * \frac{r_b}{|r_b|}\right]}{E\left[r_w^2 * \frac{r_w}{|r_w|}\right]} \quad (5)$$

In simpler terms, multiplying the variances by the correlation divided by its absolute value means that if the original correlations were negative, the resulting variance is also negative. Although this modification helps account for disagreement, it may not be theoretically optimal, as the meaning of negative variances is inherently ambiguous (Nakagawa & Schielzeth, 2010). For simplicity, we only report estimates of shared preferences; a value greater than .5 would suggest greater contributions from shared taste.

Results

Validating psychometrics The correlational patterns matched the expected stimulus psychometrics. Face ($r_{\text{intra}} = .62$, 95% CI[.52, .71]) and object ($r_{\text{intra}} = .49$, 95% CI[.39, .59]) judgments both showed high intrarater reliability, whereas the pattern judgments had almost none ($r_{\text{intra}} = .067$, 95% CI[.02, .11]) (Fig. 2A). Faces also showed interrater agreement (r_{inter}

Fig. 2 Correlation and estimation metrics across stimulus types. The columns in all four figures represent data from beauty judgments of faces, objects, and patterns. (A) Descriptive correlations. The x-axis represents the intrarater, interrater, and rater-to-group (RtG) correlations. The y-axis represents the estimates (error bars represent 95% confidence intervals). Values above .5 (black horizontal lines) represent more shared contributions, and values less than .5 represent more idiosyncratic contributions. (B) Variance partitioning coefficients (VPCs). The x-axis represents clusters important for the shared versus idiosyncratic

contributions (rater, stimulus, and Rater $\times$ Stimulus). For block-related clusters, see Supplementary Fig. 1. The y-axis represents how much variance each cluster explains. (C) Beholder indices: Estimates of the b_1 index (black) and the b_2 index (dark grey), whose values are interpreted similar to the correlation indices; values above .5 (black horizontal line) mean more of a shared contribution, and values less than .5 mean more of an idiosyncratic contribution. (D) Correlation indices: Proportions of shared contribution estimates from the correlation indices

= .34, 95% CI[.32, .35]), whereas objects ($r_{\text{inter}} = .003$, 95% CI[−.02, .03]) and patterns ($r_{\text{inter}} = -.002$, 95% CI[−.01, .006]) did not. When examining the rater-to-group (RtG) correlations, faces showed a boost in agreement ($r_{\text{RtG}} = .57$, 95% CI[.49, .65]) that was not present in the objects ($r_{\text{RtG}} = .039$, 95% CI[−.06, .14]) or patterns ($r_{\text{RtG}} = .004$, 95% CI[−.03, .05]), congruent with the signal-boosting effects of between-rater aggregation.

Estimation The VPC metrics suggest a pattern of results consistent with the psychometrics. Face judgments were largely explained by the rater VPC (i.e., potentially idiosyncratic variance) and Rater $\times$ Stimulus VPC (i.e., idiosyncratic variance), yet they also contained a sizable stimulus VPC (i.e., shared variance) (Fig. 2B). The stimulus VPC was correctly close to zero for the object and pattern judgments. The object and pattern data were largely explained by rater and Rater $\times$ Stimulus VPCs. However, the pattern data only contained rater variance, suggesting that unreliable judgments can contribute to this particular variance component. The findings also show that observing large idiosyncratic contributions through the Rater $\times$ Stimulus VPC is dependent on collecting reliable judgments, which was the case for the face and object data (Figs. 2A and 2B). Because the beholder indices are derived from the variance components, they reflect the VPC results but provide a simpler summary metric (Fig. 2C).

The correlation index appropriately estimated that faces contained more idiosyncratic judgments (i.e., 64% of the reliable variance) and that there were close to no shared judgments in the object and pattern data (Fig. 2D). Although this index provided reasonable estimates, it is important to examine how consistent they are with the VCA estimates. If one were to consider the correlation index and beholder index values as being comparable, then it would seem the correlation index estimates more idiosyncratic contributions by about 0.1 units in the face data, and less in the object or pattern data. This might be due to how each method deals with repeated measures. Mixed models naturally account for repeated measures, but for the correlation index one has to average repeated measures in the process of calculating pairwise interrater correlations. In our case, we averaged them within raters first, then calculated agreement, likely increasing the idiosyncratic contributions by improving rater signal.

Discussion

After validating the hypothesized psychometric properties of beauty judgments for different kinds of stimuli, the analyses showed that the VCA metrics provided meaningful estimates. The correlation index, which only combines intra- and interrater correlations, provided estimates similar to those for the beholder indices, which are derived from the VCA. Yet,

the averaging procedures required in order to address repeated measures in the correlation index may have led to an overestimation of idiosyncratic contributions (see Supplementary Fig. 17). We returned to this issue in Study 3. The VCA estimates relate to the psychometrics (i.e., stimulus VPC to interrater agreement, and Rater $\times$ Stimulus VPC to intrarater reliability), although the rater VPC remains ambiguous, since even unreliable pattern judgments contributed to this variance.

Study 2

The previous study showed that the VCA can provide reasonable estimates for stimuli that represent edge cases of psychometric properties. In Study 2, we conducted a simulation analysis to directly manipulate the magnitudes of reliability and agreement, for greater control of the investigated psychometric space, and examined the additional impacts of stimulus and sample size on estimations.

Method

Simulation procedure The simulation procedure was as follows: First, the data sets were simulated from a prespecified variance–covariance matrix using a multivariate normal distribution (Genz et al., 2018) with differing numbers of stimuli and raters. From these data sets, we estimated VPCs. To take advantage of the high level of control over psychometric properties, a variance–covariance matrix was created with varying levels of interrater agreement and intrarater reliability. The matrix structure was compound symmetric, such that the diagonal contained all 1s and the off-diagonals contained all interrater correlations between all raters (interrater agreement) and between each rater’s first and second blocks (intrarater reliability) (Supplementary Table 1). To input realistic off-diagonal values that could mimic those in a real data set (i.e., not everyone is correlated in the exact same way), an algorithm was implemented that added noise around the specified agreement and reliability correlations. Importantly, this algorithm ensured that the matrix stayed positive-definite (Hardin, Garcia, & Golan, 2013).

For researchers interested in using the simulation, we have created a web application to which the variables listed in the next section (and more) can be submitted in order to create and download a simulated data set: <https://joelem.shinyapps.io/SimulateRatingData/>.

Simulation variables The simulation manipulated the following variables: number of raters (nine levels: 20 to 100, in increments of 10), number of stimuli (six levels: 10 to 60, in increments of 10), average interrater agreement (ten levels: 0 to .9, in increments of .1), and average intrarater reliability (ten levels: 0 to .9, in increments of .1). Variables that stayed

consistent were number of repeated measures (2), the standard deviation of participant averages (value of 1), and number of simulations to run per each combination of variables ($n = 120$). The ratings were continuous integers from 1 to 9. We ran 120 simulations per each combination of variables in order to obtain better estimates of their effects due to the stochastic nature of the simulation procedure. Importantly, nonmeaningful combinations of variables were ignored (e.g., when values of interrater agreement were greater than the intrarater reliability). Altogether, the combinations of variables led to 356,400 simulated data sets.

Analysis For each data set, we calculated the simulated reliability and agreement correlations to validate that our procedure’s output data matched the specified psychometrics, and we calculated the VPCs in order to estimate the shared and idiosyncratic contributions, as specified in the previous studies. To simplify reporting, we report psychometric correlations and VPC estimates from the combination of most number of stimuli and raters and therefore represent the most precise estimates. To show the relative and combined impacts of the number of stimuli and raters on estimate precision, we first calculated estimation variance for every combination of reliability and agreement. Then we visualized average trends using a smoothed loess curve. The trend lines pooled across combinations of levels of reliability and agreement ($N = 55$); therefore, each combination of N raters and N stimuli contained 55 variance estimates. In this case, larger variance estimates indicate less precision. Collapsing across agreement and reliability can hide more nuanced patterns; therefore, interested readers can find the full results across all combinations of stimulus, sample size, reliability, and agreement in the supplementary material (Appendix B).

For completeness, we also computed beholder and correlation indices from the simulations (Appendix B). Both indices suffered from instability at low levels of reliability and interrater agreement, the correlation index produced implausible estimates (Supplementary Figs. 7 and 9). Creating a ratio of variance components (i.e., beholder indices) led to more stable estimates than did a ratio of intra- and interrater correlations (i.e., correlation index) (Supplementary Figs. 8 and 10).

Results

We initially checked that our simulations captured the specified psychometric properties. The average simulated reliability increased linearly with the specified reliability, regardless of level of agreement (Fig. 3A). The same occurred for agreement, regardless of the level of reliability (Fig. 3B). These findings validate that our procedure created data sets that matched the desired psychometrics. However, it should be noted that the ability to faithfully capture these properties

Fig. 3 Simulated interrater agreement, intrarater reliability, and their precision. (A) The y-axis represents the simulated reliability estimates from 100 raters and 60 stimuli, and the x-axis represents the specified reliability (0 to .9). The interrater agreement is represented by rainbow colored gradient from the least agreement (0 = red) to the most (.9 = pink). The columns represent 20 or 100 participants, whereas the rows represent 10 or 60 stimuli. (B) The same plot as for reliability, except that the

rainbow gradient now represents different levels of reliability. For both plots, dots represent the estimates from one simulated data set, and lines represent the averages. (C) The y-axis represents the variance in estimations of the reliability across numbers of raters (x-axis) and numbers of stimuli (least [orange] to most [pink]). (D) Plot of variance for the agreement estimates. Lines represent smoothed loess averages, and more variance means less precision

depended on either the sample or stimulus size. For reliability, increasing the number of stimuli led to a greater improvement in precision than did increasing the sample size, although both make an appreciable impact (Fig. 3C). The precision improvement seems to level off after 60 raters or 30 stimuli. Notably, having only 10 stimuli was associated with the worst precision, such that even increasing the sample size to 100 raters would only make the precision as good as having 20 stimuli and 20 raters. For agreement, the number of stimuli had the most impact on precision, such that having over 30 stimuli led to the best estimates (Fig. 3D).

After validation, we examined the impacts on VPC estimates of different levels of reliability and agreement alongside increasing numbers of raters or stimuli. The rater VPC was not impacted by either reliability or agreement (Fig. 4A), but its precision was improved more by increasing the sample size than by increasing the stimulus number, as shown by the greater drop in variance, as long as the design contained more than ten stimuli (Fig. 4D). The stimulus VPC increased with levels of agreement, with no impact from levels of reliability, suggesting that the stimulus VPC is directly linked to interrater agreement (Fig. 4B). Increasing the number of

stimuli improved estimates more than increased sample size, and required more than ten stimuli for a reasonable precision (Fig. 4E). Finally, the Rater $\times$ Stimulus VPC was dependent on both agreement and reliability, providing a more nuanced understanding of these links than was concluded in Study 1 (Fig. 4C). Greater reliability increased the VPC, while greater agreement simultaneously decreased the VPC. A larger sample size improved the precision of this VPC estimate more than did a larger number of stimuli.

Discussion

The simulations in this study clarified the relationship between psychometric correlations and the variance components and the impact of sample and stimulus set size on the precision of those estimates. Whereas the rater VPC was not related to agreement or reliability, the stimulus VPC was associated with agreement and the Rater $\times$ Stimulus VPC with both agreement and reliability. For many of the agreement- or stimulus-related estimates, an increase in the number of stimuli led to larger decreases in estimation variance (i.e., better precision) than increases in sample size. The rater-related estimates depended

Fig. 4 Simulated variance partitioning coefficients (VPCs) and their precision, for (A,D) rater VPC, (B,E) stimulus VPC, and (C,F) Rater $\times$ Stimulus ($R \times S$) VPC. (A–C) The x-axes represent the specified reliability, whereas the y-axes represent the respective VPC estimates. Columns within each plot represent numbers of participants, and rows represent numbers of stimuli. Agreement is represented by a rainbow

gradient, such that no agreement is orange and the most agreement is pink. Dots represent the estimates from simulated data sets, and lines represent averages. (D–F) The y-axes represent the estimation variance of each specific VPC across numbers of raters (x-axis) and numbers of stimuli (least [orange] to most [pink]). Lines represent smoothed loess averages

on both the sample size and the number of stimuli. However, the precision advantage of stimulus number waned, whereas the sample size had more impact on precision for designs with more than ten stimuli. These asymmetries in the contributions to estimation point to the possibility that increasing the stimulus set size might be a better experimental design trade-off than simply striving for greater sample size.

The simulations made three simplifying assumptions that should be noted when interpreting these results. First, each rater's mean was assumed to be similar across repeated measures, simply because there are many ways a mean can change across time, and that change will depend on the topic of study. This assumption is consistent with the data from Study 1: The average difference between each rater's mean on Block 1 and the mean on Block 2 was small for the faces (difference = .122), objects (difference = .04), and patterns (difference = -.08). Second, the specified interrater correlations were the same across repetitions—for example, the average interrater correlation was the same for Block 1 as for Block 2. This assumption was also consistent with the data from Study 1: The difference between the average interrater correlation on Blocks 1 and 2 was small for the faces (difference = .003), objects (difference = .002), and patterns (difference = .006).

Third, because the psychometric specifications were at the level of raters, the simulations do not directly manipulate stimulus variability (i.e., specifying different kinds of stimuli), and therefore represent a restricted stimulus range.

Study 3

The objective of this study was to test whether the estimates of idiosyncratic and shared variance would change with more than two repeated measures, and further, how aggregation or the simple inclusion of those repeated measures impacts estimation. For example, when it comes to estimating inter- or intrarater correlations from repeated measures, the common practice is to average within raters in order to increase the signal-to-noise ratio and fulfill the requirement of independent observations (Bakdash & Marusich, 2017). Conversely, averaged data reduce the power of regressions by condensing the amount of input data. A benefit of mixed-effect models is that they can handle multiple measurements by properly specifying the dependencies. Given these possibilities in data preprocessing and collection, it is important to assess how keeping

more than two repeated measurements intact or aggregating them affects the stability of shared and idiosyncratic estimates.

We focused on judgments of faces and objects, because these were the reliably rated stimuli in Study 1. Having two types of stimuli with different levels of rater agreement, the analysis tracked the changes in rater reliability, agreement, and VCA estimates when incrementally increasing the number of repeated stimuli up to ten ratings per individual. Additionally, we manipulated the way that these repeated measures were preprocessed: by aggregating or simply including the additional measures. For evidence of the generalizability of these results to other judgments besides beauty, we applied the same analysis pipeline to three more judgments (approachable, dangerous, and likable) on the same faces and objects. Although the absolute values of the estimations differed across judgments and stimuli, the effects of additional repeated measures were consistent across dimensions (Supplementary Figs. 11, 12, 13, 14, and 15).

Method

Forty-four participants were recruited from the Princeton undergraduate pool for course credit. The data were a reanalysis of data collected for Kurosu and Todorov (2017), Study 1. Participants judged either 66 faces from the KDEF (Lundqvist et al., 1998) ($n = 19$) or 66 objects ($n = 25$) on their beauty.

The experiments were run on in-lab CRT monitors and were developed using PsychoPy (Peirce, 2007). The task was a simple rating task, in which each trial contained a 500-ms blank screen followed by 500 ms of a crosshair, and ended with a self-timed question probe. The probe displayed the stimulus (20° visual angle) above a question (“How beautiful is this {object, face}?”), and a 1 (*not at all*) through 9 (*extremely*) rating scale below the question. The 66 stimuli were presented ten times each across ten blocks. For each participant, the presentation order of the stimuli was randomized for each block, and participants were given the chance to take a quick break between blocks.

Descriptive correlational analyses We estimated the correlation between two blocks within each rater and averaged across raters to get an overall measure of the sample’s reliability as we increasingly used more data. To test the impact of more than two measures (ten, in this case) on the intrarater correlations, we took a sequential-averaging approach. This approach avoided violating the independence requirement of correlations and assured that the results reflected many possible block combinations for generalizability; see Appendix C for details.

We calculated pairwise correlations between raters on the same block(s) and took the average of all pairwise correlations in order to obtain an overall measure of the sample’s

consensus as we aggregated more data. Here, we also took a sequential-averaging approach (see Appendix C).

To calculate the correlation between an individual rater’s data and the average of the rest of the group, we employed a leave-one-out approach, in which raters’ data were isolated and correlated to the group average of the rest of the raters. This process was repeated for each rater, using the same averaging stepwise approach above, as we added more data from the ten blocks. The point estimates and confidence intervals at each step were derived from the mean of all the rater’s correlations with the group.

Although the analyses above describe how the intra- and interrater correlations independently shift with aggregated data, it is also important to examine whether the relationship between them shifts, too. At each step of the sequential averaging, we correlated the individuals’ intrarater correlation with their rater-to-group correlation.

Results

First we report the impact of increased data averaging on the intrarater, interrater, and rater-to-group correlations for faces and objects (Fig. 5). The correlation values (r_N) represent the average correlation across N blocks; the 95% confidence intervals were calculated from individuals (intrarater and rater to group) and from block combinations (interrater).

Consistent with the signal-increasing goal of aggregating the data, more repeated measures increased the intrarater correlations, for both faces ($r_2 = .59$ [.49, .69], $r_{10} = .85$ [.77, .92]) and objects ($r_2 = .57$ [.48, .66], $r_{10} = .83$ [.76, .90]) (Figs. 5A and 5B). The increase was nonlinear and was still increasing at ten blocks, indicating that the intrarater correlation may require more than ten repeated measures to stabilize.

The interrater correlations for faces exhibited an increase with repeated measures ($r_2 = .38$ [.38, .39], $r_{10} = .49$), even showing an increasing trend at ten blocks, suggesting that interrater correlations, too, may require more than ten repeated measures to stabilize. The interrater correlations for objects showed a stable pattern ($r_2 = .02$ [.01, .02], $r_{10} = .02$), indicating that when the average agreement is around zero, it will remain around zero, irrespective of any type of aggregation or averaging. It is important to note that the zero correlation is largely due to similar amounts of positive and negative interrater correlations, as we will see in the next analysis. That is, the average pairwise correlation obscures variations in agreement and disagreement between pairs of raters.

The rater-to-group correlations were greater than the pairwise correlations and exhibited similar changes across block aggregation for faces ($r_2 = .60$ [.52, .67], $r_{10} = .68$ [.62, .74]), suggesting once again that it may take more than ten repeated measures to stabilize this statistic. The rater-to-group correlations remained relatively stable for objects ($r_2 = .07$ [−.03, .17], $r_{10} = .10$ [−.03, .22]). Since these point

Fig. 5 Average correlations and relationships between the intrarater and rater-to-group (RtG) correlations across repeated measures. Panels A and C represent the face data, and panels B and D the object data. (A) The x-axis represents the number of blocks used in the sequential-averaging approach. Columns display changes in the intrarater, interrater, and RtG correlations, respectively. Each point in the Intra and RtG columns represents an individual participant; each point in the Inter column represents the average pairwise correlation for every instance of N

estimates were at the level of the individual rather than block combinations, one can see why the pairwise interrater correlations were low for object beauty (Fig. 5B). Many of the individuals were negatively correlated with each other, indicating that average estimates of pairwise interrater correlations can hide the extent of disagreement between individual raters. An average pairwise interrater correlation of zero could occur because every rater is orthogonal to the rest or because positive and negative correlations cancel out.

To examine the relationship of the aggregated measures on the intra- and interrater correlations, we used the rater-to-group correlation instead of the pairwise interrater correlations in order to have one estimate of agreement at the level of the rater. For faces, there were high correlations between intrarater and rater-to-group correlations using just two blocks ($r_2 = .82$; Fig. 5C). Raters who had high intrarater correlations agreed more with the group: If agreement exists, intrarater reliability improves with an increasing number of blocks and is coupled with increasing interrater agreement. The correlations decreased as more blocks were aggregated ($r_{10} = .59$), due to truncation in the range of variance, in which the data progressively clustered toward higher reliability and agreement with more blocks (Goodwin & Leech, 2006). For objects, the correlation was weak and negative (Fig. 5D): If there is little interrater agreement to begin with, intrarater reliability improves with repeated measures but is not coupled with increased interrater agreement.

blocks combined. The red dashed lines are the averages; the solid lines are the 95% confidence intervals. (B) The same plot for the object data. (C) The x-axes represent the intrarater correlations, the y-axis represents the RtG correlation, and the columns represent the number of blocks used. Each point represents the average intrarater and RtG correlations across all N -block combinations per participant. The correlation between the intrarater and RtG correlations is given in the bottom part of each plot. D) The same plot for the object data

Variance component analyses We developed two data-preprocessing streams: *sequential addition* and *sequential averaging*. The former merely included, whereas the latter aggregated, an increasing number of repeated measures as data for the VCAs. These analyses were implemented using both ANOVA and linear mixed models. Due to similar results, we only report the linear models. See Appendix C for additional details on the analysis streams, implementations, and issues with convergence in mixed models.

Variance partitioning coefficients Using the sequential-addition approach, the average VPC per cluster stayed consistent (Figs. 6A and 6C), yet showed less variance for increasing blocks, resulting in more stable estimates. The range of some VPC estimates was overlapping, especially when only using two blocks. For example, in the face ratings, the stimulus and Rater $\times$ Stimulus VPCs highly overlapped, suggesting that either could be estimated as greater than the other, depending on which two blocks were sampled in particular. The block, Block $\times$ Stimulus, and Block $\times$ Rater VPCs did not explain much of the variance in these data.

Using the sequential-averaging approach, repeated measures had the same variance-reducing effect on the estimates (Figs. 6B and 6D). They also exhibited the same relative VPC patterns as in the sequential-addition analyses. However, the average VPC estimates

Fig. 6 Variance partitioning coefficients (VPCs) across repeated measures. The left plots (A,B) are for face judgments, and the right plots (C,D) for object judgments. The panels represent estimates from either the sequential-addition or sequential-averaging approach. The x -

axes represent the numbers of blocks used in the analysis. For the VPCs, there are six clusters: rater, stimulus, Rater $\times$ Stimulus, block, Block $\times$ Rater, and Block $\times$ Stimulus

increased nonlinearly as more blocks were averaged. This increase specifically occurred in clusters that already explained lots of variance: those for rater, stimulus, and Rater $\times$ Stimulus. The increase can be explained by directly examining the variance component estimates (Supplementary Fig. 14). When one averages repeated measures before running this analysis, the affected variances are the residuals and block-related clusters: Both decrease with more averaged data.

Beholder indices Since residuals are not used in these estimations, the sequential-averaging and sequential-addition analyses provided similar results. The mean beholder index estimates stayed consistent with increasing blocks (Fig. 7). Overall, beholder indices became more stable with an increase in repeated measures.

For faces, beauty judgments had a mean b_1 of about .5, replicating previous estimates of equal shared and idiosyncratic contributions (Hönekopp, 2006). However, the individual

Fig. 7 Beholder indices across repeated measures. The left plots (A,B) are for face judgments, and the right plots (C,D) for object judgments. The panels represent estimates from either the sequential-addition (A,C) or sequential-averaging (B,D) approach. The x -axes represent the

numbers of blocks used in the analysis. Black is b_1 , which excludes the rater variance, and gray is b_2 , which includes the rater variance. Values above .5 represent more shared contributions, and those below .5 represent more idiosyncratic contributions

b_1 estimates showed substantial variance when fewer repeated measures were used. For example, with only two measures, the estimate range was .39 to .62, depending on which two blocks were used, suggesting that b_1 could be under- or overestimated depending on how many measures were taken. The b_2 indices were well below .5 and below the b_1 estimates, suggesting a high amount of between-rater variance. The objects show a very different pattern, which shows very tightly linked b_1 and b_2 estimates well below the halfway point, indicating highly idiosyncratic ratings.

Discussion

The impact of collecting and analyzing more than two repeated measures was threefold: intra- and interrater correlations increased in a nonlinear manner; their dynamic was interdependent, such that agreement was magnified when reliability also increased; and the VCA estimates stabilized. The dynamic shifts of the correlations suggest that any index derived from them is subject to change with more repeated measures. In our case, the estimates showed a continuing trend at even ten measurements. The VCA instead showed a pattern of stabilization with more measurements, around six in our study, which makes this method ideal for handling multiple repeated measures and arriving at reliable estimations. Estimates derived using the maximum number of repeated measures typically collected in psychology studies (i.e., two) should therefore be interpreted with caution.

The manner in which the repeated measures are preprocessed also matters. Because the VPCs take into account the residuals, averaging beforehand can increase these estimates as compared to simply adding all the repetitions into the model. The beholder indices are not affected by preprocessing, since their calculation ignores the residuals. The use of either preprocessing step will therefore depend on how the researcher theorizes the residuals. The reporting of VPCs or beholder indices will depend on the specificity of the hypotheses.

General discussion

Quantifying sharedness or idiosyncrasy in judgments is a crucial enterprise for any area of research in which the relative influence of people or stimuli to judgments is in question. For example, theories of social or aesthetic perception (Hehman et al., 2017; Kurosu & Todorov, 2017; Vessel et al., 2018) and practical concerns over criminal justice (Austin & Williams III, 1977; Forst & Wellford, 1981; Hofer et al., 1999) or public health (Shoukri, 2011) depend on these estimates. In fact, so does the replicability and feasibility of research in these areas. Domains for which the stimuli explain more variance may experience greater success recapturing the

average effect of a manipulation. Domains for which the sample explains more variance might instead have a difficult time consistently capturing the impact of a manipulation due to effect heterogeneity. With such widespread consequences, it is important to identify best practices for estimating these contributions.

Our goal was to test whether a general method such as the VCA would provide meaningful estimates of shared and idiosyncratic contributions to judgments while probing data preprocessing procedures and measurement error through the use of repeated measures and the psychometrics of stimuli. First, we showed that the VCA handled stimuli that represented edge cases of rater reliability and agreement. Second, a simulation study revealed that the relative advantage of increasing either sample or stimulus set size on precision depends on the specific estimate. Generally, experimental designs with more than 30 stimuli and 60 raters can lead to reasonable precision. Third, we showed that more than two repeated measures are needed to arrive at stable estimates of shared and idiosyncratic contributions to judgments. Finally, whether repeated measures are averaged or simply added into the model as a preprocessing step will depend on assumptions about the residuals as the former can increase VPC estimates. We elaborate on these findings before providing more concrete analytic recommendations for future research.

The role of stimuli

The VCAs handled all edge cases and showed that the stimuli could lead to extreme forms of shared *and* idiosyncratic judgments depending on their psychometric properties within a specific evaluative dimension. Face beauty judgments showed equal amounts of shared and idiosyncratic contributions congruent with previous findings (Hehman et al., 2017; Hönekopp, 2006), whereas object beauty judgments were highly idiosyncratic. However, it should be noted this was specific to the set of objects used in this study, which were selected because of their psychometric properties of low interrater agreement for judgments of beauty. The same objects showed extremely high interrater agreement for judgments of dangerousness (see Supplementary Figs. 11, 13, and 15). Furthermore, other sets of novel objects have shown moderate agreement for judgments of beauty (Kurosu & Todorov, 2017).

Likewise, sets with different compositions of the same stimulus type (e.g., faces or novel objects) can also lead to varied estimates given the same judgment (e.g., beauty; Hönekopp, 2006; Kurosu & Todorov, 2017). If the estimates are highly idiosyncratic, that may suggest that the stimulus set was homogeneous in a manner relevant to the judgment. Conversely, the heterogeneity of a stimulus set may provide highly shared ratings. For example, deciding whether a supermodel or a goblin (heterogeneous set) is more attractive would lead to high agreement, however rating between two

supermodels (homogeneous set) would lead to more idiosyncratic preferences. Inferences about the eye of the beholder depend on what glasses (i.e., dimension) the eye is observing through and also what they are specifically observing. Thus, estimates of idiosyncrasy and/or sharedness are affected by the kind of stimuli set that researchers decide to use for their study.

Consequently, estimates from single studies are unlikely to capture general truths about the topic at hand. To this end, the range of possible estimates for the same research topic should be compared across varieties of contexts, stimulus sets, and sets of raters. Our results have implications for the potential to capture generalizable knowledge. Specifically, they are consistent with the idea that sample size should not be the only design consideration in studies with multiple raters and multiple stimuli: The number of stimuli matters, too (Westfall et al., 2015; Westfall et al., 2014), as well as their heterogeneity with respect to the judgment. The manner in which a larger stimulus set size can lead to greater precision than does a larger sample size suggests that a large number of stimuli is essential for accurately estimating and replicating shared and idiosyncratic contributions to judgments.

On the basis of our simulations, there were notable patterns in how precision operated for the VPCs and the intra- and inter-rater correlations. First, the VPCs showed overall less variance in terms of magnitude, so these estimates were generally more precise. Second, the VPC allowed more design leniency, such that even 20 stimuli provided reasonable VPC precision, as compared to the correlations. Third, the number of stimuli affected the precision of both intra- and interrater correlations, while the number of raters affected mostly intrarater correlations. As long as a design contained more than ten stimuli, these effects mapped unto the VPCs in a more distinct manner, in which raters were more important for the rater and Rater $\times$ Stimulus VPCs, and stimuli for the stimulus VPC. These effects are, of course, contingent on the simulation's assumptions: restricted stimulus variability and relatively similar means and correlations across repeated measures.

The link between VCA and psychometrics

In Study 2, the simulations also showed that the stimulus variance is associated with rater agreement, and the Rater $\times$ Stimulus variance is associated with rater reliability (Fig. 4). Additional analyses with real rating data in Study 3 showed that the increase in intrarater reliability was typically associated with a decrease in the residual variance and an increase in the Rater $\times$ Stimulus variances (Supplementary Fig. 16), thereby increasing idiosyncratic contributions. Likewise, the increase in interrater agreement was generally associated with a decrease in the Rater $\times$ Stimulus variances, and sometimes an increase in the stimulus variance, increasing shared contributions. The role of the rater variance remains ambiguous as

our simulation results showed no relationship to rater agreement or reliability. In the additional analyses, the rater variance slightly increased with an increase in reliability and decreased with an increase in agreement (Supplementary Fig. 16). Yet, our pattern judgments that contained no rater reliability showed a relatively large rater variance (Fig. 2B). These conflicting results suggest that the interpretation of the rater VPC will depend on whether the ratings are meaningful in the first place (i.e., reliable).

The importance of repeated measurements

We assessed how estimates of idiosyncrasy change by increasing the number of repeated measures, with the assumption that more measures reduce measurement error. For the psychometric correlations, extra measures did improve estimates as shown by the increasing intrarater reliability. Interestingly, this was coupled with a slight increase in interrater agreement for faces, indicating that to the extent that there is initial rater agreement, increasing reliability will also increase agreement. The increases in the correlations seemed to continue even at ten measures, suggesting that these estimates could be further improved. This finding may be related to the number of stimuli used in this study. If participants had rated, say, only ten stimuli, it is more likely that the correlations would have stabilized with less repetitions.

In Study 3, the VPC and beholder indices stabilized toward the estimate using ten measures, yet were very unstable at the number of repeated measures that are currently the norm in psychology studies: two measures. Although this norm may reflect a design compromise, due to worries about task engagement, time on task, and participant fatigue, our results suggest this level of measurement is insufficient. One important caveat is that unless the estimates are close to important boundaries—.5 for the beholder indices or an overlapping range between cluster VPCs—the overall conclusion will be the same. However, if the magnitude of the effect is of importance, then more measures will capture a better estimate.

The role of data preprocessing

For correlations, repeated measures can only be averaged, whereas for modeling analyses, data preprocessing can include averaging data or using all repeated measures. Our findings showed that the averaging or simply adding repeated measures for the modeling analyses did not change the estimates that do not rely on residual variance (beholder indices, variance components). However, for estimates that include the residuals in their equations (VPCs), the averaging procedure tends to increase the estimates, since averaging reduces residual variance. The approach researchers should take will depend on their assumptions about the importance of keeping

the residuals (i.e., systematic or random) intact for their research question.

Analysis recommendations

When it comes to computing intrarater, interrater, and rater-to-group correlations, one should aggregate data (within raters) to maintain the requirement of independence in correlation. Whenever intra-rater consistency/reliability is less than perfect, however, correlation-based estimates of agreement necessarily change as a function of how many repeated measures are collected. Therefore, VCAs provide better estimates for projects with multiple repeated measures (e.g., judgment studies).

When it comes to the use of the VPCs or the beholder indices, the goal of the analysis will determine the number of repeated measures and whether they are averaged or simply added into the model. If one has hypotheses about relative quantities of specific clusters (e.g., rater, stimulus) instead of a simple binary estimate, then the VPCs are more appropriate as long as enough repeated measures are collected. As is shown in the VPC analyses (Fig. 4), with two repeated measures the range of some of the variance components overlaps. The same inferential issues may occur with the beholder indices. For example, the average b_1 is around .5 for facial beauty (Fig. 5), but with only two measures, the range of this estimate can go from below to above .5, shifting the inference from more idiosyncratic to more shared contributions.

Since the VPC is dependent on the residuals, averaging beforehand can increase the estimates with increasing repeated measures. If the residuals are assumed to be meaningful and critical to report (Doherty et al., 2013), or if variation across measurement repetitions is considered important to quantify for the research question (e.g., tasks with a temporal memory or learning component), the best approach would be to simply use the disaggregated data. Mixed models are flexible enough to handle them and will keep the VPC estimates stable with enough repeated measures. If one assumes the residuals are simply noise and that block effects (i.e., changes across repetitions) are unimportant for the research question at hand, one can average to obtain relatively residual-less VPC estimates. The former maps the sources of total variance in the data, whereas the latter isolates and maps variance of interest; both may serve to compare estimates across studies or evaluate replications (Gantman et al., 2018). Whichever approach is taken, it is important to report as the residual-less estimates will be greater than those that include intact residuals and may be harder to compare across studies (for related calls toward standardization of analyses see Judd et al., 2012).

As we mentioned above, most of the analyses improved with greater measures. In our data (Study 3), variance seemed to stabilize around six measures. However, collecting that amount of data might be difficult for many reasons: participant

fatigue or lack of interest, time constraints, and issues of power such as having enough stimuli. In these cases, we recommend collecting as many measures as is feasible for the study design and explicitly quantifying measurement error by bootstrapping confidence intervals for each estimate (e.g., using the bootmer function from the lme4 R package), to get a measure of uncertainty and a range of likely estimates (Goldstein et al., 2002).

In addition to the number of repeated measures, the sample size and stimulus set size should both be given necessary consideration, due to their effects on the precision of estimates. Our simulations (Study 2) only investigated a maximum of 100 raters and 60 stimuli; therefore, there may be room for improvement. But even within this range, we found that, in general, over 30 stimuli and 60 raters lead to reasonable precision across all estimates. Readers can consult Figs. 3 and 4 and Appendix B in the supplementary materials, which contain all estimates across the full range, to provide some visual intuitions about how researchers could balance the experimental costs of additional stimuli or raters.

Finally, it should be made explicit that these analyses were run separately on stimuli that were considered a single set or condition. If one were to combine different categories of stimuli or combine stimuli from different conditions into one analysis, this likely would increase the between-stimulus variance, resulting in largely shared estimates. If there is a main effect of stimulus type or condition in one's stimulus set, this will overestimate the amount of shared judgments. It will be best to run separate analyses for each condition, with each condition cell having an adequate number of raters and stimuli. To compare across conditions, one could use the bootstrapped confidence intervals and examine overlap.

Conclusion

The amount of sharedness or idiosyncrasy in judgments is critical for the science of human behavior (Hehman et al., 2017; Hönemann, 2006; Leder et al., 2016), for the potential replicability of projects (Gantman et al., 2018), and even for understanding the (un)fairness of our justice system (Austin & Williams III, 1977; Forst & Wellford, 1981; Hofer et al., 1999). Our goal here was to find a general method that would best quantify these contributions to judgments, while probing data preprocessing procedures and measurement error through the use of repeated measures and the psychometrics of stimuli. The correlation analyses showed that estimates monotonically increased in a nonlinear manner with repeated measures. Variance component analyses provided a more rigorous examination of idiosyncrasy, by marking where the variance lies in the data (e.g., in the stimuli, the raters, or their interaction), and provided estimates that stabilized with repeated measures. The focus on VPCs or beholder indices as the output of

analysis depends on the specificity of the hypotheses and on assumptions about the residuals. Finally, consideration of the number of stimuli is just as important as consideration of the sample size in experiments designed to target these estimates. These methods are general enough to be useful in any judgment domain in which consensus and disagreement are important to quantify and in which multiple raters independently rate multiple stimuli.

Open Practices Statement We reported all manipulations and measures. None of the studies reported were preregistered, as they are exploratory in nature. All data, analyses, and supplemental material can be found at the following Open Science Framework archive: <https://osf.io/q28g6/>.

Acknowledgments We thank A. Kurosu for data collection, and T. Qian and G. Nicolas for helpful comments on the analyses.

References

- Austin, W., & Williams, T. A., III. (1977). A survey of judges' responses to simulated legal cases: Research note on sentencing disparity. *Journal of Criminal Law and Criminology*, 68, 306–310. doi:<https://doi.org/10.2307/1142852>
- Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, 8, 456:1–13. doi:<https://doi.org/10.3389/fpsyg.2017.00456>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01
- Carter, E. R., & Murphy, M. C. (2017). Consensus and consistency: Exposure to multiple discrimination claims shapes Whites' intergroup attitudes. *Journal of Experimental Social Psychology*, 73, 24–33. doi:<https://doi.org/10.1016/j.jesp.2017.06.001>
- Corbeil, R. R., & Searle, S. R. (1976). A comparison of variance component estimators. *Biometrics*, 32, 779–791.
- Cunningham, M. R., Roberts, A. R., Barbee, A. P., & Druen, P. B. (1995). "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology*, 68, 261–279. doi:<https://doi.org/10.1037/0022-3514.68.2.261>
- Doherty, M. E., Shemberg, K. M., Anderson, R. B., & Tweney, R. D. (2013). Exploring unexplained variation. *Theory and Psychology*, 23, 81–97. doi:<https://doi.org/10.1177/0959354312445653>
- Engell, A. D., Haxby, J. V., & Todorov, A. (2007). Implicit trustworthiness decisions: Automatic coding of face properties in the human amygdala. *Journal of Cognitive Neuroscience*, 19, 1508–1519. doi:<https://doi.org/10.1162/jocn.2007.19.9.1508>
- Forst, B., & Wellford, C. (1981). Punishment and sentencing: Developing sentencing guidelines empirically from principles of punishment. *Rutgers Law Review*, 33, 799–837.
- Gantman, A., Gomila, R., Martinez, J. E., Matias, J. N., Paluck, E. L., Starck, J., . . . Yaffe, N. (2018). A pragmatist philosophy of psychological science and its implications for replication. *Behavioral and Brain Sciences*, 41, e127. doi:<https://doi.org/10.1017/S0140525X18000626>
- Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press. doi:<https://doi.org/10.2277/0521867061>
- Genz, A., Bretz, F., Miwa, T., Mi, X., Friedrich, L., Scheipl, F., & Hothorn, T. (2018). mvtnorm: Multivariate normal and *t* distributions (Version 1.0-8). Retrieved from <http://CRAN.R-project.org/package=mvtnorm>
- Germine, L., Russell, R., Bronstad, P. M., Blokland, G. A. M., Smoller, J. W., Kwok, H., . . . Wilmer, J. B. (2015). Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Current Biology*, 25, 2684–2689. doi:<https://doi.org/10.1016/j.cub.2015.08.048>
- Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, 1, 223–231. doi:https://doi.org/10.1207/S15328031US0104_02
- Goodwin, L. D., & Leech, N. L. (2006). Understanding correlation: Factors that affect the size of *r*. *Journal of Experimental Education*, 74, 249–266. doi:<https://doi.org/10.3200/JEXE.74.3.249-266>
- Grammer, K., & Thornhill, R. (1994). Human (*Homo sapiens*) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology*, 108, 233–242. doi:<https://doi.org/10.1037/0735-7036.108.3.233>
- Gwet, K. L. (2014). Intrarater reliability. Wiley StatsRef: Statistics reference online. doi:<https://doi.org/10.1002/9781118445112.stat06882>
- Hardin, J., Garcia, S. R., & Golan, D. (2013). A method for generating realistic correlation matrices. *Annals of Applied Statistics*, 7, 1733–1762. doi:<https://doi.org/10.1214/13-AOAS638>
- 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, 513–529. doi:<https://doi.org/10.1037/pspa0000090>
- Heiphetz, L., & Young, L. L. (2017). Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs. *Cognition*, 167, 78–90. doi:<https://doi.org/10.1016/j.cognition.2016.05.014>
- Hofer, P. J., Blackwell, K. R., & Ruback, R. B. (1999). Effect of the federal sentencing guidelines on interjudge sentencing disparity. *Journal of Criminal Law and Criminology*, 90, 239–306.
- 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, 199–209. doi:<https://doi.org/10.1037/0096-1523.32.2.199>
- Jacobsen, T., Schubotz, R. I., Höfel, L., & Cramon, D. Y. V. (2006). Brain correlates of aesthetic judgment of beauty. *NeuroImage*, 29, 276–285. doi:<https://doi.org/10.1016/j.neuroimage.2005.07.010>
- Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, 103, 54–69. doi:<https://doi.org/10.1037/a0028347>
- Kenny, D. A. (1996). Interpersonal perception. New York, NY: Guilford Press.
- Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PLoS ONE*, 13, e0202655. doi:<https://doi.org/10.1371/journal.pone.0202655>
- Kurosu, A., & Todorov, A. (2017). The shape of novel objects contributes to shared impressions. *Journal of Vision*, 17(13), 14:1–20. doi:<https://doi.org/10.1167/17.13.14>
- Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, 126, 390–423. doi:<https://doi.org/10.1037/0033-2909.126.3.390>
- Leder, H., Goller, J., Rigotti, T., & Forster, M. (2016). Private and shared taste in art and face appreciation. *Frontiers in Human Neuroscience*, 10, 155:1–7. doi:<https://doi.org/10.3389/fnhum.2016.00155>
- Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska Directed Emotional Faces, KDEF (CD ROM). Stockholm, Sweden:

- Karolinska Institutet, Department of Clinical Neuroscience, Psychology section.
- Ma, F., Xu, F., & Luo, X. (2016). Children's facial trustworthiness judgments: agreement and relationship with facial attractiveness. *Frontiers in Psychology*, 7, 499–1–9. doi:<https://doi.org/10.3389/fpsyg.2016.00499>
- Martinez, J. E., & Paluck, E. L. (2019). Quantifying shared and idiosyncratic judgments of racism in social discourse. Unpublished manuscript.
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537–563. doi:<https://doi.org/10.1146/annurev.psych.59.103006.093735>
- Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews*, 85, 935–956. doi:<https://doi.org/10.1111/j.1469-185X.2010.00141.x>
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105, 11087–11092. doi:<https://doi.org/10.1073/pnas.0805664105>
- Paluck, E. L., & Shafir, E. (2017). The psychology of construal in the design of field experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments* (pp. 245–268). Amsterdam, The Netherlands: North Holland. doi:<https://doi.org/10.1016/bs.hefe.2016.12.001>
- Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162, 8–13. doi:<https://doi.org/10.1016/j.jneumeth.2006.11.017>
- Rhodes, G. (2006). The evolutionary psychology of facial beauty. *Annual Review of Psychology*, 57, 199–226. doi:<https://doi.org/10.1146/annurev.psych.57.102904.190208>
- Schepman, A., Rodway, P., & Pullen, S. J. (2015). Greater cross-viewer similarity of semantic associations for representational than for abstract artworks. *Journal of Vision*, 15(14), 12. doi:<https://doi.org/10.1167/15.14.12>
- Schielzeth, H., & Nakagawa, S. (2013). Nested by design: Model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution*, 4, 14–24. doi:<https://doi.org/10.1111/j.2041-210x.2012.00251.x>
- Schuetzenmeister, A. (2016). VCA: Variance component analysis (Computer software). Retrieved from <https://cran.r-project.org/package=VCA>
- Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components*. Hoboken, NJ: Wiley.
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 6, 922–932.
- Shoukri, M. M. (2011). *Measures of interobserver agreement and reliability* (2nd ed.). Boca Raton, FL: CRC Press.
- Vessel, E. A. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision*, 10(2), 18:1–14. doi:<https://doi.org/10.1167/10.2.18>
- Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition*, 179, 121–131. doi:<https://doi.org/10.1016/j.cognition.2018.06.009>
- Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, 25, 1115–1125. doi:<https://doi.org/10.1177/01461672992512005>
- Westfall, J., Judd, C. M., & Kenny, D. A. (2015). Replicating studies in which samples of participants respond to samples of stimuli. *Perspectives on Psychological Science*, 10, 390–399. doi:<https://doi.org/10.1177/1745691614564879>
- Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in that samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, 143, 2020–2045. doi:<https://doi.org/10.1037/xge0000014>
- Xie, S. Y., Flake, J. K., & Hehman, E. (2019). Perceiver and target characteristics contribute to impression formation differently across race and gender. *Journal of Personality and Social Psychology*, 117, 364–385. doi:<https://doi.org/10.1037/pspi0000160>
- Zebrowitz, L. A., Franklin, R. G., Hillman, S., & Boc, H. (2013). Older and younger adults' first impressions from faces: Similar in agreement but different in positivity. *Psychology and Aging*, 28, 202–212. doi:<https://doi.org/10.1037/a0030927>
- Zebrowitz, L. A., Montepare, J. M., & Lee, H. K. (1993). They don't all look alike: Individual impressions of other racial groups. *Journal of Personality and Social Psychology*, 65, 85–101. doi:<https://doi.org/10.1037/0022-3514.65.1.85>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.