Original Manuscript

Social Psychological and Personality Science
1-12

DOI: 10.1177/19485506211047095
journals.sagepub.com/home/spp


Consensus Enables Accurate Social Judgments

R. Thora Bjornsdottir[1], Eric Hehman[2], and Lauren J. Human[2]

Abstract

Ubiquitous to theories of social perception is an assumed relationship between an attribute's (e.g., intelligence) "signal" and judgment accuracy, with accuracy impossible without the presence and consensual use of signal. Yet this foundational assumption remains untested. Our investigation focused on consensus (quantified using intraclass correlations, ICCs), which should suggest signal availability, according to theories of accurate social perception. Study 1 confirmed that judgments of different social attributes exhibit different degrees of consensus. Study 2 specifically tested the consensus → accuracy link, anticipating that social judgments with higher consensus (target ICCs) would show greater judgment accuracy. Using 497,780 judgments of 3,847 targets from 4,162 participants across 45 data sets testing a broad variety of social judgments, we found that consensus moderated the relationship between targets' self-report and participants' judgments: Judgment accuracy was higher when consensus was higher. Results show the first empirical support for a foundational assumption of theories of social perception.

Keywords

accuracy, consensus, signal, good trait, social perception, ICCs

There has long been interest in the possibility of accurately inferring others' characteristics. Indeed, the study of physiognomy, or the belief that faces reflect individuals' character, appears throughout history (see Re & Rule, 2015; Todorov et al., 2015), and similar lay beliefs persist today (e.g., Hassin & Trope, 2000; Jaeger et al., 2020). In modern science, personality and social psychologists continue to explore what can be accurately assessed from appearance and behavior, and what cannot.

This body of evidence makes clear that not all social attributes are judged with equal accuracy. Research consistently shows differences in judgment accuracy between personality traits: Extraversion, for example, is generally judged more accurately than neuroticism (Beer & Watson, 2008; Funder & Dobroth, 1987; John & Robins, 1993). Research in nonverbal behavior also shows wide variation in social judgment accuracy: Judgments of perceptually obvious group memberships like gender and race are highly accurate (Bruce et al., 1993; Fiske & Neuberg, 1990), whereas judgments of perceptually ambiguous group memberships like sexual orientation and social class show lower accuracy (Bjornsdottir & Rule, 2017; Tskhay & Rule, 2013). What drives this variation?

Numerous frameworks theorize that accuracy in interpersonal perception depends on characteristics of perceivers, targets, and the attributes being judged (Funder, 1995, 2012; Kenny, 1994; Tanner & Swets, 1954; Zebrowitz & Collins, 1997). These frameworks share the assertion that the attribute being judged (e.g., friendliness) has a "signal" strength—a notion tied to signal detection theory, in which signal (patterns of information) must be distinguished from noise (uninformative patterns) to make correct judgments (hits and correct rejections; MacMillan & Creelman, 2004). Similarly, Brunswik's (1956) influential lens model posits that (1) certain cues (valid cues) accurately signal certain attributes and (2) perceivers use some cues (utilized cues) to infer others' attributes. If perceivers utilize enough valid cues, accuracy is possible. Other theoretical frameworks echo the importance of valid cue (i.e., signal) use in accurate judgment formation. For example, the Realistic Accuracy Model states that relevant cues or signal can only lead to accuracy if they are available, detected, and utilized by perceivers (Funder, 1995). Crucially, for a group of perceivers to show accuracy in their judgments of an attribute (with accuracy defined as perceivers' judgments corresponding to targets' ground truth), they must demonstrate *consensus* in their cue use. Consensus should therefore predict variation in accurate judgments of others' social attributes, such that greater

[1] Department of Psychology, Royal Holloway, University of London, Egham, UK
[2] Department of Psychology, McGill University, Montreal, QC, Canada

Corresponding Author:
R. Thora Bjornsdottir, Department of Psychology, Royal Holloway, University of London, Wolfson Building, Egham, Surrey TW20 0EX, UK.
Email: thora.bjornsdottir@rhul.ac.uk

consensus should lead to greater accuracy—but this critical theoretical link remains untested.

Although consensus in the use of a cue is agnostic to its validity (i.e., cue validity and utilization are separate parts of the lens model of perception; Brunswik, 1956) and consensus is therefore insufficient (though necessary) for accuracy (Blackman & Funder, 1998), there is reason to believe that consensus should occur more often for valid cues (signal) than invalid cues (noise) and thereby predict accuracy in social judgments. Existing theory, including the ecological theory of social perception (McArthur & Baron, 1983), posits that individuals' social perceptions serve an adaptive function. Perceivers should therefore detect useful social information (i.e., detect valid cues/signal) from their environments and use this in their judgments. This does not suggest that people's social perceptions should be completely accurate, but simply *accurate enough* to be adaptive in navigating their social environments—and accurate more often than *in*accurate (Haselton & Funder, 2006). Similarly, this suggests that consensus should occur more often for valid than invalid cues (otherwise we would expect more systematic inaccuracy in our social perceptions, which would not make them adaptive), such that consensus should suggest the presence of signal more often than not. This central theoretical claim remains to be empirically tested, however.

The presence of signal should be another source of variation in social judgment accuracy: attributes with stronger signal should elicit more consensus and be judged more accurately. However, we focus our investigation on consensus, first, because signal cannot lead to accuracy without consensual use of that signal. Second, whereas consensus is objectively measurable, signal availability is difficult to directly quantify. Although researchers can systematically explore the specific cues that provide signal to a particular attribute (Brunswik, 1956), it is not possible to provide an overall measure of signal availability that is comparable between social attributes. Research in personality has quantified signal availability using subjective judgments of how "easy" a trait is to judge, finding that more easily observable traits are indeed judged more accurately (Krzyzaniak & Letzring, 2019), lending support to theoretical claims that greater signal should lead to greater judgment accuracy. However, subjective ease of judgment may provide a murky view of signal availability. There may be judgments which feel difficult but are accurate, and those that feel easy, but which give rise to inaccuracy. Indeed, research demonstrates that there are judgments perceivers find easy, but for which they show low consensus (e.g., attractiveness; Hehman et al., 2017), which should lead to low accuracy. Given the difficulty in measuring signal, and that consensus should suggest signal (as discussed above), we focus the present research on consensus.

One approach to quantifying consensus is calculating intraclass correlations (ICCs) for various social judgments. ICCs calculated from cross-classified multilevel models quantify the percentage of variance in judgments arising uniquely from targets, perceivers, and their interaction (Judd et al., 2012; Kenny & Albright, 1987; Raudenbush & Bryk, 2002; Shrout & Fleiss, 1979). This maps conceptually onto the ideas outlined in the Social Relations Model: A perceiver's judgment of a target is due to how the target tends to be judged, how the perceiver tends to judge others, and the unique judgment a certain perceiver makes of a certain target (Kenny, 1994). Target ICCs represent the extent to which between-target differences drive a given judgment: High target ICCs indicate that between-target differences (in appearance, behavior) elicit differences in judgments, denoting greater consensus among perceivers and thus suggesting the presence of signal. We therefore anticipate that greater consensus (higher target ICC) should predict greater judgment accuracy.

Consistent with previous theory (Funder, 1995; Kenny, 1994; Tanner & Swets, 1954), when there is a strong signal, perceivers should better be able to detect and use this signal, resulting in higher consensus and enabling their judgments to more strongly relate to targets' ground truth. Target ICCs (i.e., consensus) should therefore significantly moderate accuracy (the relationship between target ground truth and perceiver judgments). Higher target ICCs should enable stronger relationships between target truth and perceiver judgment, whereas lower target ICCs should result in weaker relations between target truth and perceiver judgment. To date, though a strong assumption central to numerous theoretical frameworks, this relationship has gone untested—a test critical to a robust science of accuracy.

We hypothesized that higher target ICCs would predict greater judgment accuracy. One reason why this relationship has not been tested previously may be that this requires a very large amount of data across numerous studies. Target ICC is a property of a "set" of stimuli, which becomes the unit of analysis—meaning that numerous studies assessing different types of accuracy (e.g., neuroticism, political affiliation) across numerous sets of stimuli are required for a comprehensive test. To this end, we combined 58 studies from 45 data sets testing perceptions of 24 different social attributes. Specifically, we tested whether study-level target ICC moderated the relationship between target ground truth and perceiver judgments (i.e., judgment accuracy).

Study 1

A central assumption of our primary planned analysis is that different judgments have meaningfully different target ICCs. This has been demonstrated among different social trait judgments (Hehman et al., 2017; Xie et al., 2019), but not among the ambiguous group memberships on which much accuracy research focuses (e.g., sexual orientation). We therefore sought to empirically demonstrate that different judgments exhibit different patterns of ICCs, allowing us to test the generalizability of any findings relating consensus to accuracy. We focused on judgments of perceptually ambiguous groups (sexual orientation, political affiliation, social class, religiosity) rather than perceptually obvious groups (which should be near-exclusively target-driven).

Method

We preregistered this study (original: https://osf.io/4cqtu, replication: https://osf.io/rmgb4) and make the data available on the Open Science Framework (OSF; https://osf.io/edqyr). This study received approval from the Research Ethics Board of McGill University.

Original study. We recruited 384 perceivers from Mechanical Turk, excluding the data of participants whose responses were overly repetitive or consistently faster than 400 ms, using a preregistered data cleaning procedure (https://osf.io/65tpb), which retained 369 perceivers (162 female, 195 male, 12 unreported gender; $M_{age} = 36.81$ years, $SD = 11.30$). Participants judged targets on one of six attributes, resulting in an average of 64 perceivers[1] rating each target on each attribute. Previous research demonstrates that ratings of targets "stabilize" with around 40–50 ratings (Hehman et al., 2018; Jones et al., 2021).

Targets consisted of 100 neutral expression photos of White men from the Face Research Lab London Set (DeBruine & Jones, 2017) and the Chicago Face Database (Ma et al., 2015), ranging in age from 18 to 50 years. This sample size, together with our number of perceivers, afforded sufficient power to compute cross-classified MLMs and thus ICCs (see Judd et al., 2012). We randomly assigned perceivers to rate targets on one of the following: sexual orientation (from 1 *exclusively attracted to men* to 7 *exclusively attracted to women*), political ideology (from 1 *very conservative* to 7 *very liberal*), wealth (from 1 *not at all wealthy* to 7 *very wealthy*), religiosity (from 1 *not at all religious* to 7 *very religious*), age (choosing from: *20 or less, 21–24, 25–28, 29–32, 33–36, 37–40,* or *41 or more*), or number of siblings (choosing from: *0, 1, 2,* or *3 or more*). We included the judgments of age, as this should be primarily target-driven (thus showing greater consensus), and of number of siblings, as this should be primarily perceiver-driven (as there should be no signal in the face to infer this information and we are aware of no widely held stereotypes that could drive judgments), to serve as points of comparison. After rating all of the faces in random order, participants provided basic demographic information.

Replication. Following this study, we conducted a conceptual replication. This replication differed in collecting ratings of both men's and women's faces (50 of each gender, White, aged 18–50 years, from the same databases as in the original study), and with perceivers judging each target twice. This approach enabled us to partition Target × Perceiver ICCs (the extent to which a rating depends on characteristics of both the perceiver and the target) from the residual, which is impossible to calculate with single ratings of each target. Here we recruited 768 Mechanical Turk perceivers, retaining 747 (311 female, 387 male, 49 unreported gender; $M_{age} = 37.61$ years, $SD = 12.20$) after data exclusions using the same preregistered cleaning procedure as above. We randomly assigned each participant to provide one of the six ratings of either the men's or women's faces, resulting in an average of 64 participants judging each target on each attribute (as in the original study).

Results

For each of these two data sets, we computed the target and perceiver ICCs (and additionally Target × Perceiver ICCs for the replication data set) for each of the six judgments, by computing null cross-classified multilevel models using the lme4 package in R 3.5.2 (Bates et al., 2015; R Core Team, 2018). Due to the exploratory nature of the research, we partitioned each data set before analysis into a training set and test set for cross-validation. We did this by splitting the data in half to create a hold-out test set to serve as confirmation of patterns revealed in the exploratory data set. Thus, we can have more confidence in patterns replicating across training and test sets. We based conclusions about differences in ICCs across different ambiguous groups on 95% confidence intervals for the ICCs (Xie et al., 2019; R code: https://osf.io/anwx2/).

Original study

Exploratory set. The ICCs for our comparison judgments (age and number of siblings) displayed the anticipated patterns: target ICC was substantially higher than perceiver ICC for age judgments, and we observed the opposite pattern for sibling judgments (see Figure 1A). We also observed substantial variability among our ambiguous group memberships of primary interest. Target and perceiver ICCs did not differ from one another for judgments of wealth and political affiliation, whereas perceiver ICCs were higher than target ICCs for judgments of sexual orientation and religiosity. Perceiver ICCs were similar across these four judgments, but target ICCs were higher for wealth judgments than sexual orientation and religiosity judgments, and higher for political affiliation judgments than religiosity judgments.

Confirmatory set. The patterns in the confirmatory data set largely replicated those in the exploratory set (see Figure 1B), with the exception that target ICCs for political affiliation and religiosity did not significantly differ.

Replication. We did not observe substantial differences in ICCs across ratings of men and women, and therefore collapsed across target gender (see Supplemental Material for patterns split by target gender).

Exploratory set. The ICCs for age and number of siblings again showed the expected patterns, with target ICC significantly higher than perceiver ICC for age judgments and perceiver ICC significantly higher than target ICCs for sibling judgments (see Figure 2A). Target × Perceiver ICCs were moreover greater than both target and perceiver ICCs for sibling judgments and lower than target ICCs for age judgments. Replicating the original study, target and perceiver ICCs were equivalent to one another for wealth and political affiliation judgments, whereas perceiver ICCs exceeded target ICCs for sexual orientation and religiosity judgments. Here, however,

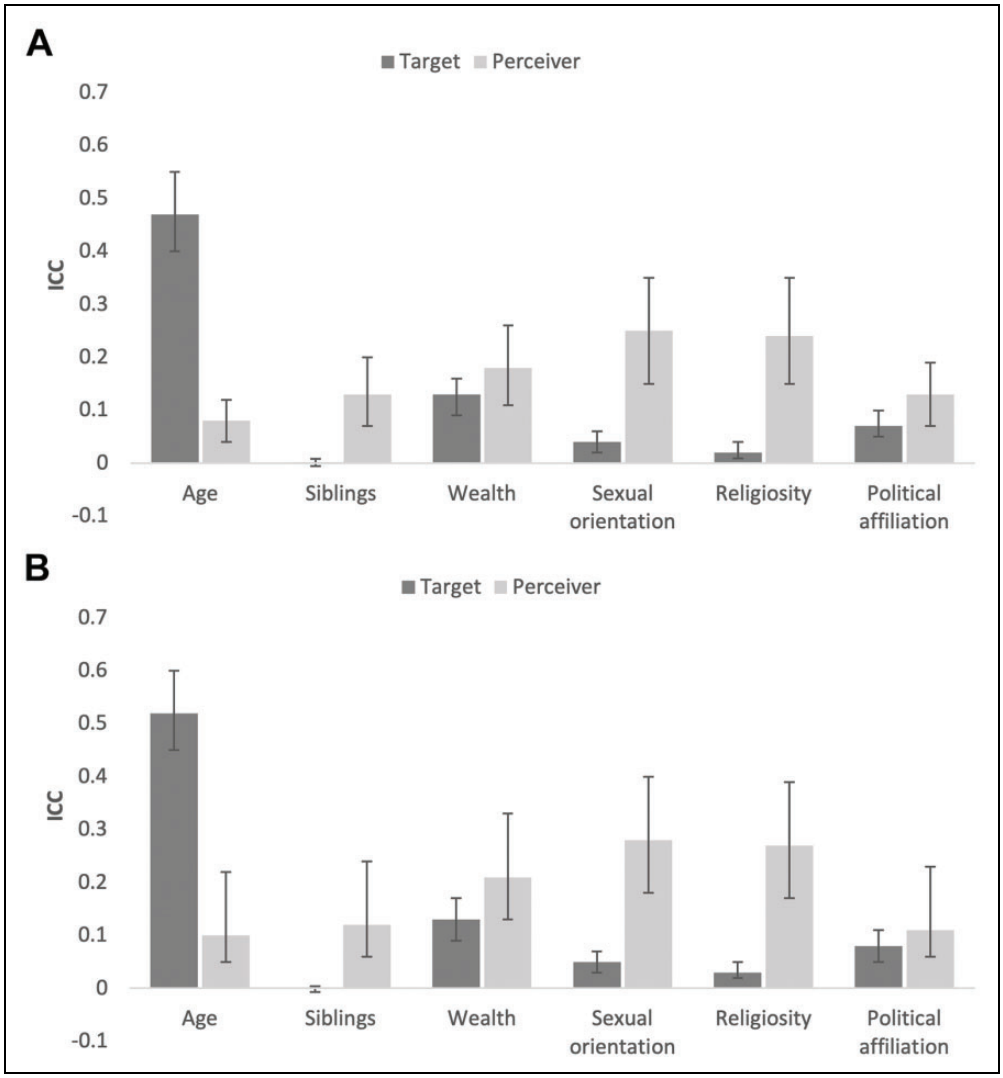

Figure 1. Target and perceiver ICCs for the (A) exploratory and (B) confirmatory data sets of judgments of men's faces in Study 1 original study.
Note. Error bars represent 95% CIs. ICC = Intraclass correlation.

we found perceiver ICCs to be lower for judgments of political affiliation compared to wealth, sexual orientation, and religiosity judgments. Similar to the original study, target ICCs were higher for wealth judgments than for the sexual orientation, religiosity, and political affiliation judgments. Target × Perceiver ICCs exceeded target ICCs for judgments of sexual orientation, religiosity, and political affiliation, and also perceiver ICCs for religiosity and political affiliation. Finally, the Target × Perceiver ICCs were highest for political affiliation and also lower for wealth than for religiosity.

Confirmatory set. The confirmatory data set replicated the patterns observed in the exploratory data set, again with a few exceptions (see Figure 2B). First, perceiver ICCs were significantly greater than target ICCs for wealth judgments in this data set. Target × Perceiver ICCs also did not differ between wealth and religiosity judgments, nor between political affiliation, sexual orientation, and religiosity judgments. Finally, Target × Perceiver ICCs did not exceed perceiver ICCs for religiosity judgments.

Discussion

Altogether, these results reveal substantial (and fairly consistent) variability in ICCs based on the particular ambiguous group membership being judged, joining results examining trait attributions (Hehman et al, 2017; Xie et al., 2019) to illustrate the large variability in consensus (i.e., target ICCs) for different types of judgments. This wide variability is critical to our hypothesis, as we posit that accuracy will scale with target ICCs—for example, we would expect greater accuracy for age judgments than wealth judgments, both of which should be more accurate than religiosity judgments. In this study, however, we simply measured *perceptions* without attempting to

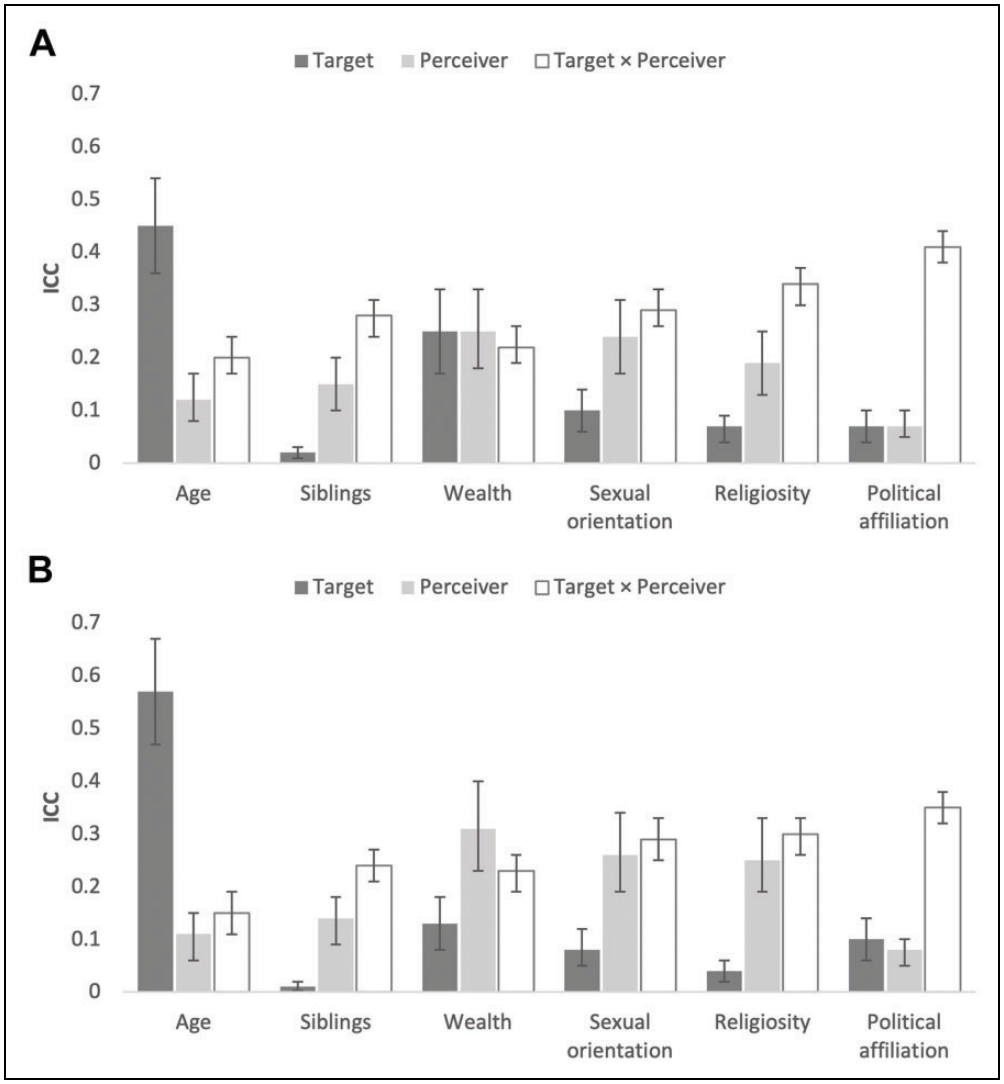

Figure 2. Target, perceiver, and Target × Perceiver ICCs for the (A) exploratory and (B) confirmatory data sets of judgments of men's and women's faces in the Study 1 replication. *Note.* Error bars represent 95% CIs. ICC = Intraclass correlation.

assess *accuracy*, leaving open the question as to whether target ICCs predict social judgment accuracy.

Study 2

To test whether the degree of consensus (i.e., target ICC) predicts greater accuracy in judging targets' attributes, we compiled data which included target self-report information (i.e., ground truth) on a social attribute and perceiver judgments of that same attribute. We hypothesized that higher target ICCs would predict greater social judgment accuracy.

Method

We preregistered this study on the OSF (https://osf.io/t3x45). All data were obtained under the supervision of the Research Ethics Boards at McGill University and the University of Toronto and the Institutional Review Boards at Dartmouth College, New York University, and the University of Delaware.

Source of data. We compiled data from studies involving social judgments for which we could compute accuracy (i.e., for which we had perceiver judgments and a ground truth for the target) from our own laboratories. We also approached colleagues for inclusion of any relevant data sets. Our final sample consisted of data from both published and unpublished studies testing accuracy in detecting 24 different social attributes, including targets' health, social group memberships (sex, sexual orientation, social class), personality (Big 5), ideology (political affiliation, racism), future outcomes (election success), and other individual differences (acculturation, intelligence, sexual preference). This sample included 58 studies from 45 data sets, 4,162 unique perceivers (both undergraduate and online participants), 3,847 unique targets, and 497,780 individual perceiver judgments of individual targets, from face photos or in-person interactions (see Table 1 for summary of data). Although we originally planned to include only studies using face images, this we expanded, as predictions derived

from these theoretical frameworks apply to social judgments from any modality.

For each study, we calculated the study-level target ICC as our measure of consensus ($M = 0.13$, $SD = 0.11$, $Range = .03–.71$; see Table 1) by running a null cross-classified multilevel model for each study, with perceiver judgment as the outcome variable. We then compiled the data into two aggregate data sets: one for studies with binary perceiver judgments (e.g., gay vs. straight categorization) and one for studies with continuous perceiver judgments (e.g., judgments of extraversion). For each of these two data sets, we then ran a cross-classified multilevel model with the target ground truth (group-mean centered by study), target ICC (grand-mean centered and *z*-scored), and the interaction between the two predicting the perceiver judgment, using glmer for the binary data set and lmer for the continuous data set (both from the lme4 package; Bates et al., 2015), and using lmerTest for significance tests (Kuznetsova et al., 2017) in R 3.5.2 (R Core Team, 2018).[2] As lme4 has a low threshold for flagging nonconvergence (McCoach et al., 2018), we confirmed parameter estimates were not an artifact of nonconvergence in the Bayesian brms package (Bürkner, 2017, 2018) for any models failing to converge in lme4. We ran our models with random intercepts for studies, perceivers, and targets, and a random slope for target ground truth across targets. This can be expressed by the equations:

Level 1:

$$Y_{ijkl} = \beta_{0jkl} + \beta_{1jkl} GroundTruth_{ikl} + \epsilon_{ijkl}$$

Level 2:

$$\beta_{0jkl} = \beta_{0000} + \beta_{0100} TargetICC_j + u_{0j00} + u_{00k0} + u_{000l}$$
$$\beta_{1jkl} = \beta_{1000} + \beta_{1100} TargetICC_j + u_{10k0}$$

At Level 1 of the model, Y_{ijkl} is a judgment of attribute i (e.g., conscientiousness) by perceiver j of target k in study l. The intercept, β_{0jkl}, is the expected value of this judgment, at the average level of targets' ground truth. β_{1jkl} represents the correspondence between a perceiver's judgment and the target's ground truth. At Level 2, the intercept, β_{0jkl}, is an outcome modeled as the grand mean of judgment, β_{0000}, and the average change in judgment for each perceiver as target ICC increases, β_{0100}. The residual, u_{0j00}, represents perceiver j's deviation from the grand mean across all targets and studies. Next, u_{00k0} is the residual of target k from the grand mean across all perceivers and studies. The final residual, u_{000l}, is the residual of study l or the difference between the grand mean and the average judgment in study l averaged across all perceivers and targets.

β_{1jkl} represents the effect of targets' ground truth on judgment and is modeled as β_{1000}, the average increase in judgment with every 1-unit increase in ground truth (across perceivers, targets, and studies). The residual, u_{10k0}, represents the variation of target k around this average relationship. Finally, the cross-level interaction, β_{1100}, is the change in this slope with every 1-unit increase in target ICC, which varies across perceivers; in other words, this represents the between-perceiver effect of target ICC on the relationship between targets' ground truth and judgment. See Supplemental Material for R syntax.

Results

Continuous perceiver judgments. For the continuous data, we observed a significant main effect of target ground truth, $b = .038$, $SE = .004$, $t(661.86) = 10.79$, $p < .001$, as well as the hypothesized significant interaction between target ground truth and target ICC, $b = .030$, $SE = .002$, $t(11286.22) = 14.42$, $p < .001$ (see Figure 3A), model conditional $R^2 = .51$. Decomposing this interaction, among studies with high target ICCs (+1 *SD*), we observed a significant main effect of target ground truth in predicting perceiver judgment, $b = .069$, $SE = .004$, $t(1288.63) = 16.19$, $p < .001$. We also observed this main effect among studies with low target ICCs (−1 *SD*), though the effect was weaker, $b = .008$, $SE = .004$, $t(818.28) = 2.01$, $p = .04$. Consistent with our hypothesis, accuracy was higher when target ICC (i.e., consensus) was greater.

Binary perceiver judgments. A similar pattern emerged for the binary data. We again observed a main effect of target ground truth, $b = .42$, $SE = .03$, $z = 16.31$, $p < .001$, and a Target Ground Truth × Target ICC interaction, $b = .66$, $SE = .02$, $z = 36.16$, $p < .001$ (see Figure 3B), model conditional $R^2 = .35$. Decomposing this interaction, when ICCs were high, target ground truth was significantly positively associated with perceiver judgments, $b = 1.08$, $SE = .03$, $z = 37.01$, $p < .001$. Distinct from the continuous data, when ICCs were low, this effect was both smaller and entirely reversed, such that target ground truth was negatively associated with perceiver judgments, $b = -.24$, $SE = .03$, $z = -6.90$, $p < .001$. Thus, we observed accuracy only when target ICCs were high, and *in*accuracy when target ICCs were low.

Discussion

In sum, we found that target ICC interacted with targets' self-reported ground truth to predict perceiver judgments, such that greater target ICCs predicted greater judgment accuracy. In other words, attributes for which there was greater consensus were judged more accurately, indicating that for social judgments to be accurate, there needs to be sufficient agreement in cue use. This pattern was more extreme for binary judgments of targets, which perhaps have less margin for error.

General Discussion

Using data from 58 studies from 45 data sets, here we demonstrated that higher consensus, operationalized as target ICC, predicts greater accuracy in social judgments. This finding indicates that variation in accuracy in social perception is partly caused by agreement among perceivers, not only demonstrating that consensus is a necessary precondition for accuracy but also providing empirical evidence that consensus often implies the use of signal (i.e., valid cues rather than invalid cues). Our

Table 1. Summary of Data Included in Study 2.

Source	Judgment	Number of Targets	Number of Perceivers	Perceiver Judgment Type	Targets	Target ICC
Tskhay & Rule (2015) Study 4b	Democrat or Republican	118	33	Binary	Face photos (professional images of candidates, grayscale)	.085
Hehman & Freeman (2014b) (unpublished data)	Male or female	75	38	Binary	Face photos (edited, high contrast Mooney-like faces)	.457
Hehman & Freeman (2014c) (unpublished data)	Male or female	75	38	Binary	Face photos (edited, high contrast Mooney-like faces)	.110
Hehman & Freeman (2015) (unpublished data)	Male or female	200	31	Binary	Face photos (computer generated, crossing sex morphology and sex reflectance)	.711
Bjornsdottir & Rule (2017) Study 1	Rich or poor	160[1]	81	Binary	Face photos (from online dating profiles, grayscale)	.116
Bjornsdottir & Rule (2017) Study 1 replication	Rich or poor	160[1]	80	Binary	Face photos (from online dating profiles, grayscale)	.151
Bjornsdottir & Rule (2017) Study 2a	Rich or poor	160[1]	71	Binary	Face photos (from online dating profiles, grayscale)	.133
Bjornsdottir & Rule (2017) Study 4a	Rich or poor	160[2]	76	Binary	Face photos (standardized, neutral, grayscale)	.092
Bjornsdottir & Rule (2017) Study 5b	Rich or poor	160[2]	93	Binary	Face photos (standardized, neutral, grayscale)	.095
Bjornsdottir & Rule (2017) Study 6b	Rich or poor	160[2]	75	Binary	Face photos (standardized, smiling, grayscale)	.069
Bjornsdottir & Rule (2017) Supplemental study	Rich or poor	160[2]	293	Binary	Face photos (standardized, neutral, grayscale)	.103
Tskhay et al. (2016) Study 1a	Sick or healthy	124	33	Binary	Face photos (from dating profiles, grayscale)	.100
Tskhay et al. (2016) Study 1b	Sick or healthy	144	37	Binary	Face photos (from dating profiles, grayscale)	.142
Tskhay & Rule (2013b) Study 1	Sexual preference	198	23	Binary	Face photos (from dating profiles, grayscale)	.041
Hehman et al. (2014); Hehman & Freeman (2014a) (unpublished data)	Who won election	100	19	Binary	Face photos (professional images of candidates, color)	.158
Bjornsdottir & Rule (2021) Study 1	Acculturation	189[3]	90[a]	Continuous	Face photos (standardized, neutral, grayscale)	.058
Bjornsdottir & Rule (2021) Study 1	Acculturation	189[3]	90[a]	Continuous	Face photos (standardized, happy, grayscale)	.077
Bjornsdottir & Rule (2021) Study 1	Acculturation	189[3]	90[a]	Continuous	Face photos (standardized, angry, grayscale)	.044
Bjornsdottir & Rule (2021) Study 2	Acculturation	189[3]	72[b]	Continuous	Face photos (standardized, neutral, grayscale, cropped to internal features)	.064
Bjornsdottir & Rule (2021) Study 2	Acculturation	189[3]	72[b]	Continuous	Face photos (standardized, happy, grayscale, cropped to internal features)	.092
Bjornsdottir & Rule (2021) Study 2	Acculturation	189[3]	72[b]	Continuous	Face photos (standardized, angry, grayscale, cropped to internal features)	.054
Bjornsdottir & Rule (2021) Study 4	Acculturation	189[3]	258[c]	Continuous	Face photos (standardized, neutral, grayscale, cropped to internal features)	.047
Bjornsdottir & Rule (2021) Study 4	Acculturation	189[3]	258[c]	Continuous	Face photos (standardized, happy, grayscale, cropped to internal features)	.068
Bjornsdottir & Rule (2021) Study 4	Acculturation	189[3]	258[c]	Continuous	Face photos (standardized, angry, grayscale, cropped to internal features)	.045

(continued)

Table 1. (continued)

Source	Judgment	Number of Targets	Number of Perceivers	Perceiver Judgment Type	Targets	Target ICC
Tissera et al. (2020) Study 1	Agreeableness	556[4]	557[d]	Continuous	Round robin interaction	.100
Bjornsdottir (2019) Study 1a	Class category	495[5]	95	Continuous	Face photos (standardized neutral, grayscale)	.145
Bjornsdottir (2019) Study 2a	Class category	330[6]	81	Continuous	Face photos (from yearbooks, grayscale)	.063
Bjornsdottir & Rule (under review) Study 3a	Class category	495[5]	104	Continuous	Face photos (standardized neutral, color)	.178
Bjornsdottir & Rule (under review) Study 3b	Class category	495[5]	108	Continuous	Face photos (standardized neutral, color)	.190
Bjornsdottir & Rule (under review) Study 3c	Class category	330[6]	84	Continuous	Face photos (from yearbooks, grayscale, cropped to internal features)	.060
Tissera et al. (2020) Study 1	Conscientiousness	556[4]	557[d]	Continuous	Round robin interaction	.107
Tissera et al. (2020) Study 1	Extraversion	556[4]	557[d]	Continuous	Round robin interaction	.457
Bjornsdottir (2019) Study 1a	Family income	475[5]	115	Continuous	Face photos (standardized neutral, grayscale)	.146
Bjornsdottir & Rule (under review) Study 3a	Family income	475[5]	94	Continuous	Face photos (standardized neutral, color)	.134
Bjornsdottir & Rule (under review) Study 3b	Family income	475[5]	99	Continuous	Face photos (standardized neutral, color)	.154
Bjornsdottir (2019) Study 2a	Future class category	313[6]	82	Continuous	Face photos (from yearbooks, grayscale)	.092
Bjornsdottir (2019) Study 2a	Future education	330[6]	83	Continuous	Face photos (from yearbooks, grayscale)	.030
Bjornsdottir (2019) Study 2a	Future occupational prestige	313[6]	72	Continuous	Face photos (from yearbooks, grayscale)	.055
Tissera et al. (2020) Study 1	Intelligence	556[4]	557[d]	Continuous	Round robin interaction	.112
Tissera et al. (2020) Study 1	Neuroticism	556[4]	557[d]	Continuous	Round robin interaction	.084
Tissera et al. (2020) Study 1	Openness	556[4]	557[d]	Continuous	Round robin interaction	.073
Bjornsdottir (2019) Study 1a	Parental education	455[5]	105	Continuous	Face photos (standardized neutral, grayscale)	.099
Bjornsdottir & Rule (under review) Study 3a	Parental education	455[5]	88	Continuous	Face photos (standardized neutral, color)	.115
Bjornsdottir & Rule (under review) Study 3b	Parental education	455[5]	89	Continuous	Face photos (standardized neutral, color)	.093
Bjornsdottir (2019) Study 2a	Parental occupational prestige	330[6]	82	Continuous	Face photos (from yearbooks, grayscale)	.083
Bjornsdottir & Rule (under review) Study 3c	Parental occupational prestige	330[6]	76	Continuous	Face photos (from yearbooks, grayscale, cropped to internal features)	.065
Tissera et al. (2020) Study 1	Political affiliation	555[4]	557[d]	Continuous	Round robin interaction	.044
Hehman et al. (2013) Study 2; Hehman & Leitner (2013) (unpublished data)	Racism	40	101	Continuous	Face photos (standardized neutral, color)	.106
Hehman et al. (2013) Study 3; Hehman & Leitner (2013) (unpublished data)	Racism	40	45	Continuous	Face photos (standardized neutral, color)	.147
Bjornsdottir (2019) Study 1a	SES ladder	455[5]	115	Continuous	Face photos (standardized neutral, grayscale)	.129
Bjornsdottir & Rule (under review) Study 3a	SES ladder	455[5]	108	Continuous	Face photos (standardized neutral, color)	.108

(continued)

Table 1. (continued)

Source	Judgment	Number of Targets	Number of Perceivers	Perceiver Judgment Type	Targets	Target ICC
Bjornsdottir & Rule (under review) Study 3b	SES ladder	455[5]	109	Continuous	Face photos (standardized neutral, color)	.153
Bjornsdottir & Rule (2020) Study 2	Sexual orientation	64[7]	29	Continuous	Face photos (standardized, neutral, grayscale)	.338
Bjornsdottir & Rule (2020) Study 2	Sexual orientation	100[8]	30	Continuous	Face photos (from online dating profiles, grayscale)	.217
Bjornsdottir & Rule (2020) Study 3	Sexual orientation	396[7]	86	Continuous	Face photos (standardized, neutral, grayscale)	.184
Bjornsdottir & Rule (2020) Study 3	Sexual orientation	373[8,9]	58	Continuous	Face photos (from online dating profiles, grayscale)	.250
Tissera et al. (2020) Study 1	Sexual orientation	556[4]	557[d]	Continuous	Round robin interaction	.137
Tskhay & Rule (2015) Study 4a	Sexual orientation	90[9]	29	Continuous	Face photos (from online dating profiles, grayscale)	.147

Note. Target/perceiver numbers with the same superscript indicate the same or overlapping samples of targets/perceivers across studies.

pattern of results emerged across a broad variety of social judgments, including personality traits and social group memberships, across studies in which perceivers made binary judgments of targets and in which they provided judgments along continuous scales, and across studies where judgments were based on photographs and brief interactions, supporting the generalizability of our findings.

Our findings build upon previous work within personality research, providing an alternative to subjective measures of the visibility (or "observability") of social attributes (ease of judging an attribute; see Krzyzaniak & Letzring, 2019), focusing instead on the more objectively quantifiable link between consensus and accuracy. Our work also expands personality-focused models of person perception to social attributes more broadly. For example, Funder's (1995, 2012) Realistic Accuracy Model proposes the idea of the *good trait* or a trait that is accurately perceived. Our findings not only help quantify what helps make a good trait (consensus) but also extend this concept to judgments beyond traits: Judgments of some social attributes (e.g., social group memberships, personality traits, ideologies) show greater consensus, are thus judged more accurately, and could therefore be considered *good attributes* in Funder's framework.

Our results also dovetail with Kenny's (1994) Social Relations Model—which posits that perceiver judgments of a target are a function of the target, the perceiver, and the relationship between the two. We empirically demonstrated in Study 1 that target ICCs (i.e., consensus), perceiver ICCs, and Target × Perceiver ICCs systematically vary for social judgments other than traits (Hehman et al., 2017). Future research could explore the precise role of Target × Perceiver interactions (which remain under-explored; Hehman et al., 2019; Letzring & Funder, 2019; Rogers & Biesanz, 2019) in predicting accuracy, building upon both our findings and research on what makes a good target or a good judge (e.g., Human & Biesanz, 2013; Letzring, 2008).

Importantly, our findings provide empirical support to theories highlighting the adaptive nature of our social perceptions (e.g., Haselton & Funder, 2006; McArthur & Baron, 1983). A notable caveat to greater consensus predicting greater judgment accuracy is that the utilized cues must be *valid* (i.e., be true signals) for this to be true. Consensus on its own does not lead to accuracy (e.g., trustworthiness judgments from faces; Rule et al., 2013). In other words, in the language of the Realistic Accuracy Model (Funder, 1995, 2012), though cues may be available, detected, and utilized, this does not lead to accuracy unless the cues are relevant/valid. However, consensus positively predicted social judgment accuracy, indicating that consensus implied *valid* cue/signal use. Consensus led to accuracy in social judgments more often than not, suggesting that perceivers are sensitive to and use signal (valid cues)—allowing for social judgments that are accurate enough to successfully navigate social environments. This is no doubt partly attributable to our specific data set (in which the studies often tested social perceptions that had a theoretical basis for accuracy), but we also believe that this finding illustrates a broader pattern in social perception.

However, one question for future research to explore is the relationship between target ICCs and judgment (in)accuracy among attributes for which there is no valid signal but there is variation in consensus. It is tenable that higher target ICCs among such judgments might lead to even greater inaccuracy or overconfidence in judgments. Another crucial step would be to develop methods to systematically quantify the *presence* versus the *use* of signal or to disentangle the valid utilized cues from invalid utilized cues in target ICCs.

Finally, an important nuance is that target ICCs are multiply determined and arise from features beyond the social attribute being judged. Consensus may be driven by perceptual characteristics of a target and how clearly an attribute appears to be exhibited (as in Study 1). However, consensus is a property of a *set* of stimuli, and features of the design and stimulus presentation

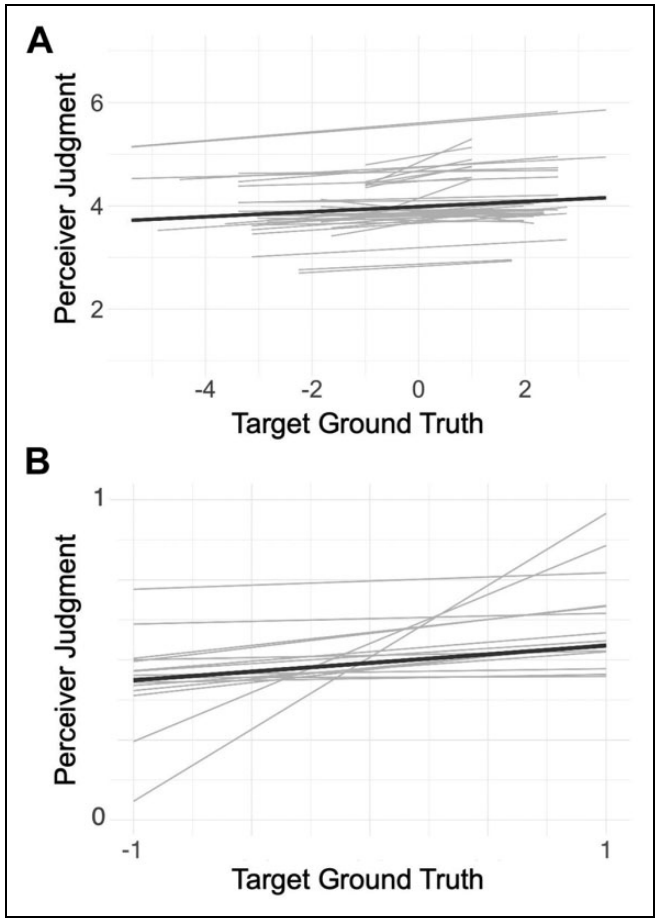

Figure 3. Relation between target ground truth and perceiver judgment, for (A) continuous and (B) binary perceiver judgment data.
Note. The darker line illustrates the grand slope and ligher lines represent slopes for individual studies.

would also influence it. For example, research indicates that participants are more sensitive to stimulus differences in binary forced-choice tasks, compared to sequential ratings of individual targets (Burton et al., 2019). Alternatively, a stimulus set including targets ranging low to high on an attribute would elicit more consensus than a set including targets ranging only medium to high. We therefore caution against drawing conclusions about the exact degree of consensus for a given social attribute based on the present results, and instead encourage researchers to think of consensus as multiply determined. Regardless, our primary finding illustrates that higher consensus, regardless of source, enables greater judgment accuracy.

This opens doors for future research to more precisely determine the relative degree of consensus for different social judgments, for example, by collecting self-report and perceiver judgment data on a variety of attributes for the same set of targets (similar to Study 1, albeit with targets' self-reported ground truth available in order to assess accuracy). Such an investigation would valuably inform future research, for example, providing a glimpse into how much researchers might expect (impressions of) different social attributes to influence social interactions or outcomes. Moreover, quantifying the degree of consensus is an important step toward determining whether it may be fruitful to search for specific aspects of attributes, perceivers, or targets that predict accuracy. If judgments of the attribute do not have sufficient consensus, such attempts are unlikely to lead to robust findings.

Altogether, our findings pinpoint one contribution to variations in social judgment accuracy: the degree of consensus. Social attributes with greater consensus, as measured by target ICCs, are perceived more accurately. Our findings thus both provide empirical support for numerous theories of accurate social perception and help to provide a more complete picture of the predictors of accuracy in social perception.

Authors' Note

All studies were preregistered (Study 1: https://osf.io/4cqtu; Study 1 replication: https://osf.io/rmgb4; Study2: https://osf.io/t3x45) and Study 1 data are available on the OSF (https://osf.io/edqyr/). Requests for Study 2 data can be sent via email to the lead author.

Acknowledgments

The authors thank Konstantin Tskhay for sharing his data.

Author Contributions

RTB and EH conceived the research. All authors developed the methodology. RTB performed data curation. RTB and EH analyzed the data. RTB wrote the original draft. All authors contributed in writing—review and editing.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

R. Thora Bjornsdottir https://orcid.org/0000-0002-1016-3829
Eric Hehman https://orcid.org/0000-0003-2227-1517
Lauren J. Human https://orcid.org/0000-0001-8384-2075

Supplemental Material

The supplemental material is available in the online version of the article.

Notes

1. Our preregistration misstated we would recruit 60 participants per judgment.
2. This differs slightly from our preregistered analysis plan, though our interpretation of results does not differ when following that plan (see Supplemental Material).

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.

Beer, A., & Watson, D. (2008). Personality judgment at zero acquaintance: Agreement, assumed similarity, and implicit simplicity. *Journal of Personality Assessment*, *90*, 250–260.

Bjornsdottir, R. T. (2019). *Investigating the pervasiveness of social class cues in the face* [Doctoral dissertation, University of Toronto, Toronto, Canada]. Retrieved from http://hdl.handle.net/1807/97322

Bjornsdottir, R. T., & Rule, N. O. (2017). The visibility of social class from facial cues. *Journal of Personality and Social Psychology*, *113*, 530–546.

Bjornsdottir, R. T., & Rule, N. O. (2020). Emotion and gender typicality cue sexual orientation differently in women and men. *Archives of Sexual Behavior*, *49*, 2547–2560.

Bjornsdottir, R. T., & Rule, N. O. (2021). Perceiving acculturation from neutral and emotional faces. *Emotion*, *21*, 720–729.

Bjornsdottir, R. T., & Rule, N. O. (under review). *Facial cues consistently convey social class across two distinct cultural and economic contexts*.

Blackman, M. C., & Funder, D. C. (1998). The effect of information on consensus and accuracy in personality judgment. *Journal of Experimental Social Psychology*, *34*, 164–181.

Bruce, V., Burton, A. M., Hanna, E., Healey, P., Mason, O., Coombes, A., & Linney, A. (1993). Sex discrimination: How do we tell the difference between male and female faces? *Perception*, *22*, 131–152.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

Bürkner, P. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411.

Burton, N., Burton, M., Rigby, D., Sutherland, C. A., & Rhodes, G. (2019). Best-worst scaling improves measurement of first impressions. *Cognitive Research: Principles and Implications*, *4*, 1–10.

DeBruine, L., & Jones, B. (2017). Face research lab London set. *Almetric*, figshare, Dataset. https://doi.org/10.6084/m9.figshare.5047666.v5

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In *Advances in experimental social psychology* (Vol. *23*, pp. 1–74). Academic Press.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review*, *102*, 652–670.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science*, *21*, 177–182.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*, 409–418.

Haselton, M. G., & Funder, D. C. (2006). The evolution of accuracy and bias in social judgment. In M Schaller, J. A. Simpson, & D. T. Kendrick (Eds.), *Evolution and social psychology* (pp. 15–37). Psychology Press.

Hassin, R., & Trope, Y. (2000). Facing faces: Studies on the cognitive aspects of physiognomy. *Journal of Personality and Social Psychology*, *78*, 837–852.

Hehman, E., Leitner, J. B., Deegan, M. P., & Gaertner, S. L. (2013). Facial structure is indicative of explicit support for prejudicial beliefs. *Psychological Science*, *24*, 289–296.

Hehman, E., Carpinella, C. M., Johnson, K. L., Leitner, J. B., & Freeman, J. B. (2014). Early processing of gendered facial cues predicts the electoral success of female politicians. *Social Psychological and Personality Science*, *7*, 815–824.

Hehman, E., Stolier, R. M., Freeman, J. B., Flake, J. K., & Xie, S. Y. (2019). Toward a comprehensive model of impression formation: What we know, what we do not, and paths forward. *Social and Personality Psychology Compass*, *13*, e12431.

Hehman, E., Sutherland, C. A., 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*, 513–530.

Hehman, E., Xie, S. Y., Ofosu, E. K., & Nespoli, G. (2018). Assessing the point at which averages are stable: A tool illustrated in the context of person perception. *PsyArXiv*. https://doi.org/10.31234/osf.io/2n6jq

Human, L. J., & Biesanz, J. C. (2013). Targeting the good target: An integrative review of the characteristics and consequences of being accurately perceived. *Personality and Social Psychology Review*, *17*, 248–272.

Jaeger, B., Evans, A., Stel, M., & van Beest, I. (2020). Lay beliefs in physiognomy explain overreliance on facial impressions. *PsyArXiv*. https://doi.org/10.31234/osf.io/8dq4x

John, O. P., & Robins, R. W. (1993). Determinants of interjudge agreement on personality traits: The big five domains, observability, evaluativeness, and the unique perspective of the self. *Journal of Personality*, *61*, 521–551.

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., Mburu, G., . . . Coles, N. A.. (2021). To which world regions does the valence-dominance model of social perception apply? *Nature Human Behaviour*, *5*, 159–169. https://doi.org/10.1038/s41562-020-01007-2

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.

Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. Guilford Press.

Kenny, D. A., & Albright, L. (1987). Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, *102*, 390–402.

Krzyzaniak, S. L., & Letzring, T. D. (2019). Characteristics of traits that are related to accuracy of personality judgments. In *The Oxford handbook of accurate personality judgment*. Oxford University Press.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26.

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality*, *42*, 914–932.

Letzring, T. D., & Funder, D. C. (2019). The realistic accuracy model. In T. D. Letzring & J. S. Spain (Eds.), *The Oxford handbook of accurate personality judgment*. Oxford University Press USA.

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*, 1122–1135.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory. A user's guide*. Taylor and Francis Group.

McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review*, *90*, 215–238.

McCoach, D. B., Rifenbark, G. G., Newton, S. D., Li, X., Kooken, J., Yomtov, D., Gambino, A.J., & Bellara, A. (2018). Does the package matter? A comparison of five common multilevel modeling software packages. *Journal of Educational and Behavioral Statistics*, *43*, 594–627.

R Core Team. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. SAGE.

Re, D. E., & Rule, N. O. (2015). Appearance and physiognomy. In D. Matsumoto, H. Hwang, & M. Frank (Eds.), *APA handbook of nonverbal communication* (pp. 221–256). American Psychological Association.

Rogers, K. H., & Biesanz, J. C. (2019). Reassessing the good judge of personality. *Journal of Personality and Social Psychology*, *117*, 186–200.

Rule, N. O., Krendl, A. C., Ivcevic, Z., & Ambady, N. (2013). Accuracy and consensus in judgments of trustworthiness from faces: Behavioral and neural correlates. *Journal of Personality and Social Psychology*, *104*, 409–426.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Tanner, W. P., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*. *61*, 401–409.

Tissera, H., Kerr, L. G., Carlson, E. N., & Human, L. J. (2020). Social anxiety and liking: Towards understanding the role of metaperceptions in first impressions. *Journal of Personality and Social Psychology*. http://dx.doi.org/10.1037/pspp0000363

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.

Tskhay, K. O., & Rule, N. O. (2013). Accuracy in categorizing perceptually ambiguous groups: A review and meta-analysis. *Personality and Social Psychology Review*, *17*, 72–86.

Tskhay, K. O., & Rule, N. O. (2013). Accurate identification of a preference for insertive versus receptive intercourse from static facial cues of gay men. *Archives of Sexual Behavior*, *42*, 1217–1222.

Tskhay, K. O., & Rule, N. O. (2015). Emotions facilitate the communication of ambiguous group memberships. *Emotion*, *15*, 812–826.

Tskhay, K. O., Wilson, J. P., & Rule, N. O. (2016). People use psychological cues to detect physical disease from faces. *Personality and Social Psychology Bulletin*, *42*, 1309–1320.

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.

Zebrowitz, L. A., & Collins, M. A. (1997). Accurate social perception at zero acquaintance: The affordances of a Gibsonian approach. *Personality and Social Psychology Review*, *1*, 204–223.

Author Biographies

R. Thora Bjornsdottir is a lecturer (assistant professor) at Royal Holloway, University of London. Her research explores how social group memberships of both targets and perceivers affect social perception, with a focus on first impressions from faces.

Eric Hehman is an assistant professor at McGill University. His research examines how individuals perceive and evaluate one another across group boundaries (e.g., race, gender, sexual orientation, occupation, etc.), and the downstream consequences of such perceptions.

Lauren J. Human is an assistant professor and Canada Research Chair in Social Perception and Expression at McGill University. Her research examines the causes and consequences of forming accurate interpersonal impressions, with a focus on first impressions of personality.

Handling Editor: Peter Rentfrow