Behavior Research Methods
https://doi.org/10.3758/s13428-024-02489-y

ORIGINAL MANUSCRIPT

Validating a brief measure of four facets of social evaluation

Alex Koch[1] · Austin Smith[1] · Susan T. Fiske[2] · Andrea E. Abele[3] · Naomi Ellemers[4] · Vincent Yzerbyt[5]

Accepted: 1 August 2024

Abstract
Five studies ($N = 7972$) validated a brief measure and model of four facets of social evaluation (friendliness and morality as horizontal facets; ability and assertiveness as vertical facets). Perceivers expressed their personal impressions or estimated society's impression of different types of targets (i.e., envisioned or encountered groups or individuals) and numbers of targets (i.e., between six and 100) in the separate, items-within-target mode or the joint, targets-within-item mode. Factor analyses confirmed that a two-items-per-facet measure fit the data well and better than a four-items-per-dimension measure that captured the Big Two model (i.e., no facets, just the horizontal and vertical dimensions). As predicted, the correlation between the two horizontal facets and between the two vertical facets was higher than the correlations between any horizontal facet and any vertical facet. Perceivers' evaluations of targets on each facet were predictors of unique and relevant behavior intentions. Perceiving a target as more friendly, moral, able, and assertive increased the likelihood of relying on the target's loyalty, fairness, intellect, and hubris in an economic game, respectively. These results establish the external, internal, convergent, discriminant, and predictive validity of the brief measure and model of four facets of social evaluation.

Keywords Big Two · Four facets · Brief measure · Validity · Adversarial collaboration

As social beings, people evaluate themselves and others to create opportunities and solve problems. That is, they notice and infer people's attributes, to guide behavior. Ideally, they learn to precisely and efficiently evaluate a few attributes that predict many people's cognition, affect, and behavior across time and many situations. Thus, these defining attributes are important. Decades of research have converged on two attribute dimensions, the Big Two, which have been labeled agency and communion (Abele & Wojciszke, 2014), for example, or competence and warmth (Fiske et al., 2002). They capture evaluations of people's prospects of getting ahead in task performance and getting along with others, respectively.

Recently, researchers representing five Big Two models set out to compare their findings on social evaluation in the context of an adversarial collaboration (Ellemers et al., 2020). Integrating theoretical predictions and available evidence allowed them to specify consensus as well as controversies in need of a resolution (Abele et al., 2021; Koch et al., 2021). A key insight from this adversarial collaboration was that existing research focused on different contexts, modes, and types of social evaluation, and different types and numbers of targets and perceivers. Future research should test whether these differences explain the heterogeneity in previous findings, and thereby resolve controversies and integrate theorizing about social evaluation. This future research requires a validated and agreed-upon measure of the Big Two.

The adversarial collaborators distinguished between two facets of vertical evaluation (ability and assertiveness; see Carrier et al., 2014) and two facets of horizontal evaluation (friendliness and morality; see Brambilla et al., 2012; Leach et al., 2007). This consensual differentiation of each Big Two dimension into two facets (see also Abele et al., 2008, 2016) is worthwhile. People describe groups based on their morality before friendliness, and based on their ability before assertiveness (Gligorić et al. 2022, Nicolas et al., 2022). Self-rated friendliness predicts life satisfaction better than self-rated morality, and self-rated assertiveness predicts

✉ Alex Koch
alex.koch@chicagobooth.edu

[1] University of Chicago, Chicago, IL, USA
[2] Princeton University, Princeton, NJ, USA
[3] University of Erlangen, Erlangen, Germany
[4] Utrecht University, Utrecht, Netherlands
[5] University of Louvain, Louvain, Belgium

Published online: 04 September 2024

self-efficacy and self-esteem better than self-rated ability (Abele, 2022; Abele & Hauke, 2020). Then again, impressions of another individual's ability and morality better predict their esteem/reputation than impressions of their assertiveness and friendliness (Abele & Hauke, 2020).

The present research validated a brief measure of these facets of social evaluation, see Appendices 1 and 2. The measure captures the horizontal friendliness facet with the items "warm" and "friendly," the horizontal morality facet with the items "honest" and "sincere," the vertical ability facet with the items "capable" and "skilled," and the vertical assertiveness facet with the items "confident" and "determined." The brief measure may prove useful for developing theory about social evaluation and comparing findings across different paradigms and labs.

Building on existing research

Three papers (two published, one under review) made a laudable effort to validate a measure of the facet model (Abele et al., 2016; Abele & Hauke, 2020; Barbedor et al., 2024). Our validation complements this previous and ongoing research in important ways that we discuss below and in no particular order (i.e., they are equally important).

First, two of the papers (Abele et al., 2016; Barbedor et al., 2024) factor-analyzed the ratings for one type of target by different perceivers. This target-centered analysis asks, for example, "if one target is rated as more friendly by one (vs. another) perceiver, is that same target rated as more moral by the first (vs. second) perceiver?" We factor-analyze the ratings for different targets by one perceiver. This perceiver-centered analysis asks, "if one perceiver rates one (vs. another) target as more friendly, does that same perceiver rate the first (vs. second) target as more moral?" Some models of social evaluation prefer the target-centered analysis (e.g., Abele & Wojciszke, 2014; Ellemers, 2017), whereas other models prefer the perceiver-centered analysis (Fiske, 2018; Koch et al., 2016; Yzerbyt, 2018). It is important to generalize the validity of the facet model from the first to the second approach.

Second, all three papers validated an exhaustive five-items-per-facet measure (Abele et al., 2016; Barbedor et al., 2024) or four-items-per-facet measure (Abele & Hauke, 2020). We validate a more parsimonious two-items-per-facet measure, because a five-items-per-facet measure is not always feasible in studies in which perceivers evaluate several or even many targets on all four facets as well as upstream and/or downstream variables. For example, evaluating ten targets on 20 items plus ten items that measure one upstream and one downstream construct makes 300 items and takes roughly 30 min (unless the study subsamples items for each participant). Thirty minutes is a study duration that arguably erodes data quality and is costly for the researcher(s) if they aim to detect a small effect and thus need to compensate many participants.

Third, the three papers (Abele et al., 2016; Abele & Hauke, 2020; Barbedor et al., 2024) validated the facet model when perceivers rated one target per survey page on all items. Some theorizing focuses on this separate mode of social evaluation (e.g., Abele & Wojciszke, 2014; Nicolas et al., 2022), which prioritizes depth. Other theorizing focuses on a joint mode of social evaluation (i.e., rating all targets on one item per survey page; e.g., Imhoff et al., 2018; Judd et al., 2019; Koch et al., 2020a, 2020b), which prioritizes breadth. Separate versus joint evaluation can reverse preferences and have other interesting effects (Bazerman et al., 1999; Hsee et al., 1999), but we validate our brief measure of the facet model for both modes.

Fourth, the three papers (Abele et al., 2016; Abele & Hauke, 2020; Barbedor et al., 2024) validated the facet model when perceivers expressed personal evaluations of targets. We generalize the validity of the brief measure and facet model from personal evaluations to personal estimates of cultural evaluations (i.e., the perceivers' estimates of how people in society evaluate the targets, on average) and cultural evaluations proper (i.e., the perceivers' evaluations of the targets, on average). The size and sign of the statistical relation between two social-evaluative dimensions/facets can vary as a function of (dis)aggregating ratings for targets (Imhoff & Koch, 2017; Koch et al., 2020a, 2020b; Oliveira et al., 2020; Stolier et al., 2018). Thus, a measure of the facet model needs to be validated for both the personal and cultural type of social evaluations.

Fifth, ideally, the facet model applies to perceivers' evaluations of various types of targets. Two of the three papers validated the facet model across evaluations of the self (Abele et al., 2016) and other individuals, including close friends, acquaintances, and celebrities (Abele & Hauke, 2020). We replicate this and generalize the facet model to evaluations of large and society-representative samples of groups (i.e., social categories based on gender, age, race, status, beliefs, etc.; for a description of how the groups were selected, see Koch et al., 2016). We note that the third paper validated the facet model across evaluations of eight social categories (preselected from a study by Koch et al., 2016) plus intimacy and task groups (Barbedor et al., 2024), which differ from social categories (e.g., the entitativity of intimacy/task groups is higher; Lickel et al., 2000). Further, the three papers validated the facet model across perceivers' ratings for familiar and labeled targets (e.g., "think of an acquaintance of yours"). We also validated perceivers' ratings for unknown targets that they encountered by seeing a photo of them.

Table 1 Overview of the validation of the brief measure and facet model

Study	No. of perceivers	Type of evaluation	No. & type of targets	No. of items per facet	Mode of evaluation	Type of validity
1a 1b	4007	Personal & Cultural	Labels of 20 societal groups	2	Separate & Joint	Internal, convergent & discriminant, external
2	1502	Personal & Cultural	Labels of self, friend, acquaintance, celebrity	2	Separate & Joint	Internal, convergent & discriminant, external
3	1054	Cultural	Photos of 1000 strangers	2	Joint	Internal, convergent & discriminant, external
4	399	Cultural	16 labels of high/low scorers on 8 facet items	2	Joint	Predictive (sensitivity & specificity)
5a–5d	1225	Cultural	Photos of 1000 strangers	2	Joint	Predictive (sensitivity & specificity)

Personal vs. cultural = impressions by individual vs. aggregated perceivers. Separate vs. joint = impressions of one target (i.e., focus on depth) vs. many targets (i.e., focus on breadth) per survey page. Internal validity: the simple-structured model with four correlated facets measured with two items, each fitting the data well enough and better than alternative models. Convergent & discriminant validity: the facets correlate more strongly within (vs. between) the big two dimensions. External validity: the brief measure and facet model generalize across different types and numbers of targets and types of evaluation (i.e., personal vs. cultural) and modes of evaluation (i.e., separate vs. joint). Predictive validity (sensitivity & specificity): each facet predicts some behavior intentions and one behavior intention better than all other facets, with impressions of a target's friendliness, morality, ability, and assertiveness predicting that the perceiver relies on their loyalty, takes their advice, invests in their performance, and exploits their hubris, respectively

Sixth, the three papers validated the facet model by predicting perceivers' evaluations of targets on dimensions that are upstream or downstream of the perceivers' evaluations of the targets on the four facets (i.e., life satisfaction, self-efficacy, entitativity, and similarity/identification/likeability; Abele et al., 2016; Abele & Hauke, 2020; Barbedor et al., 2024). In addition, we validate the facet model by predicting perceivers' behavior intentions towards the targets that they evaluated. We note that ongoing research aims to validate the Big Two (i.e., not the facet model) by predicting perceivers' incentivized behavior towards targets (Walsh et al., 2023).

Table 1 shows the several ways in which our five studies validated the brief measure and facet model as we had suggested when we had consensually endorsed the facet model (Abele et al., 2021; Ellemers, 2017; Koch et al., 2021).

Overview of the validation

Internal validity In three studies, we modeled perceivers' impressions of targets on eight items. We selected the eight items from a list of 20 items; see Supplemental Study 1 (perceivers evaluated groups) and Supplemental Study 2 (perceivers evaluated individuals). Each manifest (i.e., measured) item loaded onto one latent (i.e., estimated) facet. Thus, we tested our assumption that the cause of the variance in each item was variance in one facet and no other facet. We estimated four facets from two measured items each, and we estimated correlations between the facets. This simple-structured model with four facets fit the data well enough, according to standard cutoffs (Hu & Bentler, 1999). Importantly, the model fits the data better than a simple-structured model with two correlated dimensions and four items per dimension (i.e., the Big Two). These eight items were the same as in the better and satisfactory model.

Convergent and discriminant validity Friendliness and morality correlated more strongly than the correlations between friendliness and the two vertical facets, and more strongly than the correlations between morality and the two vertical facets. In addition, ability and assertiveness correlated more strongly than the correlations between ability and the two horizontal facets, and more strongly than the correlations between assertiveness and the two horizontal facets. This pattern of correlations emerged across the three studies and for both the latent facets that we estimated and the manifest scales that measured the facets (e.g., the items "warm" and "friendly" formed the scale that measured the friendliness facet). This pattern corroborated our theorizing about the Big Two such that friendliness and morality are horizontal facets, whereas ability and assertiveness are vertical facets (Abele et al., 2016, 2021; Koch et al., 2021).

For simplicity and to emphasize the facets over the Big Two, the main text reports and interprets the pattern of correlations between the facets rather than a higher-order model in which the facets (do not correlate but) load on their theoretically designated dimension of the Big Two (which correlate). The supplement reports the higher-order model,

which fit as well as the simpler model of our choice (i.e., the model with four correlated facets); see Tables SS1.1, SS2.1, S1.1–S1.3, S2.1, and S3.1. The only research that compared the fit of the two models also found that they fit equally well (Barbedor et al., 2024).

Ecological and external validity The targets that the perceivers rated were labels of many groups (e.g., "Christians" and "drug addicts"), the self and labels of a few self-selected individuals (i.e., a close friend, acquaintance, and celebrity; see Abele & Hauke, 2020), or pictures of many individuals as they appeared on social media recently (see Connor et al., 2024; Gallardo et al., 2024). These perceivers and targets were fairly representative of today's US society. Further, the perceivers prioritized depth by rating all the items within the targets (i.e., one target per survey page; the separate mode of social evaluation), or they prioritized breadth by rating all the targets within the items (i.e., one item per page; the joint mode). In addition, they rated their personal impressions of the targets or cultural (i.e., society's) impressions of the targets (e.g., Fiske et al., 2002), or we computed cultural impressions by averaging the perceivers' ratings separately for each target and item.

The internal, convergent, and discriminant validity of the brief measure generalized across these different types and numbers of targets and modes of social evaluation. This established the external and ecological validity of the brief measure and facet model.

Predictive validity (sensitivity and specificity) In two additional studies, we predicted perceivers' self-rated behavior intentions towards targets in four economic games. Each game captured a different and broadly relevant interpersonal behavior. In each model, the four rivaling predictors were the perceivers' impressions of the targets on the four facets captured with our brief measure. Each facet was sensitive in the sense that it predicted some type of behavior intention. In addition, each facet was specific in the sense that it predicted one type of behavior intention better than all three other facets in at least one of the two studies. The four pairs of a facet and the type of behavior intention that it predicted best corroborated our novel theorizing.

Friendliness and loyalty. In the first game, the perceiver decided between relying on unalterable luck (i.e., their fate) and a target who would decide between earning the perceiver a bonus in an act of loyalty or killing the perceiver's bonus in an act of revenge. Results showed that the perceiver's impression of a target's friendliness predicted the perceiver's reliance on the target (i.e., their loyalty) better than the perceiver's impressions of the target's morality, ability, or assertiveness.

Morality and deception. In the second game, the perceiver decided between a fixed bonus and taking the advice of a target who had decided between giving the perceiver honest information in the best interest of the perceiver's bonus, or deceptive information in the best interest of the target's bonus (Gneezy, 2005). Results showed that the perceiver's impression of a target's morality predicted the perceiver's taking of the target's advice better than the perceiver's impressions of the target's friendliness, ability, or assertiveness.

Ability and investment. In the third game, the perceiver decided between investing a smaller or larger part of their bonus in a target who would earn the perceiver double what they invested if the target solved an intellectual puzzle correctly. The target would kill the perceiver's investment if the solution was incorrect. Results showed that the perceiver's impression of a target's ability predicted the perceiver's large investment in the target better than the perceiver's impressions of the target's friendliness, morality, or assertiveness.

Assertiveness and hubris. In the fourth game, the perceiver decided whether to make a bold (i.e., higher risk–higher reward/bonus) bet that a target would push their luck very far in a risk-taking game (Lejuez et al., 2002). Results showed that the perceiver's impression of a target's assertiveness predicted the perceiver's bold bet on the hubris of the target better than the perceiver's impressions of the target's friendliness, morality, or ability.

Open science and scope of validity

All studies had institutional review board approval. We preregistered Studies 4 (https://aspredicted.org/GW8_XKT) and 5a–5d (https://aspredicted.org/K31_VC5, https://aspredicted.org/QDN_M9Y, https://aspredicted.org/YH3_YNV, and https://aspredicted.org/T7T_Z5S, respectively). In each study, we collected all data before analyzing any of them. All study materials and data, code, and results are available on the Open Science Framework website (https://osf.io/sdvwt/). We sampled US residents from the online worker platform Prolific Academic, whose participation had been approved at a rate of at least 97%. Prolific workers' representativeness of the US population is decent (Douglas et al., 2023; Peer et al., 2017). We do not generalize our results beyond the United States. Other research validated the facet model for several European societies as well as China and Australia (Abele et al., 2016; Abele & Hauke, 2020; Barbedor et al., 2024).

Studies 1a and 1b

We aimed to corroborate the internal, convergent, and discriminant validity of a two-items-per-facet measure of social evaluation that we explored in Supplemental Study 1. People rated groups separately or jointly, and they expressed personal evaluations or estimated society's consensual (i.e., cultural) evaluations of the groups. We aimed to generalize the measure across these broadly relevant modes and types of social evaluation, to establish the external validity of the measure and show that it applies across the different social-evaluative contexts that our adversarial models of social evaluation focus on (Abele et al., 2021; Ellemers et al., 2020; Koch et al., 2021).

Method

Participants Studies 1a and 1b ran at different points in time. However, people rated the same groups on the same items, and thus we pooled people's data across Studies 1a and 1b for brevity and conciseness. Across Studies 1a and 1b, we sampled 4007 people from Prolific. We excluded 71 people who recommended that we not analyze their data,[1] leaving 3934 people (49.0% female, 49.0% male, 1.9% other; $M_{age} = 31.62$, $SD = 12.16$).

Stimuli People rated the 20 groups that other people in previous research had listed most frequently when instructed to list the groups that together form today's US society (see Koch et al., 2016). The 20 groups were defined by their gender, age, race, status, beliefs, etc. We list them in alphabetical order: Asian people, Black people, Christians, conservatives, Democrats, elderly people, gay people, Hispanic people, lesbian people, liberals, middle-class people, poor people, Republicans, rich people, students, transgender people, upper-class people, White people, women, and working-class people. These groups are social categories, according to a categorization by Lickel and colleagues (2000; these authors also mention intimacy groups [e.g., friends], task groups [e.g., a committee], and loose associations [e.g., riders on a bus]; research by Barbedor and colleagues [2024] validates the facet model for intimacy and task groups, in addition to a few social categories).

Procedure People used seven-point scales that ranged from 1 = "not at all" to 7 = "extremely" to rate each of the 20 groups on 10 items. The items "friendly" and "warm" aimed to measure the perceived friendliness of the groups, "honest" and "sincere" aimed to measure the perceived morality of the groups, "capable" and "skilled" aimed to measure the perceived ability of the groups, and "confident" and "determined" aimed to measure the perceived assertiveness of the groups. Finally, "positive" and "good" measure people's general evaluation of the groups. The analyses omit the latter two items.

People rated their personal evaluations of the groups or their estimates of society's consensual (i.e., cultural) evaluations of the groups. In addition, people rated one group on all ten items (in random order) before rating another group on the next survey page (i.e., separate evaluations), with the order of the groups being random as well. Alternatively, people rated all groups (in random order) on one item before rating them on another item on the next page (i.e., joint evaluations), with the order of the items being random as well. In the separate mode, people read "to what extent do you think of [group; e.g., Asian people] as [items]?" (personal evaluations) or "to what extent do most Americans think of [group] as [items]?" (cultural evaluations). In the joint mode, people read "to what extent do you think of the following groups as [item; e.g., FRIENDLY]?" (personal evaluations) or "to what extent do most Americans think of the following groups as [item]" (cultural evaluations).

Finally, people provided demographic information, including their gender and age.

Results

We used the cfa function of the R package lavaan (Rosseel, 2012) to run four multilevel confirmatory factor analyses (MCFAs; Pornprasertmanit et al., 2014). The four MCFAs modeled all separate evaluations (by 1955 people), all joint evaluations (by 1979 people), all personal evaluations (by 1962 people), or all cultural evaluations (by 1972 people). In each MCFA, we defined the raters as data clusters (i.e., we treated the groups as nested within the raters), to model the differences between the groups within the raters (vs. modeling the differences between the raters within the groups as in previous work).

Level 1—our main interest—had the groups as rows, the items as columns, and individual ratings as the data. The manifest items "friendly" and "warm" loaded on the latent friendliness facet, "honest" and "sincere" loaded on the morality facet, "capable" and "skilled" loaded on the ability facet, and "confident" and "determined" loaded on the assertiveness facet. For the sake of internal validity, we allowed no cross-loadings (i.e., we modeled a simple structure), but we allowed the latent facets to correlate with one another. Level 2 had the raters as rows, the items as columns, and mean ratings across the groups as the data. All eight

[1] Across all studies, we asked participants whether they had paid attention throughout the study and whether they recommend that we use their data. We consider this an important step to ensure that we are not capturing responses from inattentive participants. Importantly, the exclusion rate never exceeded 7.5% for any study.

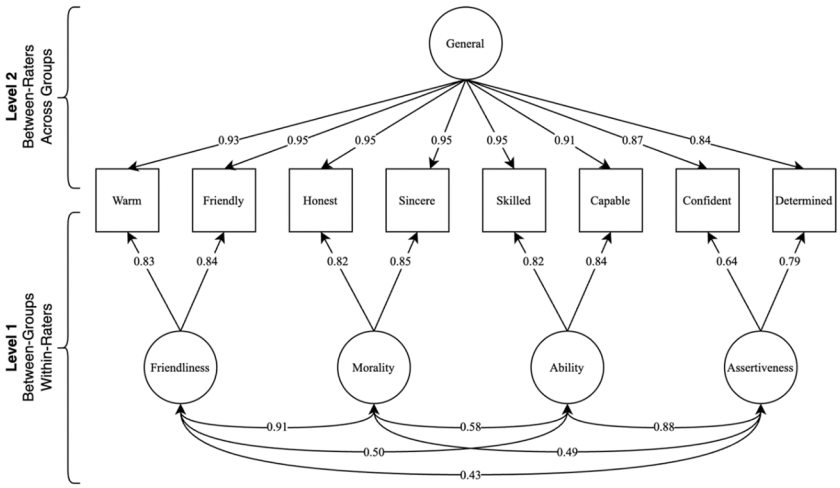

Fig. 1 Studies 1a and 1b: Parameters of a two-items-per-facet measure of cultural evaluations

Table 2 Studies 1a and 1b: Satisfactory and superior fit of a two-items-per-facet model

	χ^2	CFI	RMSEA	SRMR	AIC	$p(\chi^2)$
Two items per facet; *df* = 34						
Separate evaluations	1649	0.99	0.03	0.06	885,536	
Joint evaluations	2021	0.99	0.04	0.07	935,804	
Personal evaluations	1938	0.99	0.04	0.08	876,224	
Cultural evaluations	1042	0.99	0.03	0.05	944,668	
Four items per dimension; *df* = 39						
Separate evaluations	7923	0.96	0.07	0.08	891,800	< .001
Joint evaluations	3790	0.97	0.05	0.08	937,562	< .001
Personal evaluations	5547	0.97	0.06	0.09	879,823	< .001
Cultural evaluations	4077	0.98	0.05	0.06	947,692	< .001

$p(\chi^2)$ represents the *p*-value from a nested chi-square difference test compared to the respective two-items-per-facet model

manifest items loaded on a latent general factor because we had no hypothesis for the factor structure of level 2, and thus decided to keep it simple. Level 2's latent general factor captured that some people gave higher ratings to all groups on all items due to acquiescence, complaisance (i.e., wanting to be liked), philanthropy, or a combination of these (Rau et al., 2021). Figure 1 shows the estimated parameters of the measure of cultural evaluations. We visualize these parameters in the main text because the measure of cultural evaluations fit the data best. However, we also visualize the estimated parameters of the measures of personal, separate, and joint evaluations in Figs. S1.1–3 in the supplement.

We report multiple fit indices (chi-square [χ^2], comparative fit index [CFI], root mean square error of approximation [RMSEA], standardized root mean square residual [SRMR], and Akaike information criterion [AIC]) to account for the limitations inherent in relying on a single index (Kline, 2005). We compare the fit of the measure against the standard cutoffs of ≥ 0.95 for CFI and ≤ 0.06 for RMSEA and SRMR (Hu & Bentler, 1999). Table 2 shows that the two-items-per-facet measure fit the data well in both social-evaluative modes (i.e., separate vs. joint) and types (i.e., personal vs. cultural), except for the SRMR index in the joint mode and personal type.

Table 2 also shows the fit of a measure in which the manifest items "honest," "sincere," "friendly," and "warm" loaded on a latent horizontal dimension, and "capable," "skilled," "confident," and "determined" loaded on a vertical dimension that we allowed to correlate with the horizontal dimension. The fit of the four-items-per-dimension measure (which refrained from differentiating the Big Two into the four facets) was worse than the fit of the two-items-per-facet measure in both modes and both types of social evaluation according to nested chi-square difference tests (all *p*s < 0.001). Figures S1.4–7 show the estimated parameters of all four four-items-per-dimension measures.

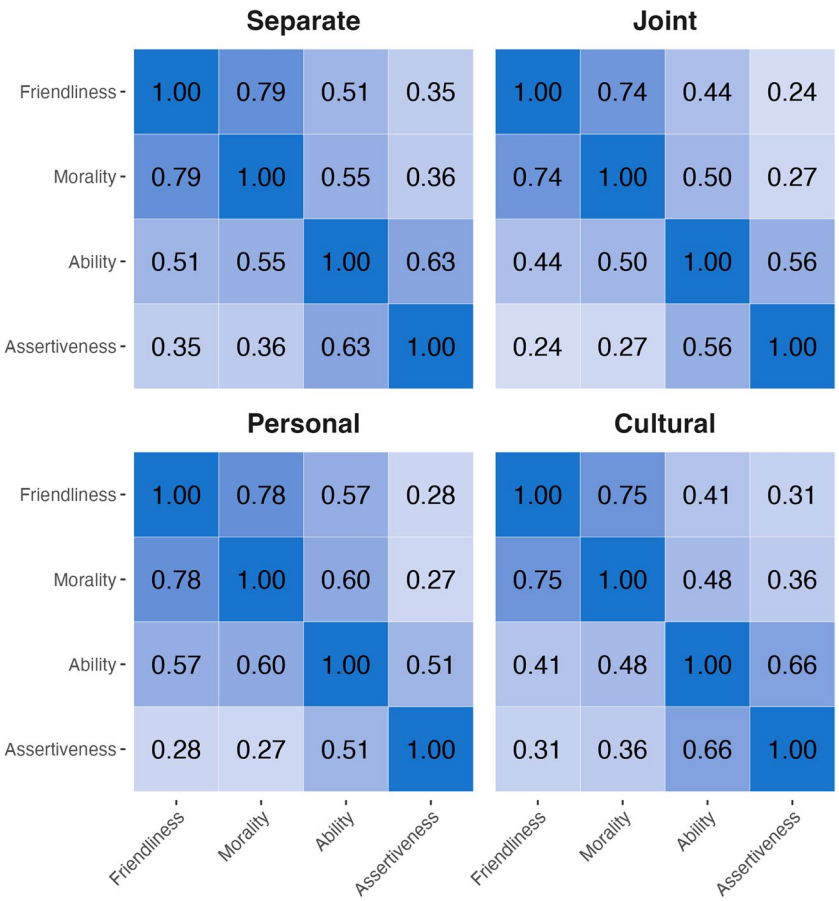

Fig. 2 Studies 1a and 1b: Correlations between the four facet scales centered within the raters. *Note*. More intense hues indicate larger correlations

Next and separately for each mode of social evaluation, we computed a friendliness scale by averaging the two manifest friendliness items separately for each group within each rater. Likewise, we computed a morality, ability, and assertiveness scale. For each scale, we centered the differences between the groups within each rater as in the MCFAs. Figure 2 shows the correlations between the within-centered facet scales in both modes and both types of social evaluation (see Tables S1.3 and S1.4 for the correlations in Study 1a separately from Study 1b). To conclude sufficient convergent and discriminant validity, the correlations between the friendliness and morality (i.e., horizontal) facets and the correlations between the ability and assertiveness (i.e., vertical) facets had to be positive and larger than all correlations between one horizontal facet and one vertical facet. The data confirmed this, except that in the personal mode of social evaluation, the correlation between the morality and ability facets, and between the friendliness and ability facets, was slightly larger than the correlation between the ability and assertiveness facets. However, the correlations between the latent facets in the two-items-per-facet measure confirmed the sought-after pattern in both modes and both types of social evaluation; see Fig. 1 and Figs. S1.1–3.

Figures S1.8–23 and Tables S1.2–5 show the results of the two studies when analyzing their people/data separately. The results fully replicate the below pattern of results

Discussion

Based on goodness of absolute and relative model fit as well as correlations between both manifest scales and latent factors, Studies 1a and 1b largely confirmed the validity of our efficient two-items-per-facet measure of four correlated facets of social evaluation (ability, assertiveness, morality, and friendliness; Abele et al., 2021; Koch et al., 2021), and demonstrated its robustness across two modes and two types of evaluating many groups (i.e., separate or joint evaluations, and personal or cultural evaluations).

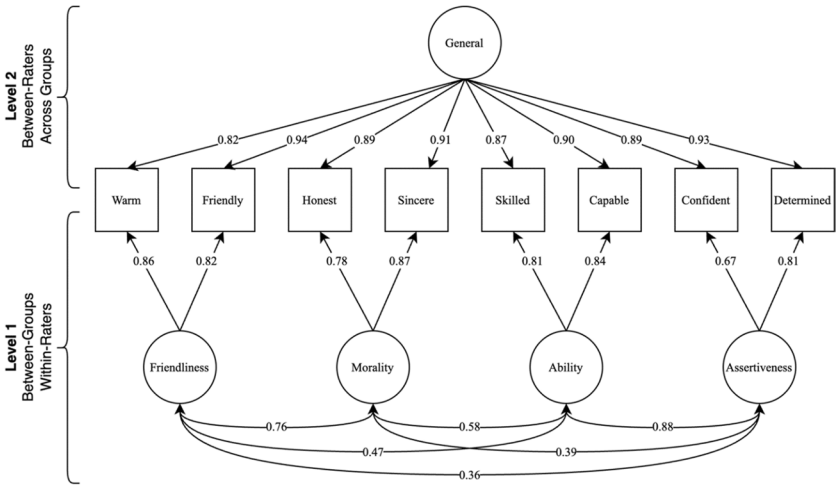

Fig. 3 Study 2: Parameters of a two-items-per-facet measure of estimated cultural evaluations

Study 2

We aimed to generalize our two-items-per-facet measure across separate, joint, personal, and cultural evaluations of several individuals per rater, to further corroborate the external validity of the measure. Put differently, Study 2 aimed to replicate the results of Studies 1a and 1b, except that people evaluated several individuals as in previous work (Abele & Hauke, 2020; see also Supplemental Study 2), instead of many groups.

Method

Participants We sampled 1502 people from Prolific. We excluded 54 people who recommended that we not analyze their data, leaving 1448 people (49.2% female, 49.0% male, 1.7% other; $M_{age} = 31.46$, $SD = 11.55$).

Stimuli and procedure People began by typing into text boxes the names of a same-sex close friend, acquaintance, and celebrity. These targets of social evaluation differ in terms of their closeness to the perceiver and positivity in the eyes of the perceiver, two continua that characterize the focus of many, if not all, models of social evaluation (Abele et al., 2021; Koch et al., 2021). As in Studies 1a and 1b, they separately or jointly evaluated these individuals, themselves, and these groups on the same 10 items as in Studies 1a and 1b. In addition, people rated their personal evaluations of the social entities or their estimates of society's consensual (i.e., cultural) evaluations of the social entities.

At the end of the study, people provided demographic information, including their gender and age.

Results

We used the cfa function of the R package lavaan (Rosseel, 2012) to run four multilevel confirmatory factor analyses (MCFAs; Pornprasertmanit et al., 2014) on people's evaluations of the four individuals. The four MCFAs modeled all separate evaluations (by 729 people), all joint evaluations (by 719 people), all personal evaluations (by 723 people), or all cultural evaluations (by 725 people). In each MCFA, we defined the raters as data clusters (i.e., we treated the individuals as nested within the raters), to model the differences between the individuals within the raters. We specified levels 1 and 2 in the same way as for the two-items-per-facet measures in Studies 1a and 1b. Figure 3 shows the estimated parameters of the measure of cultural evaluations. Figures S2.1–3 show the estimated parameters of the other three measures (i.e., separate, joint, and personal evaluations).

We report multiple fit indices, and we compare the fit of the measure against the standard cutoffs of ≥ 0.95 for CFI and ≤ 0.06 for RMSEA and SRMR (Hu & Bentler, 1999). Table 3 shows that the two-items-per-facet measure fit the data well in both modes of social evaluation (i.e., separate and joint) and both types of social evaluation (i.e., personal and cultural), except for the SRMR index.

Table 3 also shows that a four-items-per-dimension measure (that we specified as in Studies 1a and 1b) fits the data worse than the hypothesized two-items-per-facet measure in both modes and both types of social evaluation according to nested chi-square difference tests (all $ps < 0.001$). Figures S2.4–7 show the estimated parameters for the four worse-fitting four-items-per-dimension measures.

As before, we had hypothesized that the correlation between the friendliness and morality (i.e., horizontal) facets and the correlation between the ability and assertiveness

Table 3 Study 2: Satisfactory and superior fit of a two-items-per-facet model

	χ^2	CFI	RMSEA	SRMR	AIC	$p(\chi^2)$
Two items per facet; $df = 34$						
Separate evaluations	216	0.99	0.04	0.08	65,349	
Joint evaluations	204	0.99	0.04	0.09	65,000	
Personal evaluations	248	0.98	0.05	0.09	65,150	
Cultural evaluations	164	0.99	0.04	0.07	65,306	
Four items per dimension; $df = 39$						
Separate evaluations	1005	0.92	0.09	0.12	66,127	< .001
Joint evaluations	779	0.93	0.08	0.12	65,565	< .001
Personal evaluations	967	0.92	0.09	0.11	65,858	< .001
Cultural evaluations	834	0.93	0.08	0.11	65,967	< .001

$p(\chi^2)$ represents the p-value from a nested chi-square difference test compared to the respective two-items-per-facet model

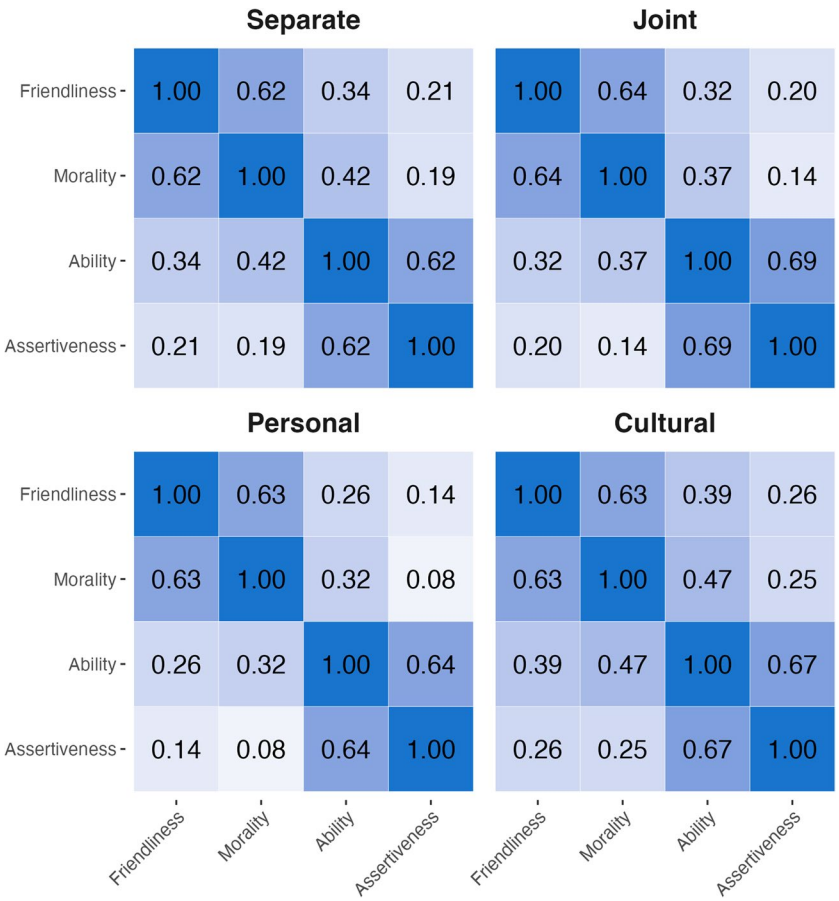

Fig. 4 Study 2: Correlations between the four facet scales centered within the raters. *Note*. More intense hues indicate larger correlations

(i.e., vertical) facets would be positive and larger than any correlation between one horizontal facet and one vertical facet. The data confirmed this pattern of correlations in both modes and both types of social evaluation when we computed within-centered facet scales according to the two-items-per-facet measure; see Fig. 4. The correlations between the latent facets in the four two-items-per-facet measures also confirmed the hypothesized pattern; see Fig. 3 and Figs. S2.1–3.

Discussion

Based on goodness of absolute and relative model fit as well as correlations between both manifest scales and latent factors, Study 2 largely confirmed the validity our efficient two-items-per-facet measure of four correlated facets of social evaluation (ability, assertiveness, morality, and friendliness; Abele et al., 2021; Koch et al., 2021). We confirmed the internal, convergent, and discriminant validity of the measure across two modes and two types of evaluating several individuals (i.e., separate or joint evaluations, and personal or cultural evaluations). This further corroborated the external validity of the measure. Studies 3–5 captured cultural evaluations because the measure of cultural (vs. personal, separate, and joint) evaluations fit the data best in Studies 1 and 2.

Study 3

We aimed to generalize the two-items-per-facet measure from personal estimates of cultural evaluations (Studies 1a–2) to cultural evaluations proper, and from evaluations of labels of familiar and groups or individuals (Studies 1a–2) to evaluations of unknown individuals that people encountered by seeing a photo of them. The latter generalization matters because people constantly meet strangers (e.g., on the street, at professional and private events, and when browsing social media and news platforms online) that they need to evaluate precisely and swiftly to make opportunities and solve problems.

Method

Participants We sampled 1054 people from Prolific. We excluded 69 people who recommended that we not analyze their data, leaving 985 people (41.5% female, 56.3% male, 1.7% other, 0.4% preferred not to answer; $M_{age} = 41.28$, $SD = 13.09$).

Stimuli Each person rated a random selection of 100 out of 1000 individuals based on their public-facing Facebook profile picture in 2021. Most of these pictures provide a glimpse at an individual's real life (e.g., their workplace, social network, habits, hobbies, etc.). Thus, seeing someone's Facebook profile picture is a more ecologically valid way of encountering them, compared to a passport photograph. We created the set of 1000 pictures using a quasi-random procedure. In each of 1000 trials, we first searched for a randomly selected US city using Facebook's search engine. Second, we selected the first Facebook page that was not the city's page, was located in the United States, and had at least 300 likes. Third, we selected the profile picture of the individual who liked that page most recently if they were the only (or focal) person in the picture, if their gender, age, and race were discernible, and if they resided in the United States as indicated in their profile's "About Info." We coded the gender (woman or man), age (young, middle-aged, or old), and race/ethnicity (White, Black, Latino/a, East Asian, or South Asian) of each picture (62.5% women, 37.5% men; 32.7% young, 51.8% middle-aged, 15.5% old; 81.5% White, 8.9% Black, 5.1% Latino/a, 2.8% East Asian, 1.7% South Asian).

Procedure On separate survey pages, people saw a picture of an individual, and below they read "To what extent do you consider this person to be CAPABLE [or another item]?" They used a seven-point scale that ranged from 1 = "not at all" to 7 = "extremely" to rate the individual. They rated 100 individuals in random order and on one and the same item that we randomly selected from the eight items that we examined in Studies 1a–2 plus "positive" and "good." At the end of the survey, people provided demographic information including their gender and age.

Results

For each of the 1000 individuals that people had rated, we computed their mean rating on the eight items that we examined in Studies 1a–2 (e.g., "friendly" and "warm"). Next and based on a matrix that had the individuals as rows, the items as columns, and their mean ratings as the data, we used the cfa function of the R package lavaan (Rosseel, 2012) to run a single-level confirmatory factor analysis (CFA). (It was neither possible nor appropriate to run a multilevel confirmatory factor analysis [MCFA] in Study 3 because Study 3's data had only one row per individual [vs. several/many rows per group/individual] in Studies 1 and 2 and Supplemental Studies 1 and 2.) The manifest items "friendly" and "warm" loaded on the latent friendliness facet, "honest" and "sincere" loaded on the morality facet, "capable" and "skilled" loaded on the ability facet, and "confident" and "determined" loaded on the assertiveness facet. We allowed no cross-loadings (i.e., we modeled a simple structure), but we allowed the latent facets to correlate with one another. Figure 5 shows the estimated parameters of this two-items-per-facet measure of cultural evaluations.

We report multiple fit indices, and we compare the fit of the measure against the standard cutoffs of ≥ 0.95 for CFI and ≤ 0.06 for RMSEA and SRMR (Hu & Bentler, 1999). Table 4 shows that the two-items-per-facet measure fit the data well enough according to the SRMR index. A four-items-per-dimension measure (that we specified as in Studies 1a–2) fit the data worse according to nested chi-square difference tests (all $ps < 0.001$). Figure S3.1 shows the estimated parameters of the worse-fitting measure.

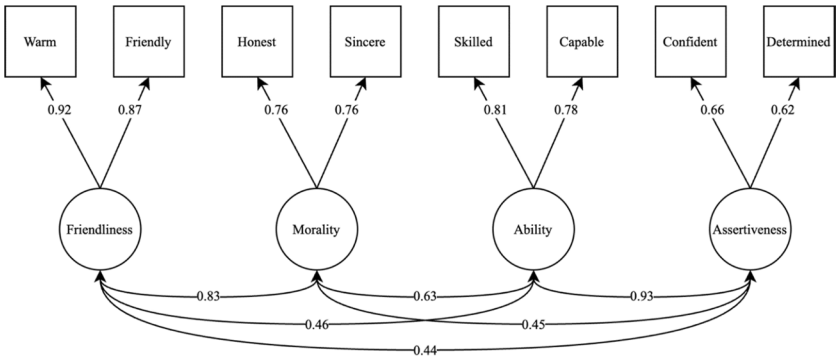

Fig. 5 Study 3: Parameters of a two-items-per-facet measure of cultural evaluations

Table 4 Study 3: Barely satisfactory and superior fit of a two-items-per-facet model

	χ^2	CFI	RMSEA	SRMR	AIC	$p(\chi^2)$
Two items per facet; $df = 14$	230	0.94	0.12	0.06	− 12,512	
Four items per dimension; $df = 19$	394	0.90	0.14	0.08	− 12,362	< .001

$p(\chi^2)$ represents the p-value from a nested chi-square difference test compared to the respective two-items-per-facet model

As in Studies 1a–2, we had hypothesized that the correlations between the friendliness and morality (i.e., horizontal) facets and the correlations between the ability and assertiveness (i.e., vertical) facets would be positive and larger than any correlation between one horizontal facet and one vertical facet. The data confirmed this; see Fig. 6. The correlations between the latent facets in the two-items-per-facet measure also confirmed this; see Fig. 5.

Discussion

Based on goodness of absolute and relative model fit as well as correlations between both manifest scales and latent factors, Study 3 largely generalized the internal, convergent, and discriminant validity of the two-items-per-facet measure from personal estimates of cultural evaluations (see Studies 1a–2) to cultural evaluations proper, and from evaluations of labels of familiar groups or individuals (Studies 1a–2) to evaluations of unknown individuals that people encountered by seeing a photo of them.

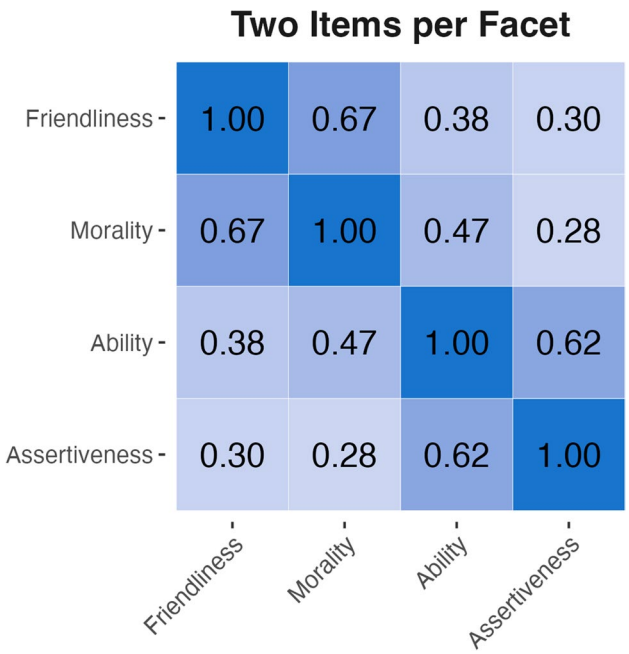

Fig. 6 Study 3: Correlations between the four facet scales. *Note*. More intense hues indicate larger correlations

Study 4

Studies 1a–3 examined items and rating scales that confirmed the facet model's internal, convergent, and discriminant validity (Abele et al., 2021; Koch et al., 2021) across a variety of to-be-evaluated social entities and two modes and two types of social evaluation. Studies 4 and 5 aimed to establish the predictive validity of the facet model. Previous and ongoing validations addressed predictive validity by showing that perceivers' impressions of targets on the

four facets predict other impressions, including the targets' life satisfaction, self-efficacy, entitativity, and likeability as seen by the perceivers (Abele et al., 2016; Abele & Hauke, 2020; Barbedor et al., 2024). In Studies 4 and 5, we went beyond these validations by showing that perceivers' impressions of targets on each facet predicted some behavioral intentions of the perceivers towards the targets (i.e., sensitivity). In addition, each facet predicted a unique behavioral intention better than all three other facets in at least one study (i.e., specificity). We measured the behavioral intentions through economic games because they capture people's willingness to act on their impressions (i.e., put their money where their mouth is) in abstracted ways that speak to many real-life situations (Thielmann et al., 2021).

We reasoned that impressions of a target's friendliness, morality, ability, and assertiveness may guide a perceiver to expect loyalty, fairness, performance, and risk-taking from a target, respectively. Accordingly, in four economic games, we predicted that the perceivers would decide to make their bonus (for participating in the study) dependent on targets' loyal trickery, fair advice, analytical success, and hubristic gambling, respectively.

Method

Participants We sampled 399 people from Prolific. We excluded three people who recommended that we not analyze their data, leaving 396 people (38.9% female, 59.3% male, 0.5% other, 1.3% preferred not to answer; $M_{age} = 41.55$, $SD = 13.25$).

Stimuli Each person envisioned 16 anonymous individuals that we described in terms of one personality trait. One individual was "a friendly person," whereas another individual was "a person who lacks friendliness." Seven other individuals were "a warm [or honest, sincere, capable, skilled, confident, or determined] person," whereas seven other individuals were "a person who lacks friendliness [or warmth, honesty, sincerity, capability, skill, confidence, or determination]." In sum, each person envisioned one high scorer and one low scorer on each of the eight items in our brief measure of the four facets; see Studies 1a–3.

Procedure People played four hypothetical games in random order. Each game was economic (i.e., about winning money) and dyadic, which means that people played each game with each of the 16 anonymous individuals in random order and on the same survey page. In each game played with each individual, people made a choice between two options; see below.

At the end of the survey, people used seven-point scales (1 = "not at all," 7 = "extremely") to rate the self on the eight items and "positive" and "good." Finally, people provided demographic information, including their gender and age.

Loyalty game. People read "You rolled a 6-sided dice and win $10 only if you rolled a 5 or 6. As of now, you do not know the outcome of your dice roll. You can publicly reveal the outcome of the dice roll. Or, you can ask each of 16 persons to secretly learn and report the outcome. They can tell the truth or a lie (no one will ever know). If they report that you rolled a 5 or 6, you get $10. Do you delegate reporting the outcome to each person?" Righteously revealing the outcome was coded as 0, whereas delegating the reporting of the outcome, and thereby expecting loyal trickery, was coded as 1.

Deception game. People read "Each of 16 persons advises you to choose Option B. They all flipped a coin. If the coin landed on heads, the payoffs of Option B will be $3 for you and $7 for them. If the coin landed on tails, the payoffs of Option B will be $7 for you and $3 for them. Everyone messages you 'Tails! Option B pays out $7 for you and $3 for me. Pick Option B.' If you choose Option A, you and they will receive $5 each. How much do you trust the advice of each person?" Skeptically choosing option A was coded as 0, whereas expecting true and fair advice, and thus choosing option B was coded as 1.

Investment game. People read "Your task is to invest $1.50 or $3.50 out of $5 in each of 16 persons. For each person, if they solve the below problem correctly, you will receive double the money you invested plus the money you withheld. If they solve the below problem incorrectly, you will lose the money you invested but receive the money you withheld. They receive $1.50 regardless of their performance. [A randomized medium-difficulty SAT question was inserted here]. How much do you invest in each person?" Pessimistically investing just $1.50 was coded as 0, whereas optimistically expecting analytical success and investing $3.50 was coded as 1.

Hubris game. Adapted from the Balloon Analogue Risk Task (BART; Lejuez), people read "Each of 16 persons is given a balloon to inflate. They earn $1 for each time they pump the balloon, but they lose their earnings if the balloon pops. At each turn, they can pump or stop and collect their earnings. Without them knowing, you decide when the balloon pops. If the balloon pops, you keep all their earnings. If the balloon doesn't pop, you earn nothing. You can pop the balloon on the 4th or 6th pump. If they pop the balloon on the 4th pump, you collect $4 and they collect nothing. If they pop the balloon on the 6th pump, you collect $6 and they collect nothing. If the balloon does not pop, they keep their earnings. On which pump do you pop the balloon?" Cautiously popping the balloon on the fourth pump already was coded as 0, whereas expecting hubristic gambling, and thus boldly popping the balloon on the sixth pump, was coded as 1.

Results

We coded the 16 individuals' high and low scores on the eight items in a way that treated the four facets as orthogonal (i.e., independent). We coded the friendliness of the two individuals who lacked friendliness and warmth as − 0.5, and those who were friendly and warm as 0.5, and we coded the friendliness of the other 12 individuals as 0. We coded the morality of the two individuals who lacked honesty and sincerity as − 0.5, and those who were honest and sincere as 0.5, and we coded the morality of the other 12 individuals as 0. We coded the ability of the two individuals who lacked capability and skill as − 0.5, and those who were capable and skilled as 0.5, and we coded the ability of the other 12 individuals as 0. Finally, we coded the assertiveness of the two individuals who lacked confidence and determination as − 0.5, and those who were confident and determined as 0.5, and we coded the assertiveness of the other 12 individuals as 0.

We used the lmer function of the R package lme4 (Bates et al., 2015) to run a linear mixed model with random intercepts for the people and the 16 individuals that they played with. We predicted the people's binary choices in the loyalty game from the individuals' orthogonal scores on the four facets. Three additional models were the same, except that we predicted people' binary choice in the deception, investment, and hubris game. We subjected each model to dominance analysis (Azen & Budescu, 2003), which we ran using the domin function of the R package domir (Luchman, 2023). Dominance analysis compares regression coefficients based on their sign, size, and scatter.

Figure 7 and the preregistered dominance analysis in Table S4.1 show that in the loyalty game, the friendliness of an individual best predicted that a participant would make their bonus dependent on the individual's loyal trickery. In the deception game, the morality of an individual best predicted that a participant would make their bonus dependent on the individual's fair advice. In the investment game, the ability of an individual best predicted that a participant would make their bonus dependent on the individual's analytical success. In the hubris game, the assertiveness of an individual best predicted that a participant would make their bonus dependent on the individual's hubristic gambling.

Dominance analysis (Azen & Budescu, 2003) compares regression coefficients descriptively (i.e., no inferential statistics). To substantiate the above interpretations with inferential statistics, we used the linearHypothesis function of the R package car (Fox & Weisberg, 2018). Figure 7 shows that in each game, the facet marked in yellow (that we had hypothesized to emerge as the best-predicting facet) actually emerged as the best predictor according to the significance tests reported in Table S4.2. The only exception was the investment game. The ability facet predicted investment better than the assertiveness facet, but this comparison of regression coefficients did not reach statistical significance; $p = 0.065$.

Discussion

Study 4 validated the two-items-by-facet measure by showing that evaluations on each facet predicted broadly relevant behavior towards the evaluated individuals better than evaluations on all three other facets (except the ability facet in the investment game). However, we described the individuals based solely on their high or low score on one item/facet (e.g., "is friendly"), and we orthogonalized the individuals' facet scores. Thus, we take Study 4's results as a proof of concept and acknowledge that the results need to be generalized to real, nuanced social entities whose facet scores vary and covary naturally (i.e., according to the facet model; Abele et al., 2021).

Studies 5a–5d

Studies 5a–5d aimed to validate the facet model by showing that cultural evaluations of real and nuanced individuals on each facet predict a relevant and specific behavior towards these unknown (vs. familiar in Studies 1a–2) and encountered (vs. envisioned in Studies 1a–4) individuals better than cultural evaluations of the individuals on all three other facets.

Method

Participants In Studies 5a–5d, we sampled 304, 307, 307, and 304 people from Prolific. We excluded six, three, three, and one people, respectively, who recommended that we not analyze their data. This left 298 people in Study 5a (43.3% female, 55.0% male, 1.3% other, 0.3% preferred not to answer; $M_{age} = 40.30$, $SD = 13.02$). It left 304 people in Study 5b (44.7% female, 53.0% male, 2.0% other, 0.3% preferred not to answer; $M_{age} = 40.32$, $SD = 12.45$). It left 304 people in Study 5c (50.0% female, 47.4% male, 1.6% other, 1.0% preferred not to answer; $M_{age} = 41.47$, $SD = 13.70$). And it left 303 people in Study 5d (37.6% female, 61.1% male, 1.0% other, 0.3% preferred not to answer; $M_{age} = 40.61$, $SD = 12.04$).

Stimuli Each person encountered a random selection of 100 out of 1000 individuals as they appeared in their public-facing Facebook profile picture in 2021. Study 3 describes their gender, age, and race distribution and the quasi-random inclusion of each individual in the set.

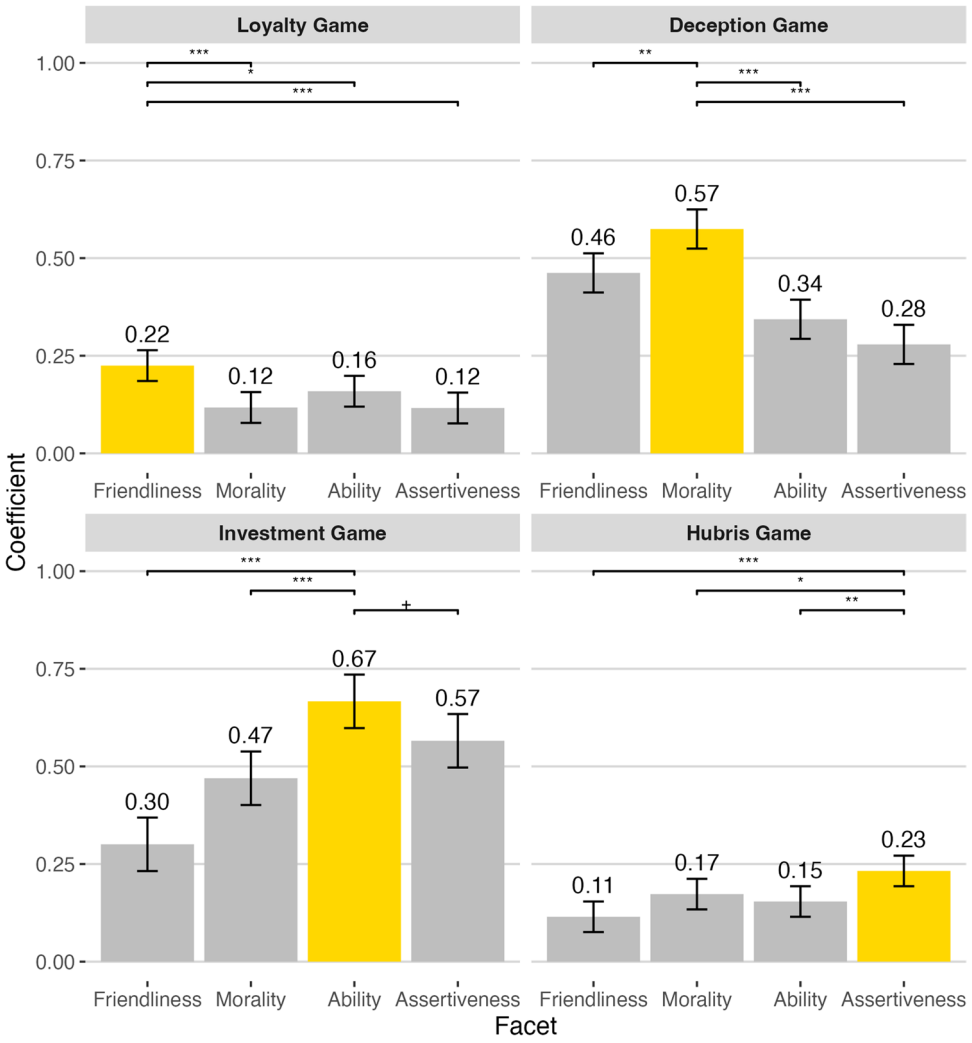

Fig. 7 Study 4: Expecting loyalty, fairness, success, and hubris from envisioned individuals based on their score on ability, assertiveness, morality, and friendliness. *Note*. Error bars represent 95% confidence intervals. The highlighted columns represent the strongest predictors of people's binary choice in each game according to dominance analysis. *** $p < .001$; ** $p < .010$; * $p < .050$; $^{+}$ $p < .100$

Procedure On each of 100 randomly ordered survey pages, people encountered one picture of an individual. Below, they read the instructions of one hypothetical economic game, namely the loyalty, deception, investment, and hubris game that we described in Study 4. Below, they made the same binary choice as in Study 4. That is, they delegated the reporting of a secret dice roll to the individual (vs. revealed it themselves), trusted the advice of the individual (vs. ignored it), invested a larger (vs. smaller) sum in the individual's analytical success, or bet against the individual boldly (vs. cautiously). So, in Studies 5a–5d we used a between-subjects design (vs. the within-subjects design of Study 4 in which people played all four games with the 16 individuals that they envisioned in that study).

Results

We used the lmer function of the R package lme4 (Bates et al., 2015) to run a linear mixed model with random intercepts for the people and the individuals that they played with. We predicted the people's binary choices in the loyalty game from the individuals' correlated scores on the four facets. The individuals' friendliness scores were their mean ratings on the "friendly" and "warm" items that we computed across all people in Study 3 who had rated one of the two items. The individuals' morality, ability, and assertiveness scores were their mean ratings on "honest" and "sincere," "capable" and "skilled," and "confident" and "determined" across all people in Study 3 who had rated one of the two

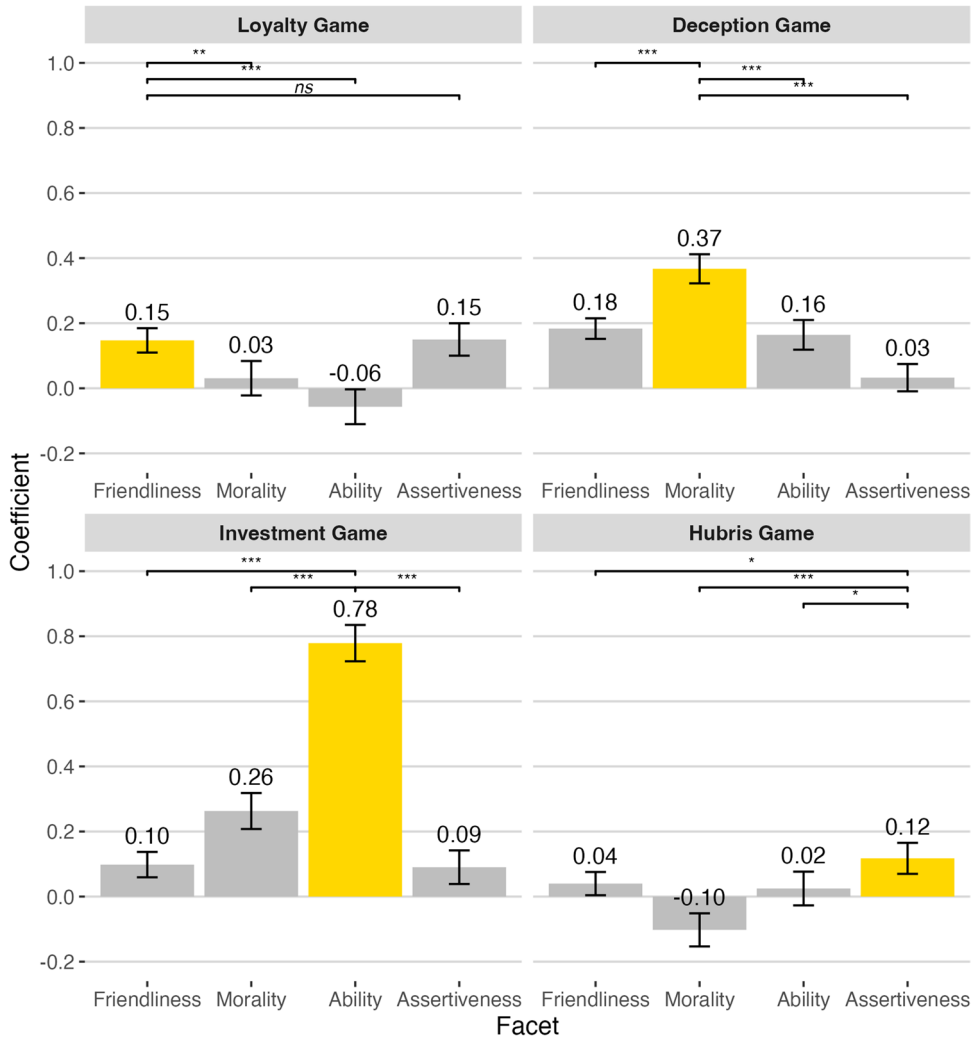

Fig. 8 Studies 5a–5d: Expecting loyalty, fairness, success, and hubris from envisioned individuals based on their score on ability, assertiveness, morality, and friendliness. *Note*. Error bars represent 95% confidence intervals. Highlighted columns represent the relatively most important predictor of each game according to dominance analysis. *** $p < .001$; ** $p < .010$; * $p < .050$; $^{+}$ $p < .100$

respective items. In sum, we predicted the binary choices of the people in Study 5a–d from the cultural evaluations that we had captured with the help of the people in Study 3. In three additional models, we predicted people's binary choice in the deception, investment, and hubris game. We subjected each model to dominance analysis (Azen & Budescu, 2003), which we ran by using the domin function of the R package domir (Luchman, 2023). Again, dominance analysis compares regression coefficients based on their sign, size, and scatter.

Figure 8 and the preregistered dominance analysis in Table S5.1 show that in the loyalty game, the friendliness of an individual best predicted that a participant would make their bonus dependent on the individual's loyal trickery. In the deception game, the morality of an individual best predicted that a participant would make their bonus dependent on the individual's fair advice. In the investment game, the ability of an individual best predicted that a participant would make their bonus dependent on the individual's analytical success. In the hubris game, the assertiveness of an individual best predicted that a participant would make their bonus dependent on the individual's hubristic gambling.

Again, dominance analysis (Azen & Budescu, 2003) compares regression coefficients descriptively (i.e., no inferential statistics). To substantiate the above interpretations with inferential statistics, we used the linearHypothesis function of the R package car (Fox & Weisberg, 2018). Figure 8 shows in each game, the facet marked in yellow (that we had hypothesized to emerge as the best-predicting facet) actually

emerged as the best predictor according to the significance tests reported in Table S5.2. The only exception was the loyalty game. The friendliness facet did not predict loyalty better than the assertiveness facet; $p = 0.930$.

Discussion

Studies 5a–5d aimed to validate the facet model by showing that cultural evaluations of real and nuanced individuals on each facet predicted a relevant and specific behavior towards these unknown (vs. familiar in Studies 1–2a) and encountered (vs. envisioned in Studies 1a–4) individuals better than cultural evaluations of the individuals on all three other facets. The only exception was the friendliness facet in the loyalty game, which predicted loyalty but did not predict loyalty better than the assertiveness facet.

General discussion

The present research validated a recently endorsed model of social-evaluative dimensions that distinguishes the two horizontal facets friendliness and morality from the two vertical facets ability and assertiveness (Abele et al., 2021; Koch et al., 2021). In three studies and two supplemental studies, perceivers evaluated different types and numbers of targets in different modes (i.e., separate or joint evaluation) and ways (i.e., personal or cultural impressions). Across these five studies, an efficient two-items-per-facet measure fit the data well enough and better than a four-items-per-dimension measure (i.e., Big Two model) and five-items-per-facet measure (see Supplemental Studies 1 and 2). In addition, the efficient measure confirmed the hypothesized pattern of statistical relations between the facets, namely larger correlations between the friendliness and morality facets, and between the ability and assertiveness facets, compared to the correlations between any one horizontal facet and any one vertical facet. Apart from corroborating the external, internal, convergent, and discriminant validity of the facet model in these ways, the present research corroborated its predictive validity across two additional studies. Evaluations on each facet predicted a broadly relevant behavior intention towards the evaluated social entities (i.e., making one's bonus dependent on their loyalty, fairness, success, and hubris) better than evaluations on all other facets in at least one of the two studies.

The present findings advance the validation of the facet model in several important ways. We validated the facet model for the within-perceiver perspective by factor-analyzing differences between targets within the perceivers. Other research (Abele et al., 2016; Abele & Hauke, 2020; Barbedor et al., 2024) validated the facet model for the within-target perspective by factor-analyzing differences between perceivers within the targets. Validating the facet model for the within-perceiver perspective required an efficient two-items-per-facet measure (vs. the exhaustive five-items-per-facet measure in the previous work) that is more feasible in studies in which perceivers evaluate several or even many social entities on the facets. The tabular analyses in the online supplement (Tables S1.5–7, Table SS2.1, and see Supplemental Studies 1 and 2 as well) show that the efficient measure validated the facet model for the within-target perspective as well.

In addition, we validated the two-items-per-facet measure across envisioned and encountered (i.e., actually seen) individuals and groups as the targets of evaluation, and across several modes and types of social evaluation. These modes include separate, joint, personal, and cultural evaluation, which covers not just various real-life situations but also the standard paradigms in various research programs that examine social evaluation (Abele & Wojciszke, 2014; Ellemers, 2017; Fiske, 2018; Koch et al., 2016; Yzerbyt, 2018). We note that the efficient two-items-per-facet measure described cultural evaluations better than personal evaluations, and separate evaluations slightly better than joint evaluations, according to the fit indices that we considered in Studies 1 and 2.

Limitations and future research

First, we validated the facet model for evaluations of US targets by US perceivers. Other research validated the facet model for European contexts as well as China and Australia (Abele et al., 2016; Barbedor et al., 2024). Future research should generalize the two-items-per-facet measure to other national contexts, especially those that are not WEIRD (White, educated, industrialized, rich, and democratic; Muthukrishna et al., 2020).

Second, in Supplemental Studies 1 and 2, we reduced an initial list of five items per facet to two items per facet to achieve a satisfactory level of internal, convergent, and discriminant validity. Future research may examine whether this gain incurred a loss in content validity (i.e., covering every nuance of the content of each facet).

Third, we validated the two-items-per-facet measure through multilevel confirmatory factor analyses that treated the evaluators as data clusters. Thus, we can infer from our results that the two-items-per-facet measure was an adequate description of how an evaluator differentiated the targets, on average. We did not run multigroup confirmatory factor analysis that treated the evaluators as groups, not least because the ratio of targets to items was (very) low in our studies (e.g., four individuals versus eight items in Study 2).

Thus, we cannot infer from our results that the two-items-per-facet measure was an adequate description of how each evaluator differentiated the targets. In fact, previous research showed that sign and size of the perceived correlations between various social-evaluative dimensions vary across evaluators (Hehman et al., 2019; Stolier et al., 2018, 2020). Accordingly, it is unlikely that our two-items-per-facet measure is an adequate description of how each evaluator differentiates the targets that we examined.

Fourth, we predicted personal behaviors towards the targets from cultural evaluations rather than personal evaluations by those that showed the behaviors, which would have been less bold but more straightforward.

Finally, we predicted hypothetical behaviors and not actually incentivized behaviors towards the targets. Future research may address the above shortcomings.

Conclusion

The present endeavor secured considerable progress with validating a recent model of four social-evaluative facets (friendliness, morality, ability, and assertiveness) endorsed by an ongoing adversarial collaboration (Ellemers et al., 2020). Our validation work is sufficient to comfortably begin treating the two-items-per-facet measure that we contribute as a useful tool to capture social evaluation. We hope to facilitate more research that decomposes the Big Two into the four basic facets of social evaluation (e.g., studies that aim to resolve the controversies between different research programs; Abele et al., 2021; Koch et al., 2021).

Appendix 1

The items in the two-items-per-facet measure of social evaluation that we validated

Dimension	Facet	Item
Horizontal	Morality	"Honest"
		"Sincere"
	Friendliness	"Warm"
		"Friendly"
Vertical	Ability	"Skilled"
		"Capable"
	Assertiveness	"Confident"
		"Determined"

Appendix 2

The questions in the two-items-per-facet measure of social evaluation that we validated

	Personal mode	Cultural mode
Joint mode	"To what extent do you think of the following [groups or individuals displayed one below another] as [item]"	"To what extent do most Americans think of the following [groups or individuals displayed one below another] as [item]"
Separate mode	"To what extent do you think of [group or individual] as [items displayed one below another]"	"To what extent do most Americans think of [group or individual] as [items displayed one below another]"

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-024-02489-y.

Authors' note Authors' note: We preregistered Studies 4 (https://aspredicted.org/GW8_XKT) and 5a–5d (https://aspredicted.org/K31_VC5, https://aspredicted.org/QDN_M9Y, https://aspredicted.org/YH3_YNV, and https://aspredicted.org/T7T_Z5S, respectively). In each study, we collected, respectively). In each study, we collected all data before analyzing any of them. All study materials and data, code, and results are available on the Open Science Framework website (https://osf.io/sdvwt/).

Authors' contributions Everyone except Austin Smith ideated the research project and designed Studies 1a, 1b, 2, and the two supplemental studies. Alex Koch, Austin Smith, and Susan Fiske designed Studies 3–5. Alex Koch and Austin Smith collected and analyzed the data, and wrote the manuscript. Everyone else revised the manuscript.

Funding Templeton World Charity Foundation, TWCF-2022–30553 ("Next-generation Adversarial Collaboration") awarded to Andrea Abele, Naomi Ellemers, Susan Fiske, Alex Koch, and Vincent Yzerbyt.

Data availability Yes.

Declarations

Ethics approval Given by the University of Chicago, IRB22-1858.

Consent to participate Requested and given in all reported studies.

Consent for publication Not applicable.

Conflicts of interest None.

References

Abele, A. E. (2022). Evaluation of the self on the big two and their facets: Exploring the model and its nomological network. *International Review of Social Psychology, 35*(1), 1–15. https://doi.org/10.5334/irsp.688

Abele, A. E., Cuddy, A. J. C., Judd, C. M., & Yzerbyt, V. Y. (2008). Fundamental dimensions of social judgment: A view from different perspectives. *European Journal of Social Psychology, 38*, 1063–1065. https://doi.org/10.1037/0022-3514.89.6.899

Abele, A. E., Ellemers, N., Fiske, S. T., Koch, A., & Yzerbyt, V. (2021). Navigating the social world: Toward an integrated framework for evaluating self, individuals, and groups. *Psychological Review, 128*(2), 290–314. https://doi.org/10.1037/rev0000262

Abele, A. E., & Hauke, N. (2020). Comparing the facets of the big two in global evaluation of self versus other people. *European Journal of Social Psychology, 50*(5), 969–982. https://doi.org/10.1002/ejsp.2639

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Frontiers in Psychology, 7*, 1810. https://doi.org/10.3389/fpsyg.2016.01810

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. In J. M. Olson & M. P. Zanna (Eds.), *Advances in Experimental Social Psychology* (Vol. 50, pp. 195–255). Cambridge, MA: Academic Press. https://doi.org/10.1016/B978-0-12-800284-1.00004-7

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8*(2), 129–148. https://doi.org/10.1037/1082-989X.8.2.129

Barbedor, J., Schneider, J., Yzerbyt, V., & Abele, A. (2024). A novel approach to the evaluation of groups: Type of group and facet of evaluation matter. *Manuscript under review.*

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bazerman, M. H., Moore, D. A., Tenbrunsel, A. E., Wade-Benzoni, K. A., & Blount, S. (1999). Explaining how preferences change across joint versus separate evaluation. *Journal of Economic Behavior & Organization, 39*(1), 41–58. https://doi.org/10.1016/S0167-2681(99)00025-6

Brambilla, M., Sacchi, S., Rusconi, P., Cherubini, P., & Yzerbyt, V. Y. (2012). You want to give a good impression? Be honest! Moral traits dominate group impression formation. *British Journal of Social Psychology, 51*(1), 149–166. https://doi.org/10.1111/j.2044-8309.2010.02011.x

Carrier, A., Louvet, E., Chauvin, B., & Rohmer, O. (2014). The primacy of agency over competence in status perception. *Social Psychology, 45*(5), 1–10. https://doi.org/10.1027/1864-9335/a000176

Connor, P., Antonoplis, S., Nicolas, G., & Koch, A. (2024). Unconstrained descriptions of Facebook profile pictures support high-dimensional models of impression formation. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/01461672241266651

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE, 18*(3), e0279720. https://doi.org/10.1371/journal.pone.0279720

Ellemers, N. (2017). *Morality and the regulation of social behavior: Groups as moral anchors*. Routledge.

Ellemers, N., Fiske, S. T., Abele, A. E., Koch, A., & Yzerbyt, V. (2020). Adversarial alignment enables competing models to engage in cooperative theory building toward cumulative science. *Proceedings of the National Academy of Sciences, 117*(14), 7561–7567. https://doi.org/10.1073/pnas.1906720117

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2), 67–73. https://doi.org/10.1177/0963721417738825

Fiske, S. T. (2002). What we know now about bias and intergroup conflict, the problem of the century. *Current Directions in Psychological Science, 11*(4), 123–128. https://doi.org/10.1111/1467-8721.00183

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression* (3rd ed.). Sage publications.

Gallardo, R., Smith, A., Zak, U., Lopez, D., Kirgios, E., & Koch, A. (2024). Being in the minority boosts in-group love: Explanations and boundary conditions. *Manuscript under review*.

Gligorić, V., van Kleef, G. A., & Rutjens, B. T. (2022). Social evaluations of scientific occupations. *Scientific Reports, 12*(1), 18339. https://doi.org/10.1038/s41598-022-23197-7

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review, 95*(1), 384–394. https://doi.org/10.1257/0002828053828662

Hehman, E., Stolier, R. M., Freeman, J. B., Flake, J. K., & Xie, S. Y. (2019). Toward a comprehensive model of face impressions: What we know, what we do not, and paths forward. *Social and Personality Psychology Compass, 13*(2), e12431. https://doi.org/10.1111/spc3.12431

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576–590. https://doi.org/10.1037/0033-2909.125.5.576

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Imhoff, R., & Koch, A. (2017). How orthogonal are the Big Two of social perception? On the curvilinear relation between agency and communion. *Perspectives on Psychological Science, 12*(1), 122–137. https://doi.org/10.1177/1745691616657334

Imhoff, R., Koch, A., & Flade, F. (2018). (Pre)occupations: A data-driven model of jobs and its consequences for categorization and evaluation. *Journal of Experimental Social Psychology, 77*, 76–88. https://doi.org/10.1016/j.jesp.2018.04.001

Judd, C. M., Garcia-Marques, T., & Yzerbyt, V. (2019). The complexity of relations between dimensions of social perception: Decomposing bivariate associations with crossed random factors. *Journal of Experimental Social Psychology, 82*, 200–207. https://doi.org/10.1016/j.jesp.2019.01.008

Kline, R. B. (2005). *Principles and practice of structural equation modeling*. (2nd ed.). Guilford.

Koch, A., Dorrough, A., Glöckner, A., & Imhoff, R. (2020a). The ABC of society: Perceived similarity in agency/socioeconomic success and conservative-progressive beliefs increases intergroup cooperation. *Journal of Experimental Social Psychology, 90*, 103996. https://doi.org/10.1016/j.jesp.2020.103996

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110*(5), 675–709. https://doi.org/10.1037/pspa0000046

Koch, A., Imhoff, R., Unkelbach, C., Nicolas, G., Fiske, S., Terache, J., Carrier, A., & Yzerbyt, V. (2020b). Groups' warmth is a personal matter: Understanding consensus on stereotype dimensions reconciles adversarial models of social evaluation. *Journal of Experimental Social Psychology, 89*, 103995. https://doi.org/10.1016/j.jesp.2020.103995

Koch, A., Yzerbyt, V., Abele, A., Ellemers, N., & Fiske, S. T. (2021). Social evaluation: Comparing models across interpersonal, intragroup, intergroup, several-group, and many-group contexts. In B. Gawronski (Ed.), *Advances in Experimental Social Psychology* (Vol. 63, pp. 1–68). Cambridge, MA: Academic Press. https://doi.org/10.1016/bs.aesp.2020.11.001

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*(2), 234–249. https://doi.org/10.1037/0022-3514.93.2.234

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., Strong, D. R., & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*(2), 75–84. https://doi.org/10.1037/1076-898X.8.2.75

Lickel, B., Hamilton, D. L., Wieczorkowska, G., Lewis, A., Sherman, S. J., & Uhles, A. N. (2000). Varieties of groups and the perception of group entitativity. *Journal of Personality and Social Psychology, 78*(2), 223–246. https://doi.org/10.1037/0022-3514.78.2.223

Luchman, M. J. (2023). Package 'domir'. Available from https://bioconductor.statistik.tudortmund.de/cran/web/packages/domir/domir.pdf

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond western, educated, industrial, rich, and democratic (WEIRD) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychological Science, 31*(6), 678–701. https://doi.org/10.1177/0956797620916782

Nicolas, G., Bai, X., & Fiske, S. T. (2022). A spontaneous stereotype content model: Taxonomy, properties, and prediction. *Journal of Personality and Social Psychology, 123*(6), 1243–1263. https://doi.org/10.1037/pspa0000312

Oliveira, M., Garcia-Marques, T., Garcia-Marques, L., & Dotsch, R. (2020). Good to Bad or Bad to Bad? What is the relationship between valence and the trait content of the Big Two? *European Journal of Social Psychology, 50*(2), 463–483. https://doi.org/10.1002/ejsp.2618

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research, 49*(6), 518–543. https://doi.org/10.1080/00273171.2014.933762

Rau, R., Carlson, E. N., Back, M. D., Barranti, M., Gebauer, J. E., Human, L. J., Leising, D., & Nestler, S. (2021). What is the structure of perceiver effects? On the importance of global positivity and trait-specificity across personality domains and judgment contexts. *Journal of Personality and Social Psychology, 120*(3), 745–764. https://doi.org/10.1037/pspp0000278

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Stolier, R. M., Hehman, E., & Freeman, J. B. (2018). A dynamic structure of social trait space. *Trends in Cognitive Sciences, 22*(3), 197–200. https://doi.org/10.1016/j.tics.2017.12.003

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

Thielmann, I., Böhm, R., Ott, M., & Hilbig, B. (2021). Economic Games: An Introduction and Guide for Research. *Collabra: Psychology*, *7*(1), 19004. https://doi.org/10.1525/collabra.19004

Walsh, J., Vaida, N., & Fiske, S. (2023). Stereotype content predicts economic discrimination even under incentives. *Manuscript under review.*

Yzerbyt, V. Y. (2018). The dimensional compensation model: Reality and strategic constraints on Warmth and Competence in intergroup perceptions. In A. E. Abele & B. Wojciszke (Eds.), *The agency-communion framework* (pp. 126–141). Routledge.