Supplementary Materials

Sample R code to run and derive ICCs from cross-classified multilevel models

The below can be fully copied and pasted into R. Comments are denoted with a '#' at the beginning of the line, and anything without a # should be run as code.

```
# Downloading and installing lme4, an R package for multilevel modeling

install.packages("lme4")
library(lme4)

# Building the cross-classified model: First we present a basic model in which
# m = the created model, dv = dependent variable, c1 = cluster 1, c2 = cluster 2, and
# dataset = name of dataset
# Thus, dv is cross-classified by both c1 and c2

m <- lmer(dv ~ 1 + (1 | c1) + (1 | c2), data = dataset)

# To demonstrate, here is an example model in the context of the current paper:
# dv is ratings on trustworthiness
# c1 is participant ID (the rater)
# c2 is stimulus ID (the target)
# Thus, ratings of trustworthiness are cross-classified by both rater and target

m <- lmer(trustworthiness ~ 1 + (1 | rater) + (1 | target), data = dataset)

# Analyses for calculating ICCs
# The below command will return a section labeled 'random effects'.

summary(m)

# For example:

# Random effects:
#  Groups        Name        Variance Std.Dev.
#  Rater         (Intercept) 0.4456   0.6675
#  Target        (Intercept) 0.3034   0.5508
#  Residual                  1.1926   1.0921
# Number of obs: 17356, groups:  ParticipantID, 459; StimName, 413

# Of the three bolded, underlined numbers, Rater Variance is the tau(τ) for rater,
Target Variance is the tau(τ) for target, and Residual Variance is sigma squared(σ2)

# To calculate rater ICC, use this equation:

      # raterICC = rater tau / (sigma squared + rater tau + target tau)

# To calculate target ICC:

      # targetICC = target tau / (sigma squared + rater tau + target tau)

# The resulting numbers represent the percentage of variance from between-rater
(rater-ICC) or between-target (target-ICC) respectively.
```

Method for establishing a corridor of stability

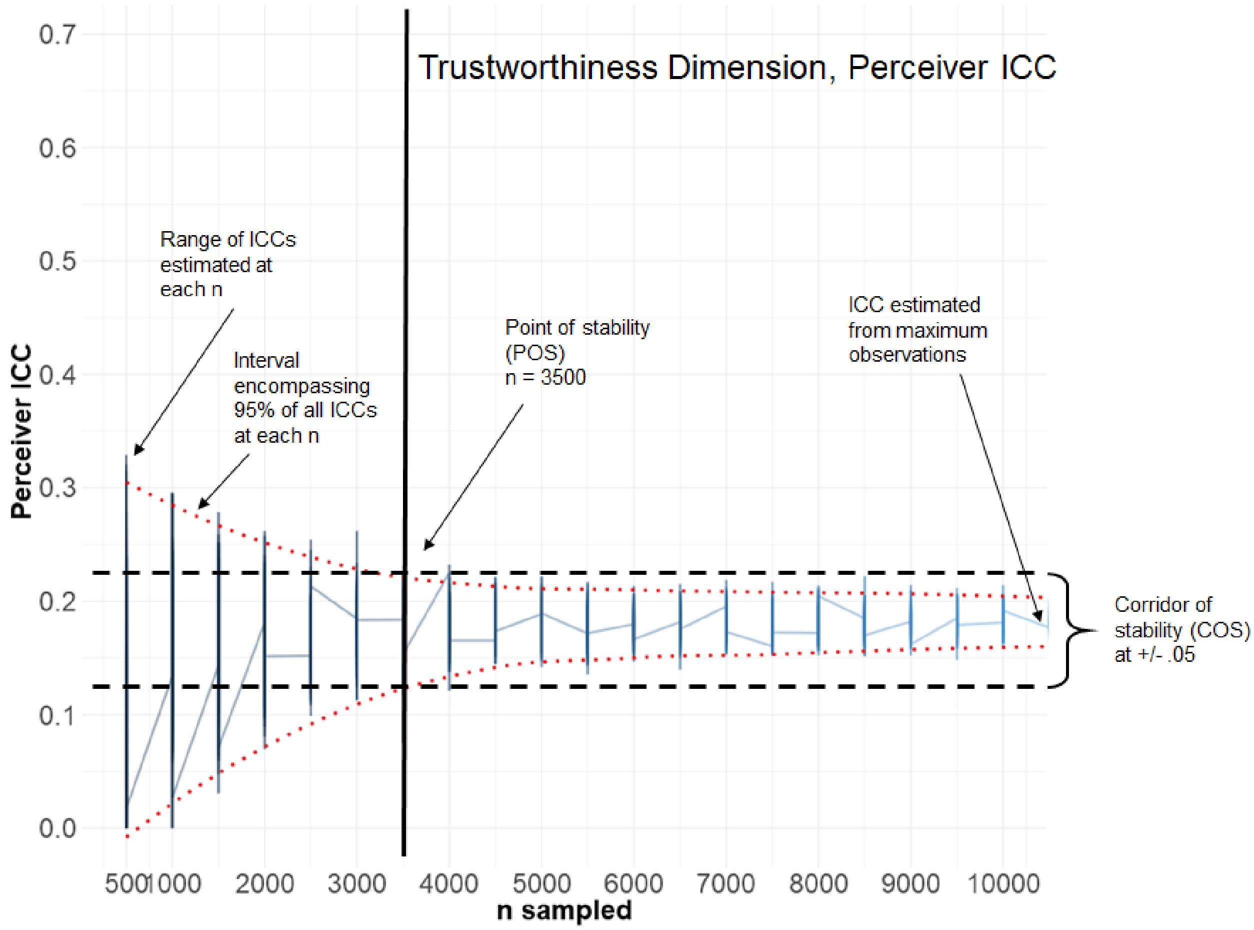
We implemented a sequential sampling approach adapted from Schönbrodt and Perugini 2013 to assess how many observations might be necessary for stable ICCs, which ensures we have a large enough sample to estimate precise ICCs. We use "stability" to reflect the N at which ICC estimates do not meaningfully change with the incorporation of additional observations – in other words, when we have sufficient sample size to limit aberrant ICCs that are the result of sampling variability.

In this approach, we defined an acceptable "corridor of stability" (COS) within which sampling variations are acceptable. Defining the COS is an arbitrary decision made by any researcher, but given that ICCs are correlations, we use the boundaries established by Schönbrodt and Perugini (+/- r = .1), and a second even more conservative corridor of +/- r = .05.

Specifically, we randomly sampled (with replacement) observations from our data, and calculated ICCs from those observations. We did so with an increasing *n* of observations, 500 times at each *n*. The "point of stability" (POS) was defined as the *n* at which 95% of the ICCs fell within the COS and did not again exceed the boundaries of the COS, consistent with previous research (Schönbrodt and Perugini, 2013). We do so for the three dimensions reported in the manuscript. Results are below, for both +/- .1 and the more conservative +/- .05 corridor.

Dimension	Corridor (+/-)	Perceiver ICC POS	Target ICC POS
Trustworthiness	0.1	2000	1500
Trustworthiness	0.05	3500	3000
Dominance	0.1	2000	1500
Dominance	0.05	6000	3500
Youthful/Attractive	0.1	1500	1500
Youthful/Attractive	0.05	3500	3000

Here is a visualization of the sampling approach:



As the figure and table depict, all ICCs are stable by 6,000 observations at latest, even using a conservative +/- .05 corridor of stability. In the current research, ICC estimates come from a number of observations that exceed these limits. This is with the exception of when minority perceivers observe other minority targets. Accordingly, these estimates are not present in the main text.

Correlation Matrix: Exploratory Dataset

Correlations between trait ratings (averaged across all participants for each stimulus) used to form the dimensions (trustworthiness, dominance, youthful/attractiveness).
Thus, *N* reflects the # of stimuli for which the various traits were each collected.
The large degree of variability in *N* is due to different trait ratings and different stimuli being included across different studies.

Supplementary Table 1

Means, standard deviations, and correlations with confidence intervals for ratings from the exploratory dataset

Trait	*M*	*SD*	aggressive	assertive	attractive	caring	competent	dominant	friendly
aggressive	3.67	0.57							
assertive	3.96	0.59	.51** [.37, .63]						
attractive	3.57	0.72	-.38** [-.56, -.15]	.41** [.20, .59]					
caring	3.75	0.63	-.78** [-.85, -.70]	-.07 [-.25, .11]	.62** [.46, .75]				
competent	4.29	0.54	-.45 [-.78, .08]	.59* [.11, .85]	.66** [.60, .72]	.84** [.58, .95]			
dominant	4.16	0.77	.63** [.51, .73]	.81** [.73, .86]	-.04 [-.17, .09]	-.31** [-.46, -.14]	.18 [-.03, .38]		
friendly	3.74	0.67	-.83** [-.88, -.77]	-.18 [-.34, .00]	.47** [.36, .57]	.91** [.87, .93]	.75** [.64, .83]	-.42** [-.51, -.32]	
healthy	4.51	0.70	-.40** [-.54, -.24]	.35** [.18, .50]	.71** [.57, .81]	.56** [.43, .67]	.83** [.56, .94]	.04 [-.14, .22]	.55** [.41, .66]
intelligent	4.20	0.60	-.55**	-.00	.37**	.62**	.73**	-.32**	.58**

			[-.66, -.41]	[-.18, .18]	[.29, .45]	[.50, .72]	[.67, .78]	[-.47, -.15]	[.44, .68]
likable	3.84	0.59	-.60*	.42	.81**	.87**	.87**	-.16	.87**
			[-.85, -.13]	[-.12, .76]	[.70, .87]	[.66, .96]	[.81, .91]	[-.36, .06]	[.80, .91]
smart	3.97	0.57	-.58**	-.04	.61**	.61**	.70**	-.29**	.60**
			[-.68, -.44]	[-.21, .14]	[.44, .74]	[.48, .71]	[.29, .89]	[-.45, -.12]	[.48, .71]
strong	4.05	0.71	.32**	.71**	.13**	.04	.16**	.71**	-.07
			[.15, .47]	[.61, .79]	[.03, .23]	[-.14, .22]	[.07, .38]	[.64, .76]	[-.19, .05]
trustworthy	3.86	0.83	-.79**	-.18*	.66**	.84**	.79**	-.40**	.81**
			[-.85, -.71]	[-.35, -.00]	[.58, .73]	[.78, .88]	[.68, .86]	[-.50, -.30]	[.77, .85]
warm	4.22	0.72	-.79**	-.14	.49**	.91**	.44**	-.35**	.94**
			[-.85, -.71]	[-.31, .04]	[.41, .57]	[.88, .94]	[.34, .53]	[-.49, -.18]	[.91, .96]
youthful	4.02	1.06	-.24*	-.28**	.42**	.04	.00	-.38**	.15*
			[-.40, -.06]	[-.44, -.11]	[.33, .50]	[-.14, .22]	[-.23, .24]	[-.47, -.27]	[.03, .26]

Trait	*M*	*SD*	healthy	intelligent	likable	smart	strong	trustworthy	warm
intelligent	4.20	0.60	.60**						
			[.47, .70]						
likable	3.84	0.59	.75**	.91**					
			[.39, .91]	[.76, .97]					
smart	3.97	0.57	.60**	.82**	.74**				
			[.47, .70]	[.76, .87]	[.36, .91]				

strong	4.05	0.71	.25**	-.12	-.03	-.24**			
			[.07, .41]	[-.24, .00]	[-.26, .20]	[-.40, -.07]			
trustworthy	3.86	0.83	.57**	.74**	.86**	.66**	-.09		
			[.43, .68]	[.65, .81]	[.79, .91]	[.55, .75]	[-.21, .02]		
warm	4.22	0.72	.52**	.47**	.83**	.55**	.05	.78**	
			[.38, .64]	[.38, .54]	[.56, .94]	[.41, .66]	[-.13, .22]	[.69, .84]	
youthful	4.02	1.06	.33**	.13*	.25*	.24**	-.26**	.26**	.08
			[.16, .48]	[.01, .25]	[.02, .46]	[.07, .46]	[-.35, -.17]	[.15, .37]	[-.10, .25]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates p < .05. ** indicates p < .01.

Correlation Matrix: Confirmatory Dataset

Supplementary Table 2

Means, standard deviations, and correlations with confidence intervals for ratings from the confirmatory dataset

Trait	*M*	*SD*	aggressive	assertive	attractive	caring	competent	dominant	friendly
aggressive	3.91	0.80							
assertive	4.08	0.69	.59** [.55, .64]						
attractive	3.21	0.73	-.09* [-.18, -.00]	.27** [.18, .35]					
caring	3.81	0.72	-.65** [-.69, -.61]	-.26** [-.32, -.19]	.38** [.30, .45]				
competent	4.28	0.67	-.39** [-.45, -.33]	.05 [-.03, .12]	.44** [.36, .51]	.50** [.45, .55]			
dominant	4.04	0.88	.70** [.65, .73]	.80** [.77, .82]	.20** [.11, .28]	-.32** [-.39, -.25]	-.11** [-.18, -.04]		
friendly	3.89	0.80	-.65** [-.69, -.60]	-.26** [-.33, -.19]	.30** [.22, .38]	.80** [.77, .82]	.42** [.35, .48]	-.32** [-.39, -.25]	
healthy	4.32	0.81	-.33** [-.39, -.27]	.11** [.04, .18]	.64** [.58, .69]	.47** [.42, .53]	.61** [.56, .65]	-.01 [-.08, .07]	.44** [.37, .50]

intelligent	4.07	0.71	-.53** [-.58, -.48]	-.13** [-.20, -.06]	.39** [.32, .47]	.52** [.46, .57]	.66** [.62, .70]	-.32** [-.38, -.25]	.46** [.40, .52]
smart	4.16	0.66	-.51** [-.56, -.46]	-.10** [-.17, -.03]	.46** [.39, .53]	.51** [.46, .56]	.66** [.62, .70]	-.32** [-.38, -.25]	.46** [.39, .51]
strong	4.08	0.90	.49** [.43, .55]	.69** [.65, .73]	.25** [.17, .34]	-.06 [-.13, .02]	-.06 [-.14, .01]	.77** [.74, .80]	-.05 [-.13, .02]
trustworthy	3.99	0.59	-.68** [-.72, -.63]	-.33** [-.41, -.25]	.45** [.38, .52]	.71** [.67, .76]	.56** [.49, .61]	-.42** [-.49, -.35]	.67** [.62, .72]
warm	3.75	0.82	-.64** [-.68, -.59]	-.23** [-.30, -.16]	.37** [.29, .44]	.81** [.78, .83]	.45** [.40, .51]	-.29** [-.36, -.22]	.82** [.79, .84]
youthful	4.05	1.00	-.44** [-.50, -.38]	-.38** [-.45, -.32]	.34** [.26, .42]	.23** [.16, .30]	.25** [.18, .32]	-.45** [-.51, -.39]	.28** [.21, .35]

Variable	*M*	*SD*	healthy	intelligent	smart	strong	trustworthy	warm
intelligent	4.07	0.71	.63** [.58, .67]					
smart	4.16	0.66	.62** [.58, .66]	.81** [.78, .83]				
strong	4.08	0.90	.10* [.02, .17]	-.32** [-.38, -.25]	-.28** [-.35, -.21]			
trustworthy	3.99	0.59	.55** [.49, .61]	.70** [.65, .74]	.69** [.64, .73]	-.25** [-.33, -.17]		

warm	3.75	0.82	.50** [.45, .55]	.51** [.46, .56]	.51** [.46, .56]	-.02 [-.09, .06]	.74** [.70, .78]	
youthful	4.05	1.00	.43** [.37, .49]	.41** [.35, .47]	.45** [.39, .51]	-.43** [-.49, -.37]	.39** [.31, .47]	.28** [.21, .35]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Method for comparing ICCs across multiple categories

Although our aims were primarily descriptive in nature, for selected comparisons we tested whether ICCs for one group differed from one another by bootstrapping 95% confidence intervals (CIs) around their ICCs and examining the overlap in intervals. These intervals allow us to make quantitative comparisons between groups. While there were many potential comparisons to explore, we were interested in trends on a broader level. For instance, instead of comparing ratings of Asian female vs. Asian male targets specifically, we also wanted to compare ratings of female vs. male targets more generally. To compare CIs across these broader social categories (e.g., all ratings of males vs. all ratings of females), we averaged the ICC intervals of groups belonging to some relevant category (e.g., target gender) while collapsing across other categories (i.e., target race). This allowed us to compare ICCs for *all* ratings of male vs. female targets, and examine, for example, target gender independent of target race.

To clarify, we averaged across multiple bootstrapped ICC intervals (e.g., across Asian, Black, and White female targets) to create averaged CIs for *all* female targets. We used this approach instead of building a separate null model for female targets and bootstrapping new intervals around that model, because we have unequal numbers of observations in each category (e.g., more ratings of White female targets than Black or Asian). If we had built a separate null model for "all" female targets, our ICC estimates would be skewed by the imbalance in ratings, and would primarily represent intervals for White female targets relative to other-race female targets. To summarize, we created subsets of our data for comparisons across broader social categories, estimated null models from each subset, and compared ICCs across different models.

Method for developing maximal models for variance explained approach

For purposes of examining how much variance was explained by our different predictors, we returned to our complete, non-subsetted dataset. This analysis models all responses together as one large, conditional model with predictors, rather than separate null models. Instead of creating subsets for each social evaluation, we include all perceiver and target groups as moderators in the model. Specifically, we contrast-coded four predictors of interest: perceiver gender, perceiver racial majority/minority status, target gender, and target racial majority/minority status. We then built one maximally identified model with all moderators included as fixed and random effects to estimate the variance explained when *all* perceiver/target characteristics were in the model.

However, this maximal model consistently failed to converge. This is often the case when some of the variances being estimated by the model are near zero. Accordingly, we applied recommended remedies for convergence issues in multilevel modeling (for review, see Brauer & Curtin, 2017). Following these recommendations, we modified the maximal model until we produced one that converged consistently across different dimensions and both exploratory and confirmatory datasets, or based on Brauer & Curtin's (2017) recommendations, until we could be reasonably certain that the non-convergence warnings produced by the *lme4* package were false positives. We ran a modified maximal model for each dimension (ratings on each dimension served as the DV).

Accordingly, our analyses have four moderators – perceiver racial majority/minority status, target racial majority/minority status, perceiver gender, and target gender. The sample R code below shows the modified maximal model for ratings on the trustworthiness dimension.

```
TrustworthinessMaximal <- lmer(Rating ~
        PerceiverRacialStatus * TargetRacialStatus * PerceiverGender * TargetGender +
        (TargetRacialStatus + TargetGender | PerceiverID) +
```

```
(PerceiverRacialStatus + PerceiverGender | TargetID),
data = TrustworthinessDimension)
```

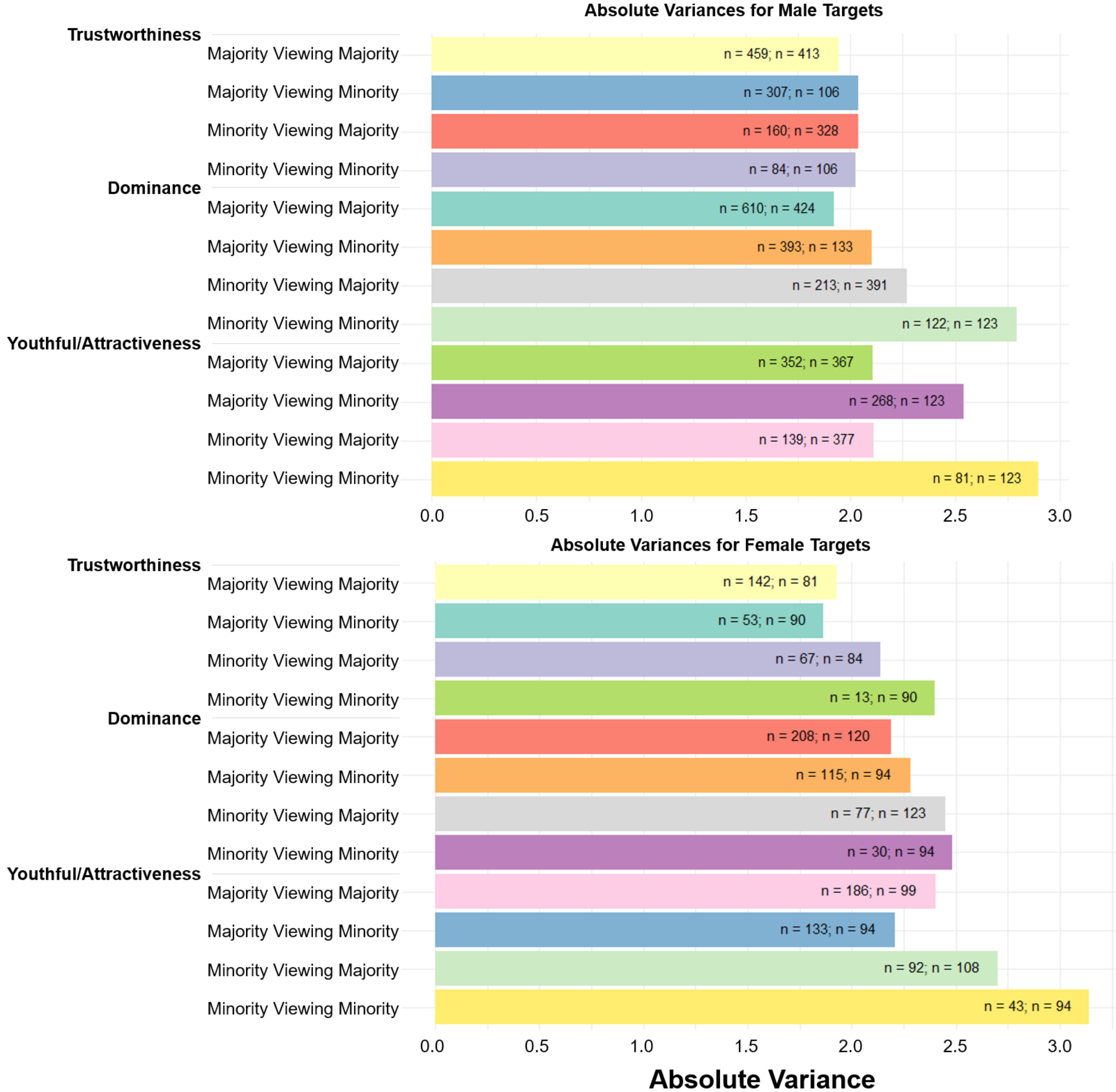
Here, our four moderators are included as fixed effects and allowed to fully interact (i.e., four-way interaction). Target racial majority/minority status and target gender are also allowed to randomly vary across perceivers (i.e., included as random effects at the perceiver level), and perceiver racial majority/minority status and perceiver gender are allowed to randomly vary across targets (i.e., included as random effects at the target level). In a maximal model, the random effects at each level would also be allowed to interact, but these models failed to converge. Thus, we included random slopes, but did not allow them to interact.

Supplementary Figure 1

Supplementary Figure 1. Example stimuli.

Absolute Variances: Exploratory Dataset

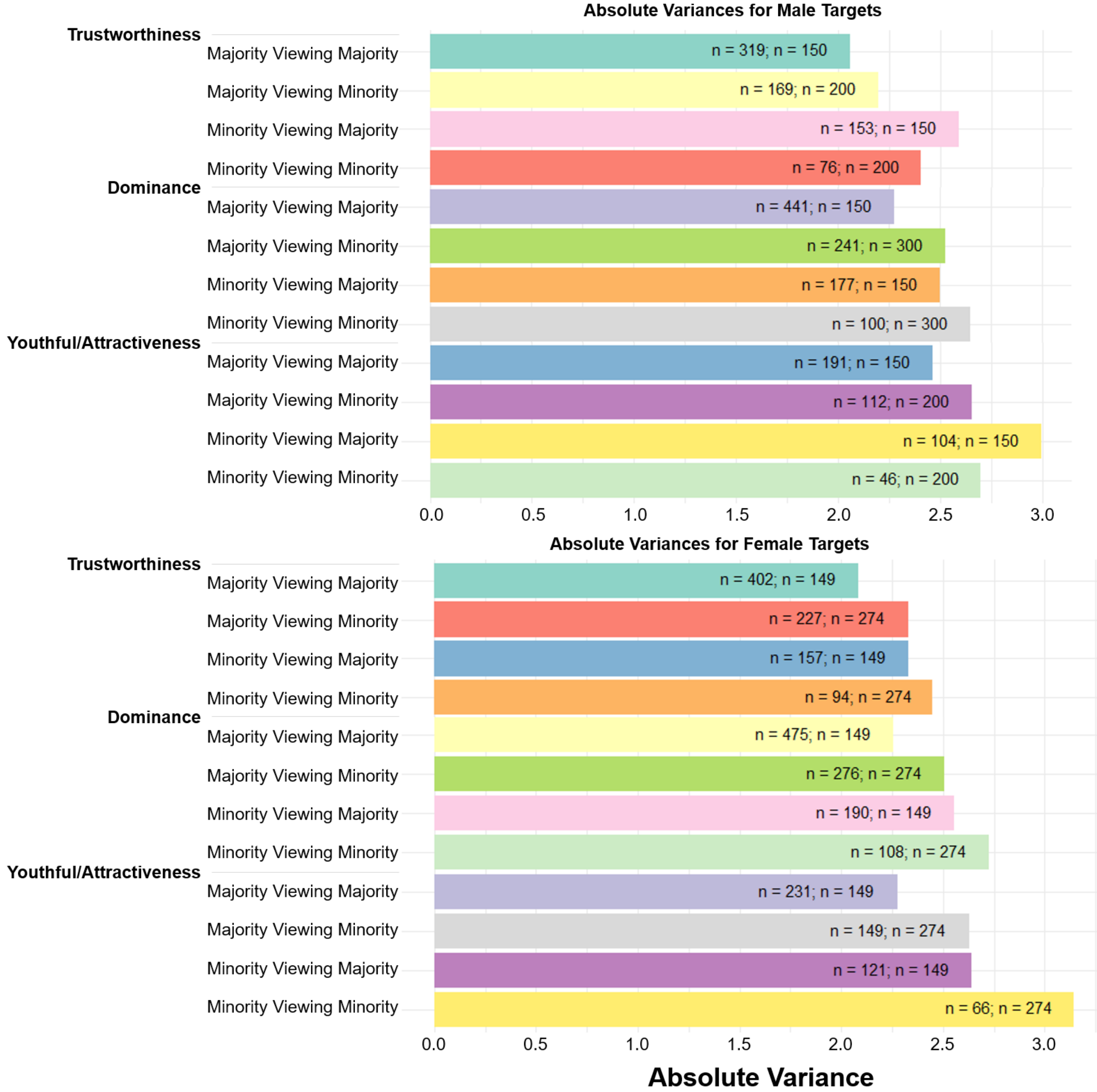
Supplementary Figure 2



Supplementary Figure 2. Absolute variance for models of impressions on trustworthiness, dominance, and youthful/attractiveness dimensions for male and female targets according to racial group status from the exploratory dataset. N for participants and N for stimuli displayed on each bar.

Absolute Variances: Confirmatory Dataset

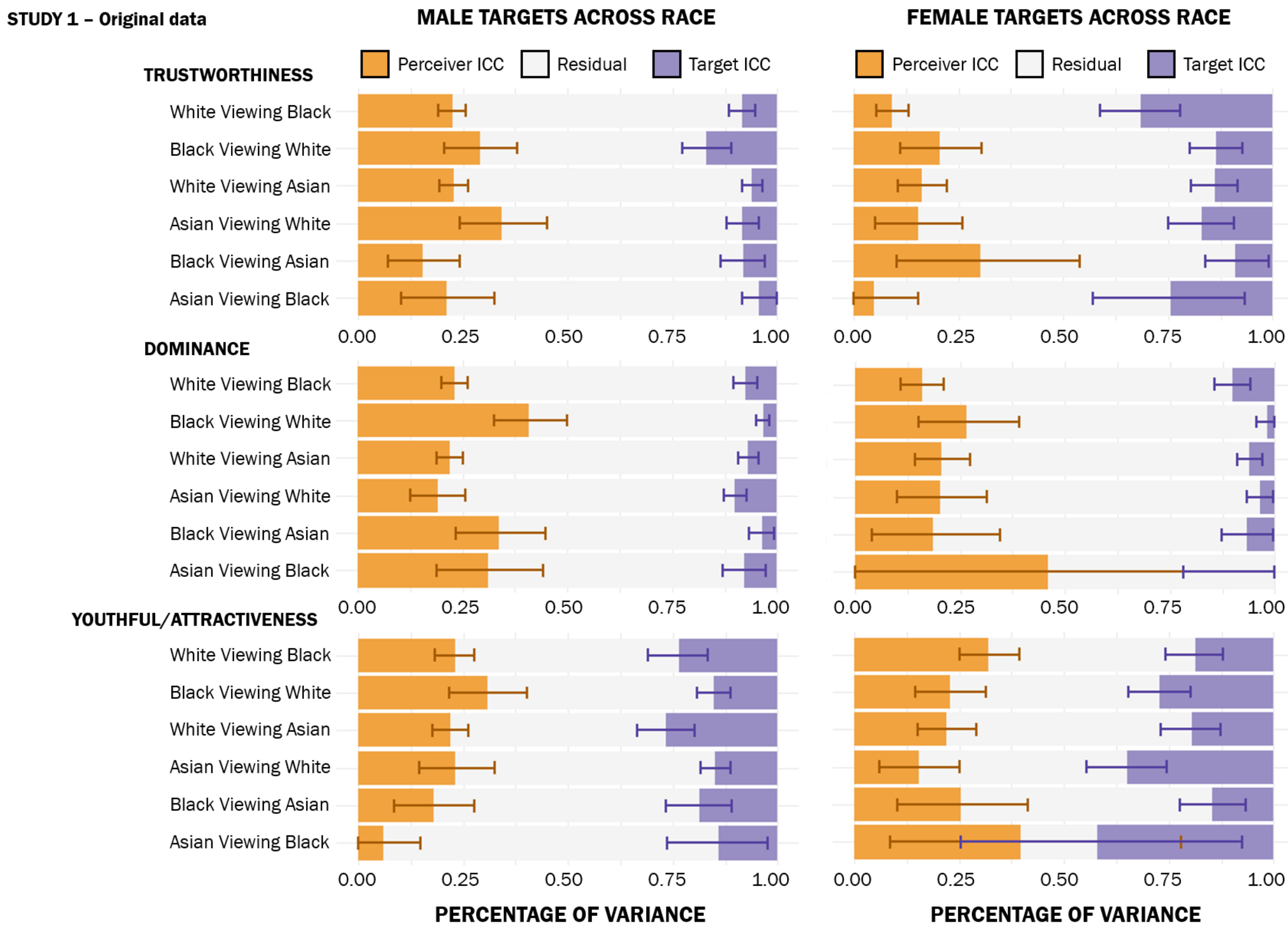
Supplementary Figure 3



Supplementary Figure 3. Absolute variance for models of impressions on trustworthiness, dominance, and youthful/attractiveness dimensions for male and female targets according to racial group status from the confirmatory dataset. N for participants and N for stimuli displayed on each bar.

Minorities viewing other minorities: Exploratory Dataset

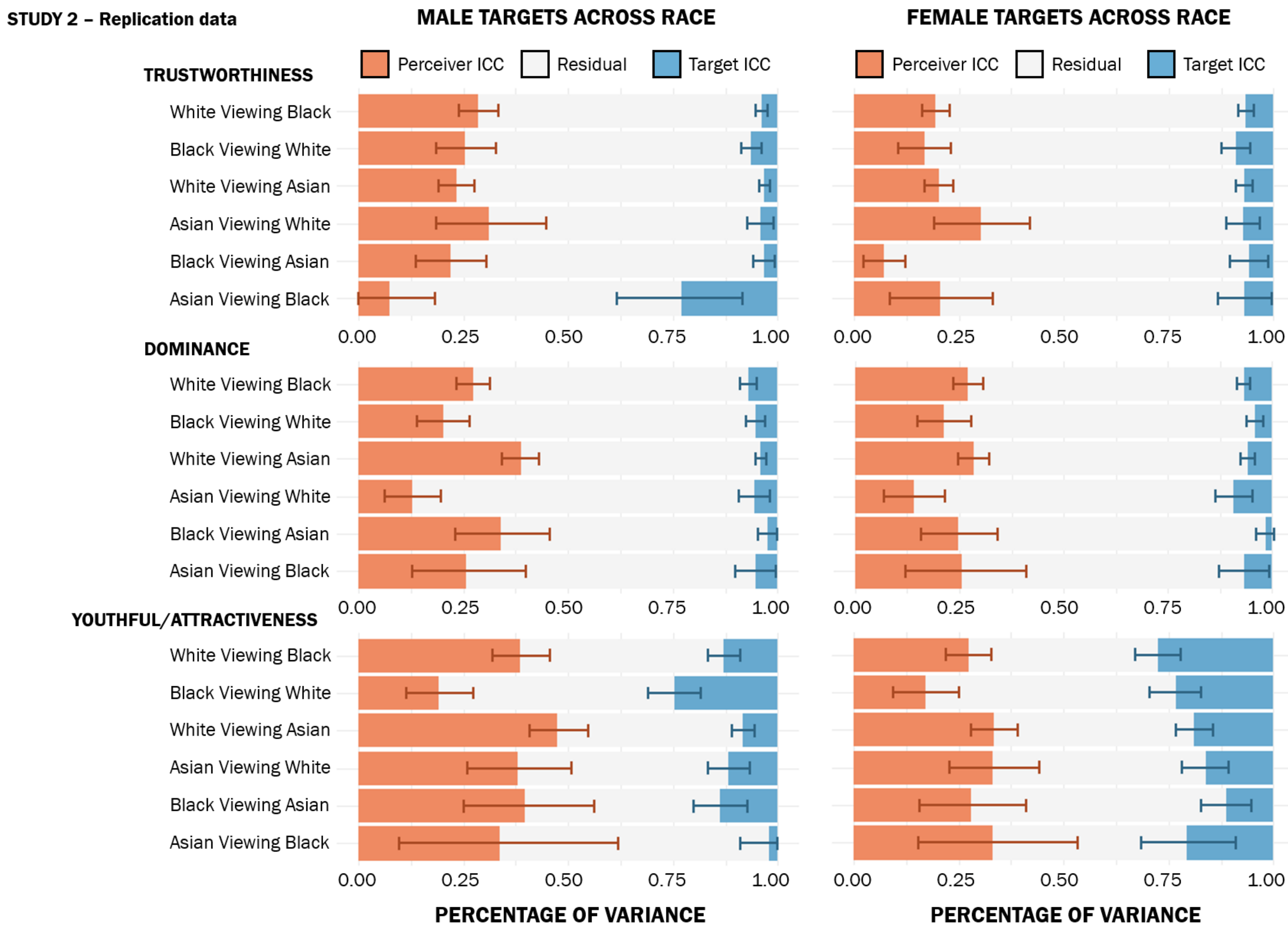
Supplementary Figure 4



Supplementary Figure 4. Relative contributions of perceiver-, target-, and residual variance to impressions for male and female targets according to racial group status. Results from the exploratory dataset. Error bars represent 95% confidence intervals.

Minorities viewing other minorities: Confirmatory Dataset

Supplementary Figure 5



Supplementary Figure 5. Relative contributions of perceiver-, target-, and residual variance to impressions for male and female targets according to racial group status. Results from the confirmatory dataset. Error bars represent 95% confidence intervals.

Table of ICC Estimates for Study 1: Social Perceptions Across Race and Gender for Black and Asian Targets

Supplementary Table 3

Intra-class correlations and 95% confidence intervals for Black and Asian perceivers' ratings of Black and Asian targets

	Study 1: Exploratory				Study 2: Confirmatory			
	Perceiver ICC	95% CI	Target ICC	95% CI	Perceiver ICC	95% CI	Target ICC	95% CI
Male Targets								
Trustworthiness								
Black viewing Asian		[.073, .243]		[.030, .134]		[.137, .306]		[.007, .058]
Asian viewing Black		[.103, .326]		[.002, .084]		[-.031, .184]		[.083, .383]
Dominance								
Black viewing Asian		[.233, .448]		[.007, .066]		[.232, .457]		[.000, .046]
Asian viewing Black		[.189, .444]		[.027, .130]		[.127, .400]		[.003, .102]
Youthful/Attractiveness								
Black viewing Asian		[.087, .277]		[.110, .265]		[.253, .563]		[.072, .200]
Asian viewing Black		[-.026, .150]		[.024, .262]		[.098, .620]		[-.074, .088]
Female Targets								
Trustworthiness								
Black viewing Asian		[.102, .539]		[.010, .162]		[.024, .123]		[.013, .104]
Asian viewing Black		[-.054, .154]		[.066, .429]		[.087, .333]		[.005, .134]
Dominance								
Black viewing Asian		[.039, .346]		[.004, .128]		[.158, .343]		[-.006, .038]
Asian viewing Black		[-.050, 1.120]		[-.406, .219]		[.121, .409]		[.007, .129]
Youthful/Attractiveness								
Black viewing Asian		[.104, .415]		[.068, .223]		[.158, .411]		[.052, .172]
Asian viewing Black		[.087, .780]		[.075, .746]		[.153, .535]		[.090, .316]

Table of ICC Estimates for Study 2: Minimal Groups

Supplementary Table 4

Intra-class correlations and 95% confidence intervals for minimal own- and other-group

	Minimal Groups			
	Perceiver ICC	95% CI	Target ICC	95% CI
Trustworthiness				
Own-Group	.251	[.175, .332]	.172	[.119, .226]
Other-Group	.200	[.134, .270]	.198	[.139, .256]
Dominance				
Own-Group	.179	[.120, .238]	.116	[.076, .156]
Other-Group	.188	[.127, .249]	.090	[.057, .124]
Youthful/Attractiveness				
Own-Group	.310	[.200, .428]	.223	[.151, .295]
Other-Group	.307	[.196, .423]	.195	[.129, .260]