Bayesian Analysis for the Gender, Status, and Emotions Project - Study 2

Suzanne Hoogeveen[1] & Julia M. Haaf[1]

[1] University of Amsterdam

Version 3, 01/2022

Bayesian Analysis for the Gender, Status, and Emotions Project - Study 2

Methods

Participants

For the primary analyses, we included the full sample of 7552 participants who completed the relevant measures for the main analyses.

Material

The dependent variable 'status conferral' was created by averaging responses to the items on *power*, *independence*, and *status* (average inter-item correlation: 0.58). The dependent variable 'competence' was created by averaging responses to the items on *competence* and *knowledgeability* (inter-item correlation: 0.72). The dependent variable 'warmth' was created by averaging responses to the items on *likeability* and *warmth* (inter-item correlation: 0.71).

As preregistered, we assessed the reliablity of the individual differences scale by calculating Cronbach's alpha. All five scales were internally consistent and surpassed the threshold of alpha > 0.4; the alpha values were 0.82 (news exposure scale), 0.83 (internal motivation not to be sexist), 0.87 (beliefs about gender inequality in the workplace), 0.74 (sexist beliefs), and 0.85 (external motivation not to appear sexist). The variable *number of studies previously participated in* was heavily right-skewed and was therefore binned into 6 groups: "0," "1-2," "3-5," "6-10," "11-20," "20+."

Procedure

Data analysis

All analyses were conducted in R. We constructed hierarchical Bayesian regression models that reflect the predictions from the 5 substantive theories, as well as the null-model and an unconstrained model that includes all main parameters from the separate theories,

which are free to vary in size and direction. Note that as perspectives 3 and 4 make equal predictions regarding the overall pattern, there were 4 different theoretical models in total. In the primary analysis, we used different ordinal constraints to capture the different theoretical predictions (see Appendix for details). The relative predictive adequacy of these models as well as the unconstrained model was compared using Bayes factors, following the approach by Haaf and Rouder (2017), Rouder, Haaf, Davis-Stober, and Hilgard (2019) and Haaf, Klaassen, and Rouder (2018).

In addition, we assessed the robustness of the findings to somewhat arbitrary analysis decisions by conducting a multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016): We intended to apply different data exclusion criteria related to language experience, a manipulation check, the validity of responses (straightlining on the included scales), perceived material quality, mode of administration (laptop, phone etc.), and a design-level manipulation check. However, as described below, straightlining criteria did not affect a substantial proportion of the sample (less than 2%) and was thus omitted as a separate path in the multiverse analysis. The preregistration for the analysis can be found at https://osf.io/sh3vx/.

Prior settings. We think small effects in the predicted direction may still be meaningful, especially with regard to gender bias where small biases can accumulate in terms of their consequences over time. We therefore used a scale of 0.25 for the effect of interest. A scale of 0.25 assumes an size effect that is 25% of the sampling noise (standard deviation), which is generally considered a small effect. For the variation between labs in the intercepts, we used a scale of 1. In the random-effects models we used a scale of 0.15 for site-specific variation in the effects of interest.

Results

Manipulation check

As a manipulation check, we assessed whether the target in the *anger* condition was indeed perceived as more angry than the target in the *not-angry* (neutral or sadness) condition and vice versa for sadness. The independent samples Bayesian t-test gives infinite evidence in favor of the hypothesis that targets with anger expressions are perceived as more angry than targets with not-angry (sadness or neutral) expressions ($BF_{+0} = \infty$, $\delta = 1.03$). Similarly, the independent samples Bayesian t-test gives infinite evidence in favor of the hypothesis that targets with not-angry expressions are perceived as more sad than targets with angry expressions ($BF_{-0} = \infty$, $\delta = -0.31$). See Figure 11 for a plot of the data.

Primary Theoretical Tests

Status Conferral. Based on the Bayes factor model comparison, for status conferral we find most evidence in favor of the baseline model that assumes a varying intercept per site and a varying intercept per design, but no experimental effects (i.e., no main effect of target gender, target emotion or a gender-by-emotion interaction). Specifically, the baseline model outperforms the Gender Stereotyping model by a factor of 66.0, the Status Signalling perspective by a factor of 200, and the Culture Change/Study Savviness model (reversed gender stereotyping) by a factor of 707, which is considered strong evidence against the theoretical perspectives (Lee & Wagenmakers, 2013). Finally, the baseline model fits the data much better than the null model ($BF_{b0} = 8.1 \times 10^{139}$), indicating that the overall status ratings differ across designs. Figure 1A visualizes the absence of experimental effects.

Competence. For competence ratings, we find most evidence for the unconstrained model, which strongly outperforms the baseline model: $BF_{ub} = 8.9 \times 10^{15}$. The pattern of results shows that, in contrast to the Status Signalling and Gender Stereotyping perspectives, both angry men ($M_{angry\ male} = 6.74$) and women ($M_{angry\ female} = 6.82$) are considered *less* competent than neutral/sad men ($M_{not-angry\ male} = 7.14$) and women ($M_{not-angry\ female} = 7.31$; see also Figure 1B).

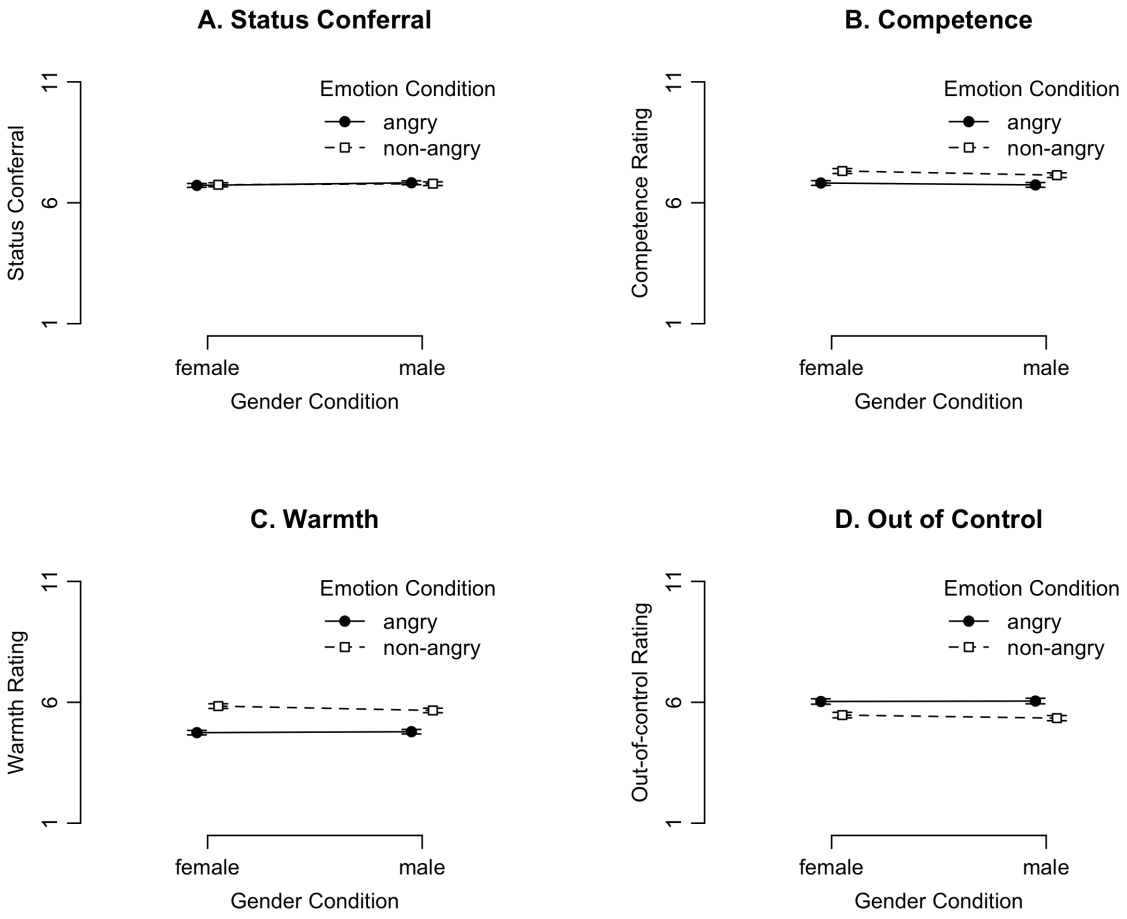


Figure 1. Descriptive plots of **A.** status conferral, **B.** competence, **C.** warmth, and **D.** Out of control per target gender and emotion.

Warmth. As expected, for perceived warmth, we find most evidence for the 'Anger suppresses warmth' perspective; angry targets ($M_{angry\ male} = 4.78$, $M_{angry\ female} = 4.74$) are perceived as less warm than sad/neutral targets ($M_{not-angry\ male} = 5.66$, $M_{not-angry\ female} = 5.84$) , regardless of gender: $\text{BF}_{wb} = 3.9 \times 10^{102}$ (see also Figure 1C).

Out of Control. For the out-of-control ratings, we expected most evidence for the Gender Stereotyping model such that anger increases out-of-control ratings for female targets, compared to non-angry targets and angry male targets. The results, however, showed most evidence for the Status Signalling perspective; $\text{BF}_{sb} = 1.1 \times 10^{27}$: both angry

Table 1
Evidence for the theoretical models vs. the baseline (intercepts only) models for the different dependent variables.

	Bayes Factors			
Comparison	Status	Competence	Warmth	Out of Control
Gender Stereotyping vs. Baseline	0.02	0.00	0.00	0.00
Satus Signalling vs. Baseline	0.00	0.00	0.00	$\mathbf{1.1 \times 10^{27}}$
Culture Change vs. Baseline	0.00	0.00	0.00	1.9×10^{10}
Unconstrained vs. Baseline	0.01	$\mathbf{8.9 \times 10^{15}}$	9.7×10^{101}	2.7×10^{26}
Null vs. Baseline	0.00	0.00	0.00	0.00
Warmth vs. Baseline	–	–	$\mathbf{3.9 \times 10^{102}}$	–

Note. The best-predicting model per dependent variable is indicated in bold. For status conferral, all of the theoretical models perform worse than the baseline model. Note that the baseline model includes random intercepts for labs and for designs, while the null model only includes random intercepts for labs.

men ($M_{angry\ male} = 6.05$) and women ($M_{angry\ female} = 6.03$) are considered more out of control than not-angry men ($M_{not-angry\ male} = 5.35$) and women ($M_{not-angry\ female} = 5.47$; see Figure 1D).

Primary Methodological Tests

Varying effects model. For the primary analyses, we constructed simple common effect models that assume that the effects of interest are of equal size across different designs. As preregistered, we also investigated whether the effects differ substantially per design, and if so what design-related features could explain the pattern. To this end, we built a varying effects model that allows the main effects of target gender and target emotion, as well as the interaction, to vary between designs. Note that we did not investigate random effects across sites, as study 1 indicated no evidence in favor of variability across labs.

As visualized in Figure 2, however, there is no indication that the target emotion effect and the gender-by-emotion interaction effect vary substantially of systematically across designs. Rather, the 95% credible intervals includes zero across almost all designs for the main effect of target emotion, as well as the interaction effect. For the target emotion effect,

angry targets are accorded more status than non-angry targets across 2 designs, while the opposite is true for 1 design. For the target gender effect, there are 6 designs for which the coefficient of the target gender effect does not include 0. Across those designs, male targets are accorded higher status than female targets (regardless of emotion condition) for 5 designs and females are accorded higher status than males for 1 design. See Figure 2 which designs show these effects. Corroborating the visual pattern, the Bayes factor analysis suggests that the unconstrained random effects model strongly outperforms the baseline-model that includes only varying intercepts for site and design; $\text{BF}_{rb} = 2.5 \times 10^7$. Note however, that the pattern of between design variation in the size and direction of the target gender effect is not predicted by any of the theoretical models.

Figure 3 shows the intercepts and varying effects across design for competence ratings as the outcome measure. In line with the Bayes factor analysis providing most evidence for the unconstrained model, the pattern in panel C displays a *negative* effect of anger vs. non-anger across 13 designs (i.e., an effect opposite to predictions by the Status Signalling perspective). None of the designs show a target gender effect or a gender-by-emotion interaction effect. As predicted by the 'Anger suppresses warmth' perspective, a similar pattern emerges for perceived warmth in Figure 4: across 23 designs the 95% credible interval of the coefficient for target emotion excludes zero, with non-angry targets considered warmer than angry targets. In addition, across 3 designs, female targets are considered warmer than male targets, while for 3 designs, male targets are considered warmer than female targets. Finally, in 2 designs, a gender stereotype effect is observed such that an angry male target is considered warmer than a non-angry male targets and an angry female target, while an angry female target is considered less warm than a non-angry female target and an angry male target. For out-of-control ratings, a reverse pattern for the target emotion effect is shown in Figure 5: across 16 designs the 95% credible interval of the coefficient for target emotion excludes zero, with angry targets considered more out of control than non-angry targets.

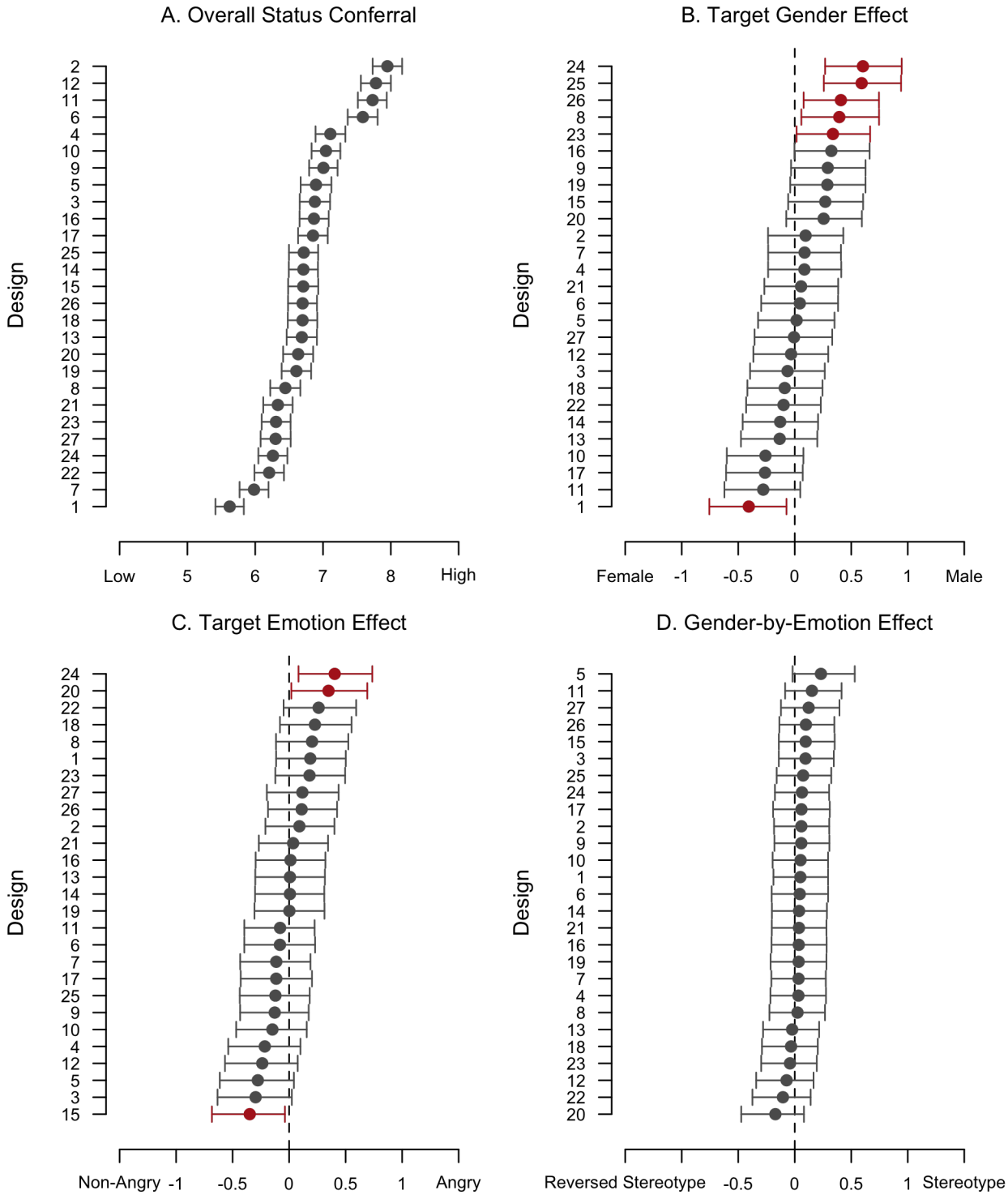


Figure 2. Estimated design-level effects (posterior means) on status conferral, in increasing order. **A.** Intercepts. **B.** Target gender effects. **C.** Target emotion effects. **D.** Gender-by-emotion interaction effects. Each dot represents a design. The horizontal lines denote the 95% credible intervals. Estimated effects for which the 95% credible interval includes zero are displayed in grey and estimated effects for which the 95% credible intervals exclude zero are displayed in red. Only the target gender effect seem to differ from zero for some designs.

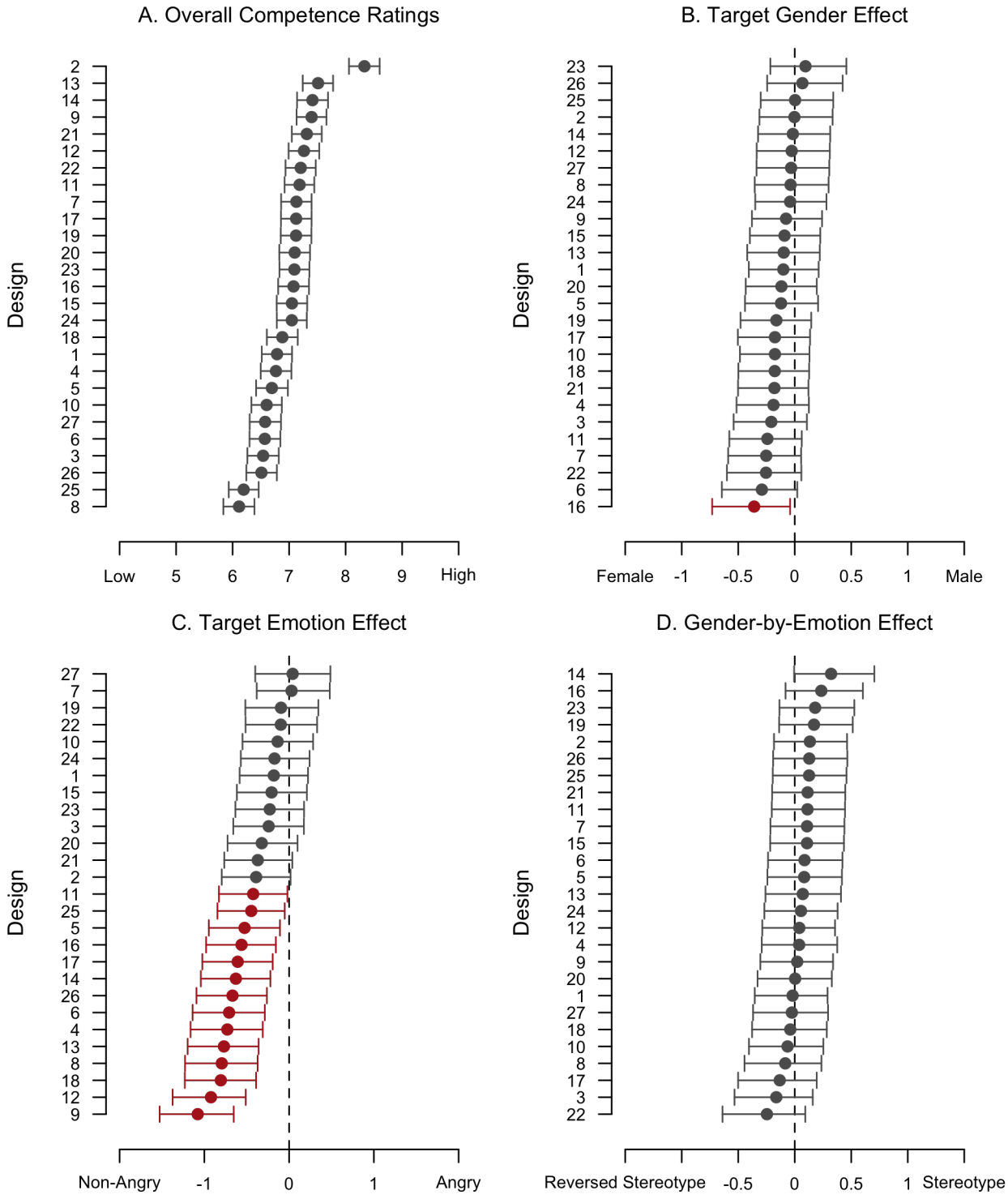


Figure 3. Estimated design-level effects (posterior means) on competence, in increasing order. **A.** Intercepts. **B.** Target gender effects. **C.** Target emotion effects. **D.** Gender-by-emotion interaction effects. Each dot represents a design. The horizontal lines denote the 95% credible intervals. Estimated effects for which the 95% credible interval includes zero are displayed in grey and estimated effects for which the 95% credible intervals exclude zero are displayed in red. A target emotion effect emerges across about half of the designs, yet in the direction opposite as predicted by the Status Signalling perspective.

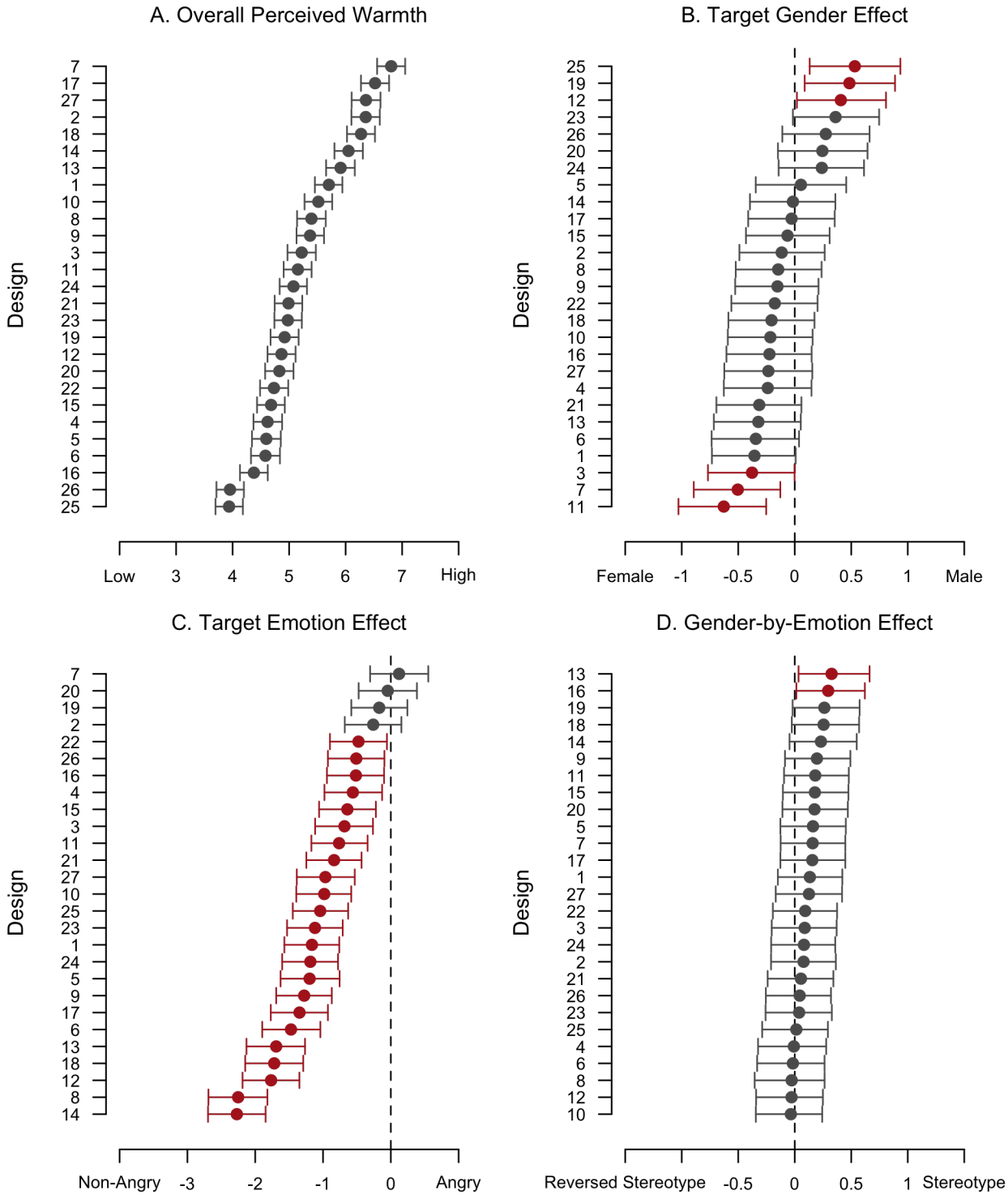


Figure 4. Estimated design-level effects (posterior means) on warmth, in increasing order. **A.** Intercepts. **B.** Target gender effects. **C.** Target emotion effects. **D.** Gender-by-emotion interaction effects. Each dot represents a design. The horizontal lines denote the 95% credible intervals. Estimated effects for which the 95% credible interval includes zero are displayed in grey and estimated effects for which the 95% credible intervals exclude zero are displayed in red. In line with the Anger Suppresses Warmth perspective, a target emotion effect emerges across most designs.

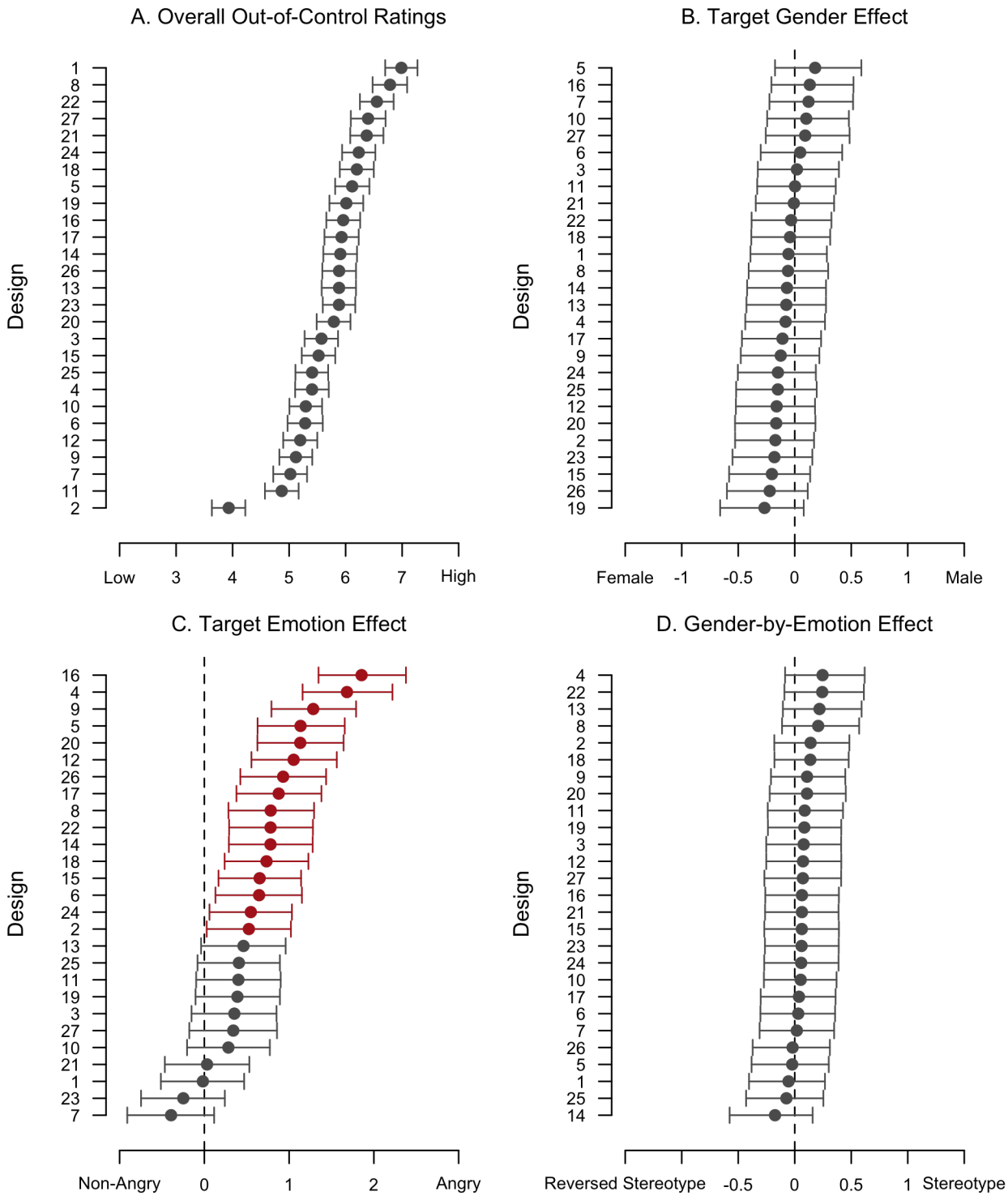


Figure 5. Estimated design-level effects (posterior means) on out-of-control ratings, in increasing order. **A.** Intercepts. **B.** Target gender effects. **C.** Target emotion effects. **D.** Gender-by-emotion interaction effects. Each dot represents a design. The horizontal lines denote the 95% credible intervals. Estimated effects for which the 95% credible interval includes zero are displayed in grey and estimated effects for which the 95% credible intervals exclude zero are displayed in red. In line with the Status Signalling perspective, a target emotion effect emerges across about half of the designs.

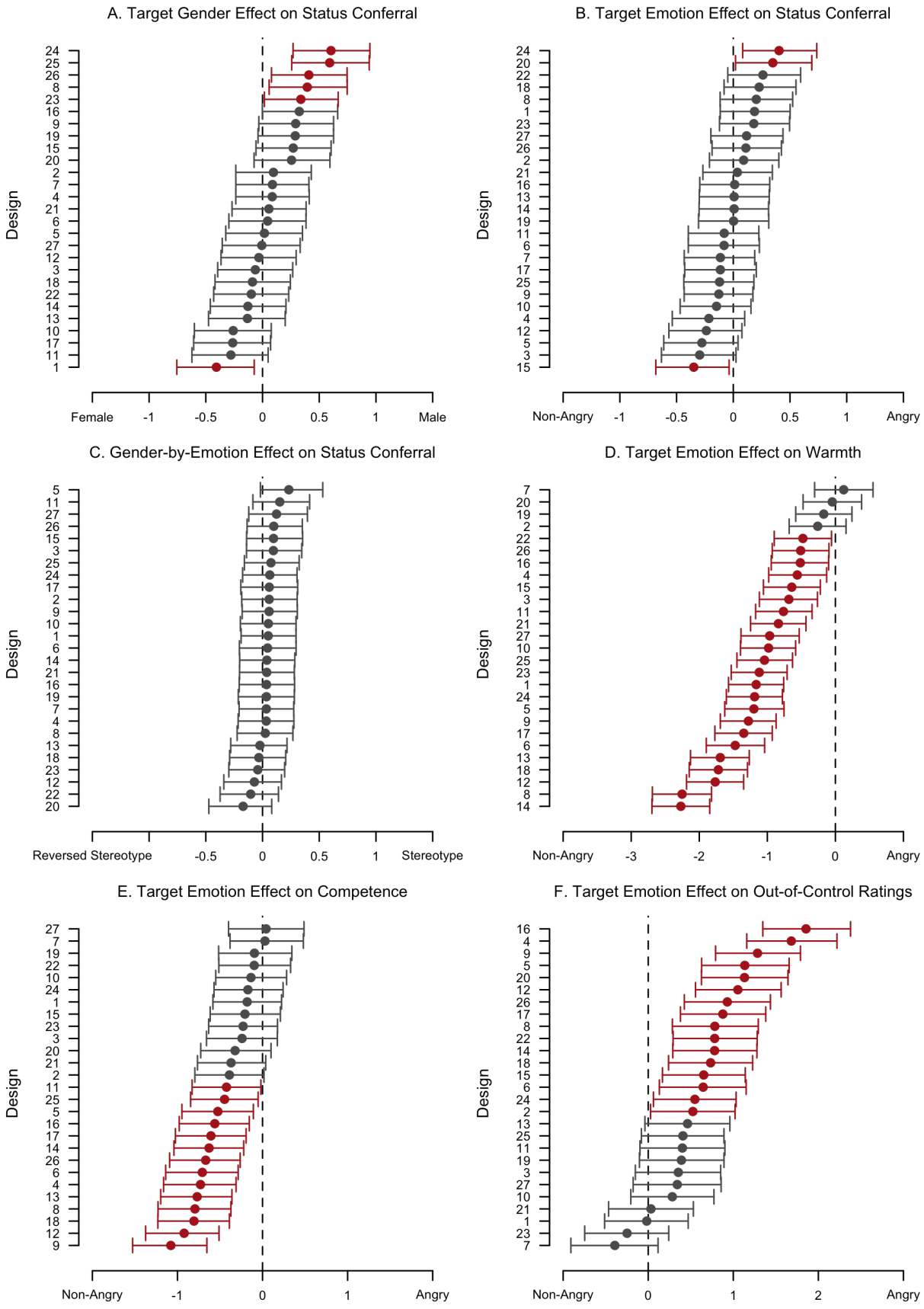


Figure 6. Estimated design-level effects (posterior means), in increasing order. **A.** Target gender effect on status conferral. **B.** Target emotion effects on status conferral. **C.** Target gender-by-emotion interaction effect on status conferral. **D.** Target emotion effects on warmth. **E.** Target emotion effect on competence. **F.** Target emotion effect on out-of-control

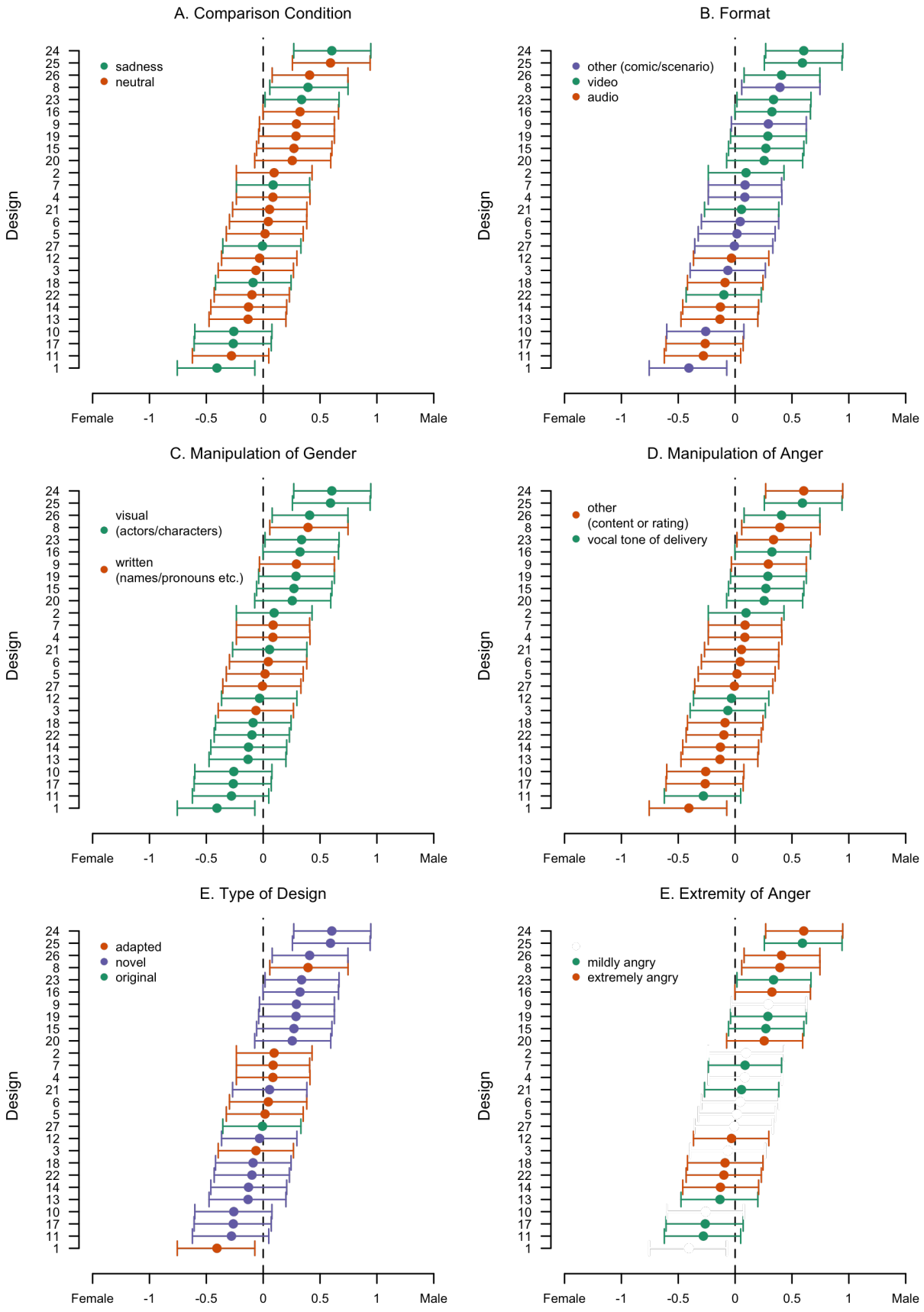


Figure 7. Varying effect of target gender per design, colored based on design-related moderators.

For status conferral as the outcome variable, we further investigated to what extent the target gender effect might be explained by design-related moderators. Based on visual inspection, it seems that only the format of the manipulation might have a systematic influence on the target gender effect. That is, when presented in a video, male targets might be accorded more status than female targets, whereas this effect is absent when presented as an audio recording or comic book or written scenario.

In order to further unpack this potential effect of format, we conducted a Bayesian ANOVA with gender and format as factors plus their interaction. Indeed, we find that the data provide most evidence for the model that includes the factors gender and format, plus their interaction. Comparing the models with and without this interaction term gives a Bayes factor of 1.1×10^5 in favor or the gender-by-format interaction. As seen in Figure 7B, the video format particularly increases status conferral to male targets. Indeed, if we look at designs with a video format and other designs separately, we find evidence in favor of a gender effect for video designs $\text{BF}_{10} = 71{,}896$ and evidence against a gender effect for the other designs (audio, scenarios, comic book): $\text{BF}_{10} = 0.11$; $\text{BF}_{01} = 9.22$.

Anger Extremity Hypothesis A. We predicted that moderate levels of anger would lead to more status conferral than extreme anger for both targets. To this end, we compared status conferral to angry targets between designs that were identical except for the extremity of the displayed anger. However, the data indicated that mild anger led to *less* status conferral compared to extreme anger ($M_{mild} = 6.64$); $M_{extreme} = 6.83$). As the effect went in the opposite direction, the Bayes factor analysis also provided evidence against the anger extremity hypothesis: $\text{BF}_{10} = 0.01$; $\text{BF}_{01} = 78.2$.

Secondary Theoretical Tests

As preregistered, we tested several additional effects and conducted relevant analyses.

Cultural Differences. The cultural differences perspective predicts that anger may

Table 2
Evidence for the cultural differences perspective.

	Bayes Factors			
	BF_{0u}	BF_{0cd}	BF_{0main}	BF_{0ci}
Assertiveness Values Lab Location	309	2.48	29.6	451
Assertiveness Values Country	131	1.08	14.5	213
Assertiveness Practices Lab Location	387	2.95	33.5	354
Assertiveness Practices Country	145	0.85	11.0	170
Disagreement Lab Location	53.3	0.75	10.4	97.8
Disagreement Country	9.85	0.39	4.81	19.2

Note. The Bayes factors reflect evidence for the null-model versus the models that include effects of culture dimension, target emotion, and their interaction. BF_{0u} gives the evidence for null model (indicated by the subscript 0) versus the unconstrained model (indicated by the subscript u). Subscript *cd* refers to the cultural differences only model, subscript *main* refers to the main effects model (cultural differences and target emotion), and subscript *ci* refers to the culture interaction model (cultural differences-by-target emotion interaction). See text for details about the different models. The null-model outperforms all more complex models for each of the different operationalisations of the culture dimension.

have positive effects in confrontation (disagreement/assertiveness) oriented cultures but negative effects in more harmony-oriented cultures. This translates into an interaction such that with increased cultural harmony-scores, the positive effect of anger decreases or becomes negative (i.e., a negative interaction coefficient). We tested the effect of the culture dimension as well as the interaction between culture dimension and target emotion on status conferral across 6 different operationalisations of the culture dimension: assertiveness values ratings of the lab's location and of the participants' indicated country of residence, assertiveness practices ratings of the lab's location and of the participants' indicated country of residence, and disagreement scores of the lab's location and of the participants' indicated country of residence. As shown in Table 2, across all operationalisations, the null-model outperforms the culture dimension only model, the culture dimension plus target emotion model, and the culture dimension-by-target emotion interaction model (i.e., the model of interest).

Main effect of target. Are male or female targets generally accorded more status? We ran a simple independent-samples t-test to answer this question. In contrast to Study 1, female targets are accorded slightly *less* status than male targets ($M_{female} = 6.73$; $M_{male} = 6.80$), yet the Bayes factor analysis provides moderate evidence against the gender effect: $\text{BF}_{10} = 0.22$; $\text{BF}_{01} = 4.56$ (Frequentist statistics: $t(7, 107.88) = -1.62$, $p = .105$). For the other dependent variables, we also found evidence in favor of the null-hypotheses; for competence $\text{BF}_{10} = 0.51$; $\text{BF}_{01} = 1.98$, for warmth $\text{BF}_{10} = 0.06$; $\text{BF}_{01} = 15.6$, and for out-of-control $\text{BF}_{10} = 0.04$; $\text{BF}_{01} = 24.4$.

Main effect of emotion. Are angry or not-angry targets generally accorded more status? A simple independent-samples t-test showed evidence against a main effect of emotion on status conferral: $\text{BF}_{10} = 0.03$; $\text{BF}_{01} = 36.9$ (Frequentist statistics: $t(7, 115.92) = 0.08$, $p = .940$). For the other dependent variables, we did find evidence in favor of a main effect of emotion; for competence $\text{BF}_{10} = 3 \times 10^{15}$, for warmth $\text{BF}_{10} = 6.2 \times 10^{86}$, and for out-of-control $\text{BF}_{10} = 1.4 \times 10^{25}$. As becomes evident from Figure 1, for competence and warmth, angry targets are considered less competent and less warm than non-angry targets, respectively, and for out of control, angry targets are considered more out of control than non-angry targets.

Gender of the rater. Do male and female raters differentially display the main effects or interactions derived from the theoretical perspectives (e.g., do male or female raters particularly show gender stereotyping effects or reversed gender biases?). The Bayes factor analysis indicates that the data is slightly more likely under the null-model than under the moderator interaction model (which assumes a three-way interaction between target gender, target emotion, and rater gender): $\text{BF}_{0m} = 1.88$. In contrast to Study 1, there was no evidence that female participants accorded more status to female targets ($M_{female} = 6.78$) than to male targets ($M_{male} = 6.85$): $\text{BF}_{10} = 0.09$; $\text{BF}_{01} = 10.73$. For male participants, the analysis indicated no evidence either way ($M_{female} = 6.57$; $M_{male} = 6.75$): $\text{BF}_{10} = 0.90$,

Table 3
Evidence for gender equality beliefs and experimental effects on status conferral (perspective 3).

	Bayes Factors			
	BF_{i0}	BF_{ie}	BF_{im}	BF_{iu}
Beliefs Gender Workplace	3.4×10^8	215	405	806
Sexist Beliefs	0.37	196	51.2	102
News Exposure	1.56	210	53,595	2,380
Internal Motivation	1.3×10^9	275	48,176	54,978

Note. The Bayes factors reflect evidence for the individual differences only model vs the interaction effects models. BF_{i0} gives the evidence for the individual differences only model (indicated by the subscript i) versus the null-model (intercept only; indicated by the subscript 0). Subscript e refers to the experimental effects model, subscript m refers to the moderation model, and subscript u refers to the unconstrained model. See text for details about the different models.

$BF_{01} = 1.11$. Again, there was no evidence for an interaction: $BF_{10} = 0.03$, $BF_{01} = 34.57$ (see also Figure 12).

Individual differences. Various individual difference moderators were included in the study in order to unpack any observed experimental effects. Specifically, scales related to beliefs about gender (in)equality as well as about self-presentation concerns were added to distinguish between perspective 3 and 4, which predict the same overall pattern of responses, yet driven by different factors.

For the moderator analyses, we constructed different models to reflect theoretical predictions. We start again from the null-model ($\mathcal{M}_0$) that includes only a random intercept per site and design. The second model ($\mathcal{M}_i$) additionally includes the *individual differences* variable of interest (e.g., sexist beliefs) but without any experimental effects or interactions. This model might be considered the baseline model for the moderation analyses. The third model ($\mathcal{M}_e$) extends the individual differences model by adding the main *experimental effects* of target emotion, target gender and its interaction. The fourth model ($\mathcal{M}_m$) is the

Table 4
Evidence for self-presentation concerns and experimental effects on status conferral (perspective 4).

	BF_{i0}	BF_{ie}	BF_{im}	BF_{iu}
External Motivation No Sexism	2.82	240	1.3×10^5	1,864
Research Study Experience	0.54	214	9,054	14,820
Participated in Similar Study	0.47	231	3,065	2,975
Taken Psychology Course	0.18	224	3,874	5,217
Awareness of Target Gender	20.8	406	1.8×10^5	2,819
Awareness of Target Emotion	0.04	216	1.1×10^5	5,310

Note. The Bayes factors reflect evidence for the individual differences only model vs the interaction effects models. BF_{i0} gives the evidence for the individual differences only model (indicated by the subscript i) versus the null-model (intercept only; indicated by the subscript 0). Subscript e refers to the experimental effects model, subscript m refers to the moderation model, and subscript u refers to the unconstrained model. See text for details about the different models.

critical moderation model that adds the interaction term between the individual differences variable and target gender, as well as the threeway interaction between the individual difference variable, target gender, and target emotion. Based on the moderation hypotheses, the sign of the interaction between the moderator and target gender was restricted (e.g., sexist beliefs are expected to be associated with a *decrease* in status conferral to female targets relative to male targets). Finally, we constructed a fully unconstrained model ($\mathcal{M}_u$) that includes the same terms as the moderation model but does not put any ordinal constraints on the parameters.

The results of the individual differences moderation analyses are given in Table 3 (for moderation effects related to beliefs about gender (in)equality; perspective 3) and Table 4 (for moderation effects related to self-presentation concerns and study savviness; perspective 4). As the Bayes factors show, there is evidence that some of the individual difference measures (i.e., beliefs about gender in the workplace, internal motivation not to appear sexist, external motivation not to appear sexist, and awareness of the target gender) are

related to status conferral in general (main effects of individual differences), but strong evidence *against* moderation effects of the individual differences on the experimental target gender, target emotion, or target gender-by-emotion effects (BF_{im}).

General public vs. students. Finally, we assessed whether students and adults from the general public differed in the extent to which they applied (reversed) gender stereotyping. The student vs. general public status was determined at the individual level (rather than the sample level) based on responses to the item 'Are you currently a university student?' However, again the individual differences-only model (main effect of student status) outperformed all other models including those with a moderating effect of student status (BF_{im} = 1,018; see also Figure 13).

Secondary Methodological Tests

As preregistered, we also investigated continuous participant ratings of anger extremity, appropriateness, dominance and warmth as moderators.

Anger extremity hypothesis B. We expected that more extreme anger, as rated by participants, should be associated with reduced status conferral from anger expressions for both female and male targets. In order to test this hypothesis, we ran a Bayesian correlation analysis with perceived anger and status conferral as variables on the subset of the data including only the anger condition. As we expected a negative relation, the interval was restricted to cover [-1,0]. The data, however, strongly favored the null-hypothesis: BF_{10} = 0.01; BF_{01} = 109.21 and the correlation coefficient was in fact slightly positive ($\rho = 0.05$).

Appropriateness hypothesis. Here we expected that less appropriate anger expressions, as rated by participants, should be associated with reduced status conferral from anger expressions for both female and male targets. That is, we expected a positive correlation between appropriateness and status conferral in the anger-condition subset of the data. The data suggested that such a correlation was indeed present: $BF_{10} = 4.2 \times 10^{17}$, $\rho =$

0.15.

Target dominance hypothesis. We expected more backlash against angry women in status conferral (i.e., a stronger interaction between target gender and emotion expression) to the extent the anger expression projects dominance. This means a moderating effect of dominance on the interaction between target gender and emotion. (Note that this hypothesis is unlikely given that we did not find evidence for a backlash effect against women in the first place.) Although the moderation-interaction model outperforms the individual-differences-only model ($BF_{mi} = 14.34$), the model performs worse than the experimental-effects-only model that excludes the crucial threeway interaction between gender, emotion, and perceived dominance: $BF_{me} = 0.06$; $BF_{em} = 15.59$.

Target warmth hypothesis. Here we expected more backlash against angry women in terms of reduced status conferral (i.e., a stronger interaction between target gender and emotion expression) to the extent the anger expression projects a lack of warmth. This translates into a moderating effect of warmth on the interaction between target gender and emotion. The data, however, provide more evidence for the individual-differences-only model than for the moderator-interaction model $BF_{im} = 106.51$.

Multiverse Analysis

As in Study 1, for the main analyses we used the intent-to-treat approach, in which few to no observations or participants are excluded (Gupta, 2011; McCoy, 2017). As preregistered we included the following paths in the multiverse:

- Language fluency.

1. Include everyone regardless of years of English experience (or other language in which the survey is administered)
2. Exclude participants with less than 5 years of English experience (or other language in

which the survey is administered)

- Manipulation check.

1. Exclude no one based on the manipulation check for target gender
2. Exclude participants who did not correctly indicate the target's gender

- Straightlining.

1. Exclude no one based on pattern of responding
2. Exclude participants who always selected the same option on all items within each of the 5 scales

- Material quality.

1. Exclude no one based on reported quality of video/audio materials
2. Exclude participants from designs with video/audio stimuli who indicated poor video/audio quality (<3 on the 7-point scale)

- Mode of administration.

1. Exclude no one based on method of taking the survey
2. Exclude participants who did not complete the study on a laptop or desktop

- Manipulation failure designs.

1. Include all designs
2. Exclude designs for which there is no compelling evidence (i.e., BF< 10) for both of the manipulation checks (target gender and target emotion)
3. Exclude designs for which there is no compelling evidence (i.e., BF< 10) for either of the manipulation checks (target gender or target emotion).

In the preregistration we specified that we would only analyze a specific

exclusion-based multiverse path when the exclusion affect at least 5% of the sample. However, the straightlining criterion did not reach this threshold; only 1.17% of the sample was excluded because of straightlining across all 5 scales. The language fluency (5.09%), target gender manipulation check item (10.23%), the material quality (7.85%), and the mode of administration (9.44%) did result in a meaningful proportion of exclusion. There was only 1 design (i.e., design 7) for which there was no compelling evidence that targets in the anger condition were perceived as angrier than in the not-angry condition ($BF_{10} = 2.67$). This, however, meant (3.74%) of the sample, which does not pass the threshold of 5%.

```
##                                    Comparison       BFs.1       BFs.2
## 1            Gender Stereotyping vs. Designs 0.005265951 0.022957574
## 2              Status Signalling vs. Designs 0.002934172 0.017090810
## 3 Culture Change/Study Savviness vs. Designs 0.002864617 0.002262296
##         BFs.3       BFs.4       BFs.5       BFs.6       BFs.7       BFs.8
## 1 0.006549733 0.033596960 0.005453602 0.003423239 0.011858325 0.004643797
## 2 0.001108746 0.020850444 0.006438586 0.000928468 0.009392627 0.002073944
## 3 0.001918251 0.002763799 0.004364849 0.002998016 0.002682945 0.002847272
##         BFs.9      BFs.10      BFs.11      BFs.12      BFs.13      BFs.14
## 1 0.041012008 0.008424213 0.002013890 0.014104497 0.004808158 0.007132289
## 2 0.046332368 0.002813369 0.001334853 0.021372164 0.002032638 0.005325789
## 3 0.002460528 0.002243294 0.005363923 0.004291009 0.003743372 0.003671794
##        BFs.15      BFs.16
## 1 0.003479007 0.015142688
## 2 0.003311039 0.004991273
## 3 0.005580447 0.001415130

##                                    Comparison BFs.1 BFs.2 BFs.3 BFs.4 BFs.5
## 1            Gender Stereotyping vs. Designs     0     0     0     0     0
```

```
## 2              Status Signalling vs. Designs     0     0     0     0     0
## 3 Culture Change/Study Savviness vs. Designs     0     0     0     0     0
##   BFs.6 BFs.7 BFs.8 BFs.9 BFs.10 BFs.11 BFs.12 BFs.13 BFs.14 BFs.15 BFs.16
## 1     0     0     0     0      0      0      0      0      0      0      0
## 2     0     0     0     0      0      0      0      0      0      0      0
## 3     0     0     0     0      0      0      0      0      0      0      0
```

```
##                                  Comparison         BFs.1        BFs.2
## 1            Gender Stereotyping vs. Designs  0.000000e+00  0.00000e+00
## 2              Status Signalling vs. Designs  0.000000e+00  0.00000e+00
## 3 Culture Change/Study Savviness vs. Designs  0.000000e+00  0.00000e+00
## 6                         Warmth vs. Designs 1.950767e+101 6.97715e+101
##           BFs.3        BFs.4         BFs.5         BFs.6        BFs.7
## 1  0.000000e+00 0.000000e+00  0.000000e+00  0.000000e+00 0.000000e+00
## 2  0.000000e+00 0.000000e+00  0.000000e+00  0.000000e+00 0.000000e+00
## 3  0.000000e+00 0.000000e+00  0.000000e+00  0.000000e+00 0.000000e+00
## 6 7.149336e+102 3.097661e+94 1.103075e+101 1.324368e+101 3.766778e+94
##           BFs.8        BFs.9       BFs.10       BFs.11       BFs.12
## 1  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## 2  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## 3  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## 6 6.531369e+100 7.003451e+95 7.384429e+94 1.321902e+99 5.852399e+95
##         BFs.13       BFs.14       BFs.15        BFs.16
## 1 0.000000e+00 0.000000e+00 0.000000e+00  0.000000e+00
## 2 0.000000e+00 0.000000e+00 0.000000e+00  0.000000e+00
## 3 0.000000e+00 0.000000e+00 0.000000e+00  0.000000e+00
## 6 1.149269e+94 5.782551e+94 6.124491e+93 3.864659e+102
```

```
##                                  Comparison        BFs.1        BFs.2
## 1            Gender Stereotyping vs. Designs 0.000000e+00 0.000000e+00
## 2              Status Signalling vs. Designs 2.237352e+24 2.876029e+27
## 3 Culture Change/Study Savviness vs. Designs 2.617518e+09 1.029790e+11
##          BFs.3        BFs.4        BFs.5        BFs.6                  BFs.7
## 1 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00                      0
## 2 8.501245e+25 1.089073e+24 1.441275e+25 6.204109e+22 948746272726215163904
## 3 2.052810e+10 2.858244e+09 1.186948e+10 9.483907e+08              150035253
##          BFs.8        BFs.9       BFs.10       BFs.11                  BFs.12
## 1 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00                      0
## 2 1.950379e+26 5.035988e+23 2.002023e+22 2.702270e+23 2808468239663169011712
## 3 3.429519e+10 4.316041e+09 7.375521e+08 2.241799e+09               461761283
##                BFs.13       BFs.14               BFs.15       BFs.16
## 1                   0 0.000000e+00                    0 0.000000e+00
## 2 8753602340931445760 1.438549e+22 19614221474964393984 1.098387e+27
## 3            29213805 8.166583e+08             53979512 1.889728e+10
```

Figure 8A displays the results of the 16 viable multiverse path for the main analysis of status conferral in which the models corresponding to the different theoretical perspectives are compared to the baseline model (random intercepts for sites and designs). For status conferral, given the most favorable set of exclusions, the data are still 21.58 times more likely under the baseline model than under the status signaling model.

Exploratory Analyses

Target Gender Effect for the Brescoll and Uhlmann (2008) design across Studies 1 and 2. Our present Study 1 finds a tendency for greater to status attribution to women, whereas Study 2 finds greater status attribution to men. If we only consider the original design (i.e., design 27) in Study 2, we find no evidence for a target gender effect:

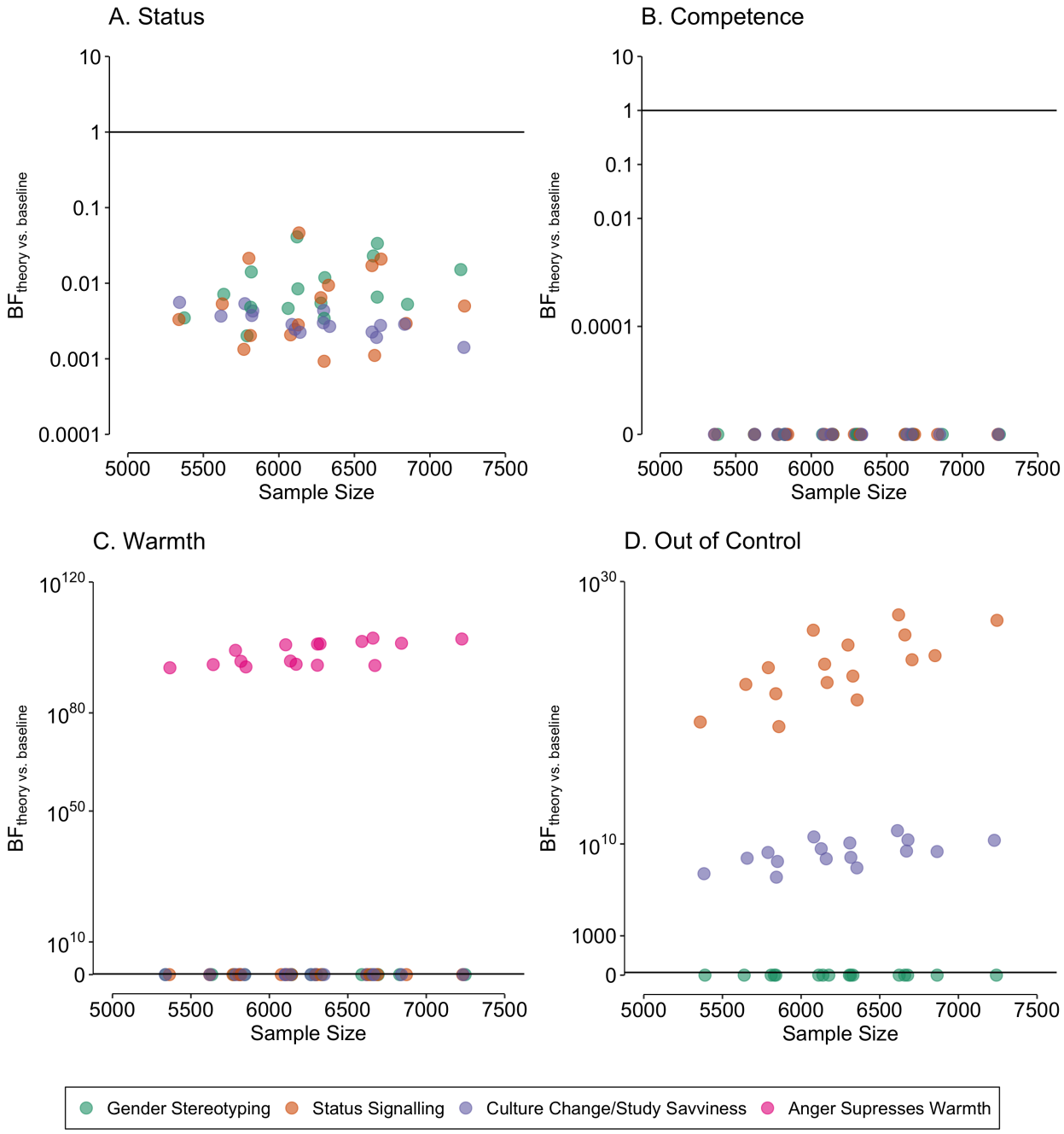
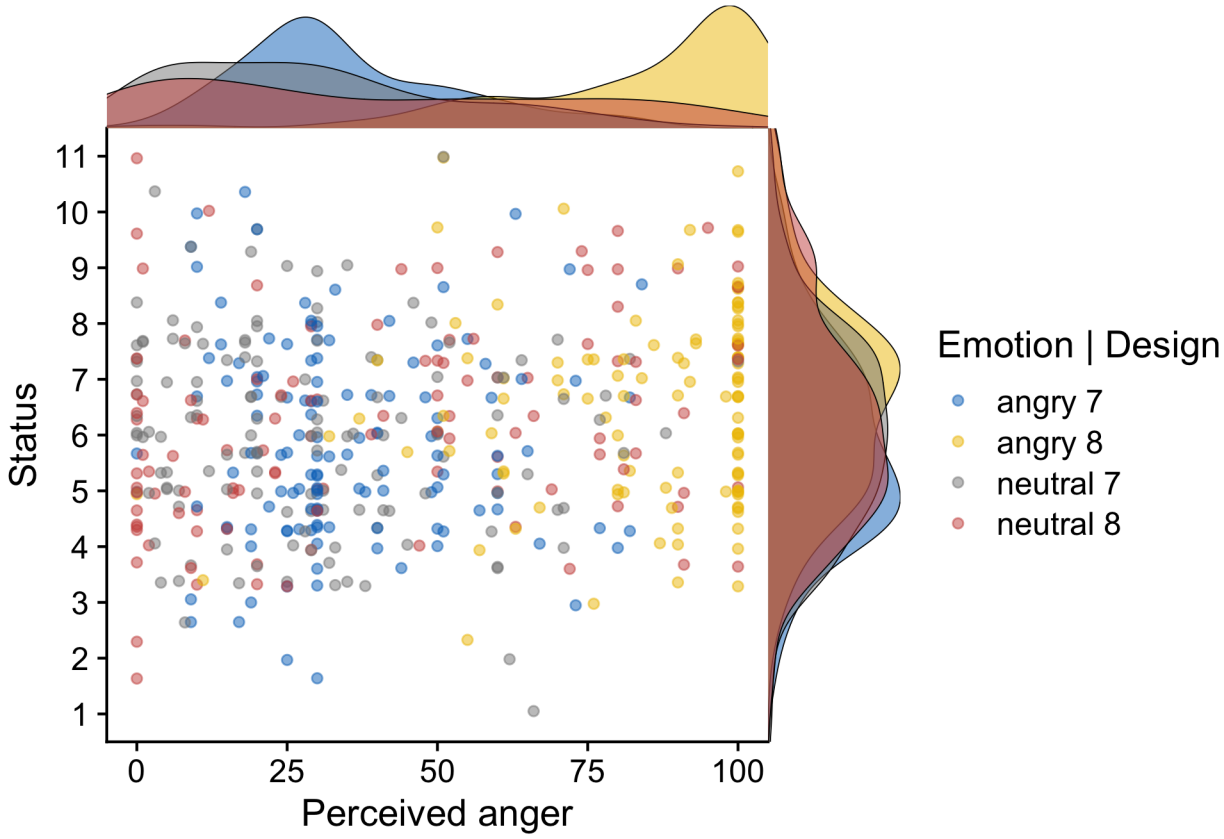


Figure 8. Results from the multiverse analysis: Bayes factors in favor of a theoretical perspective are above the horizontal line, Bayes factors against the theoretical perspectives (in favor of the intercepts-only baseline model) are below the horizontal line. The color of the points reflects different theoretical perspectives and the x-axis reflects the number of participants the analysis is based on. Note that the y-axis is on a log-scale that differs between panels. For status conferral and competence, all analyses provide evidence against the theoretical perspectives. For out-of-control ratings, the status signalling perspective is the clear winner and for warmth, the anger-suppresses-warmth perspective is the clear winner.

$BF_{10} = 0.14$; $BF_{01} = 6.94$, $M_{male} = 6.26$, $M_{female} = 6.22$. The absence of a gender difference in the original design can also be observed from Figure 7E, where the target gender coefficient is almost exactly 0.



Anger Extremity Effect for the Gaertig et al. (2019) design. We conducted a directed Bayesian t-test to see if we can replicate Gaertig et al.'s finding that moderate anger leads to more status conferral than extreme anger (regardless of target gender), by looking at their design (i.e., 7&8) in isolation. However, the data indicate *less* status conferral to moderate anger ($M_{moderate} = 5.93$) than to extreme anger ($M_{extreme} = 6.41$). The directed Bayesian t-test confirms strong evidence against the hypothesis: $BF_{+0} = 0.03$; $BF_{0+} = 39.15$. In fact, when we remove the directionality constraint, we get some moderate evidence for a difference (in the opposite direction): $BF_{10} = 6.31$.

Relatedly, we explored how the Gaertig et al. (2019) design compared to the others

that directly manipulated anger extremity. To unpack the anger extremity effect on status conferral, we estimated the overall status conferral per design, using varying intercepts per design and a fixed effect of the extremity manipulation. Figure 9 shows the estimated status conferral for the 18 designs in which extremity was manipulated directly, as well as the perceived anger extremity as rated by participants in the anger condition. Notably, the extremity effect (i.e., difference between the mild and extreme version) is the largest in the original Gaertig et al. (2019) design, yet in the opposite direction as found by Gaertig et al. (2019) Furthermore, while perceived anger extremity is consistently higher in the the extreme anger version of each design, the difference is clearly largest in the original Gaertig et al. design (7&8).

Gender Effect for the McCormick-Huhn & Shields (2019) design. Here, we assessed whether the pattern in the original McCormick-Huhn & Shields (2019) storyboard design replicates, such that angry women are accorded *more* status than angry men. In the original McCormick-Huhn & Shields (2019) study, only anger conditions were included. To replicate their results, we ran a directed Bayesian t-test with target gender as independent variable on the subset of the data with design 1 and only the anger condition. The data indeed indicate slightly more status conferral to angry female ($M_{female} = 6.02$) than to angry male targets ($M_{male} = 5.45$). However, the directed Bayesian t-test suggests that these data are quite undiagnostic and provide only anecdotal the hypothesis: $\text{BF}_{+0} = 2.76$. Note that there are only 77 participants in the angry female target condition and 71 participants in the angry male target condition.

An angry woman can get whatever she wants

Here, we demonstrate how we can get the desired result of replicating the original Brescoll & Uhlmann (2008) result if we cherry-pick one of the many reasonable analysis paths. In this case, we run the multiverse using the specified data exclusion settings again, this time using a single item 'status conferral' measure – simply taking the item on status

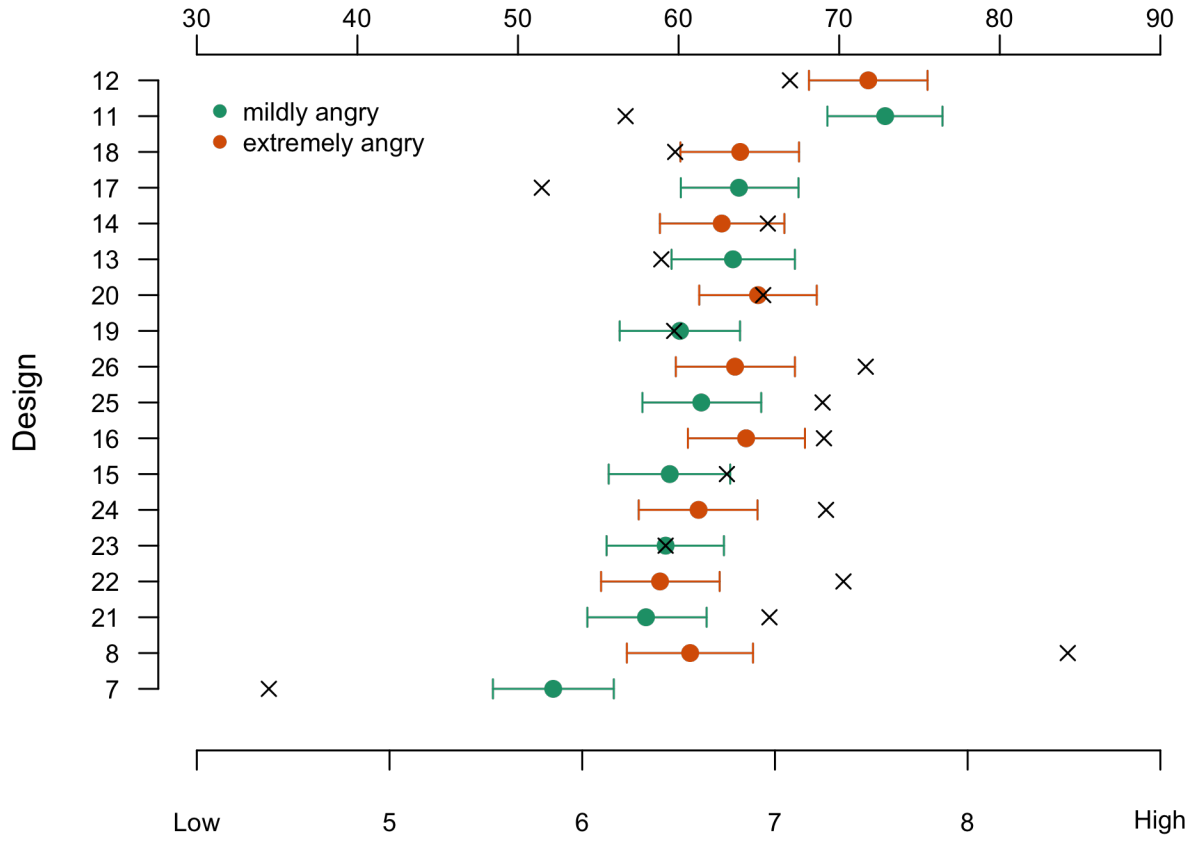


Figure 9. Estimated status conferral per design, colored based on anger extremity. Estimates are grouped per design and ordered based on the grouped average. The x's reflect the average perceived anger extremity in the anger condition per design, as rated by participants. The x-axis at the top gives the scale of the perceived anger extremity ratings (on a 0-100 scale) and the x-axis at the bottem gives scale of the estimated status conferral (on a 1-10 scale).

conferral rather than the composite of status, power, and independence.

So which are those highlighted paths? What are the corresponding specifications?

Evidence in favor of the gender stereotyping perspective only occurs when we use (1) the single-item dependent variable, (2a) *only* exclude participants who did not complete the study on a computer OR (2b) exclude participants who did not complete the study on a computer AND did not correctly recall the target gender. However, if we use any other constellation of exclusion criteria we get evidence *against* the gender stereotyping perspective, such as (2c) exclude participants who did not complete the study on a computer

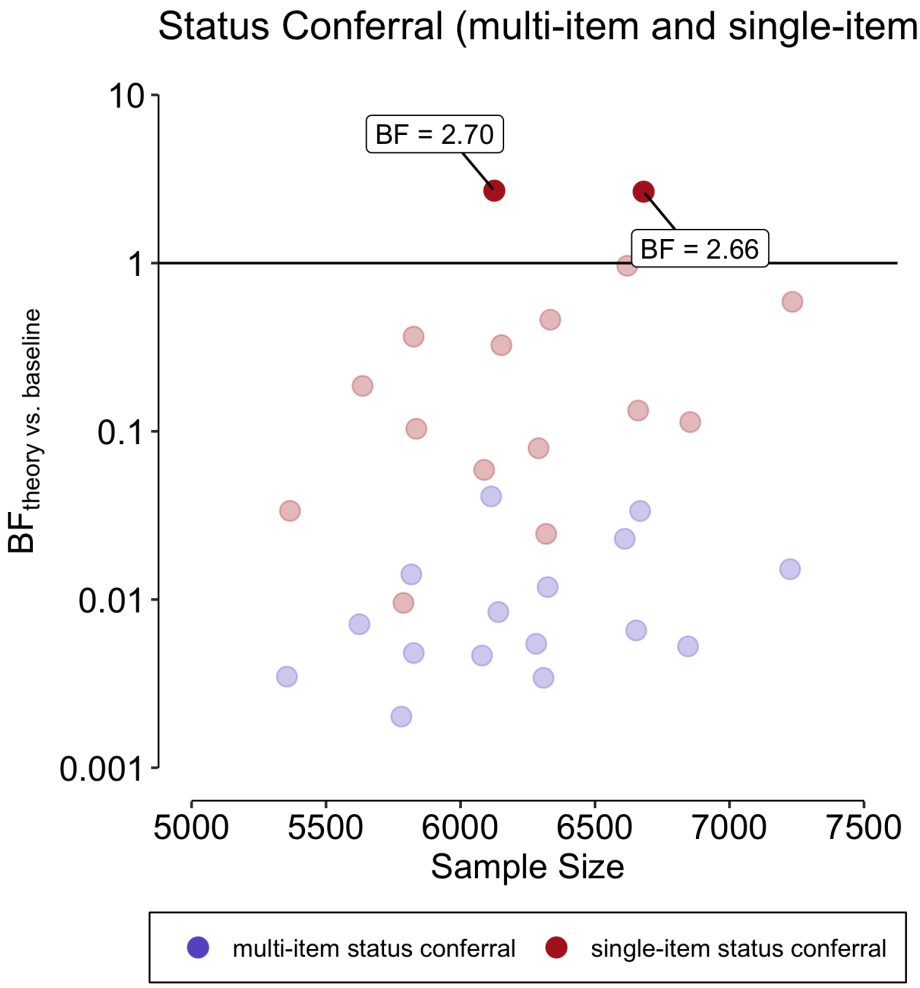


Figure 10. Results from the multiverse analysis: Bayes factors in favor of a theoretical perspective are above the horizontal line, Bayes factors against the theoretical perspectives (in favor of the intercepts-only baseline model) are below the horizontal line. The color of the points reflects different theoretical perspectives, the shape of the points reflects the version of the dependent variable, and the x-axis reflects the number of participants the analysis is based on. The majority of analyses provide evidence against the theoretical perspectives.

and exclude participants who have less than 5 years of English experience, or (2d) exclude participants who did not complete the study on a computer and participants who did not correctly recall the target gender and participants from designs with video/audio stimuli who indicated poor video/audio quality.

For those two paths where we obtained evidence supporting the gender stereotyping perspective, we get a Bayes factor of 2.66 with 6679 participants and a Bayes factor of 2.70 with 6125 participants.

Discussion

References

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, *2*, 109–112. doi:10.4103/2229-3485.83221

Haaf, J. M., Klaassen, F., & Rouder, J. N. (2018). Capturing Ordinal Theoretical Constraint in Psychological Science. *PsyArXiv*. doi:10.31234/osf.io/a4xu9

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*, 779–798. doi:10.31234/osf.io/ktjnq

Lee, M. D., & Wagenmakers, E. J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge (UK): Cambridge University Press.

McCoy, C. E. (2017). Understanding the Intention-to-treat Principle in Randomized Controlled Trials. *Western Journal of Emergency Medicine*, *18*, 1075–1078. doi:10.5811/westjem.2017.8.35985

Rouder, J. N., Haaf, J. M., Davis-Stober, C. P., & Hilgard, J. (2019). Beyond overall effects: A Bayesian approach to finding constraints in meta-analysis. *Psychological Methods*, *24*, 606–621. doi:10.1037/met0000216

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*, 702–712. doi:10.1177/1745691616658637

Appendix

Model Specification

For the main analyses, we used Bayesian hierarchical modeling with participants nested in countries/labs and in designs. As the random-effects model in Study 1 indicated evidence against variability between countries/labs, and since we're interested in between-design variability in Study 2, we used varying intercepts across sites and varying intercepts + varying slopes across designs. We first constructed an unconstrained model that includes all main parameters from the separate theories, which are free to vary in size and direction. In the primary analysis, we used different ordinal constraints to capture the different theoretical predictions (see below). Bayes factor model comparison is used to compare the models and determine what theory best predicts the empirical data. This method is based on the work by Haaf et al. (2018).

Let Y_{ijklm} be the status rating of the ith lab, the jth design, the kth participant in the lth target gender condition ($l = 1, 2$, for female and male targets, respectively) and the mth target emotion condition ($m = 1, 2$, for not-anger and anger expressions, respectively). Then

$$Y_{ijklm} \sim N(\alpha_i + \eta_j + x_{1jl}\beta_j + x_{2jm}\gamma_j + x_{3jlm}\theta_j, \sigma^2),$$

where,

- α_i is the baseline rating for the ith lab
- η_j is the baseline rating for the jth design
- β_j is the target gender effect for the jth design
- γ_j is the target emotion effect for the jth design
- θ_j is the gender-by-emotion interaction effect for the jth design

and

- x_{1jl} is the indicator for target gender ($l = 1, 2$, for female and male targets, respectively)
- x_{2jm} is the indicator for target emotion ($m = 1, 2$, for sadness/neutral and anger expressions, respectively)
- x_{3jlm} is the indicator for the gender-by-emotion interaction.

1. Null Model: angry men, not-angry men, angry women and not-angry women are all accorded equal status, and there are no differences in status conferral between designs.

   From the general model, the null-model is adjusted as follows:

$$Y_{ijklm} \sim N(\alpha_i, \sigma^2).$$

   where α_i is the baseline status conferral rating for ith lab.

2. Designs Baseline Model: angry men, not-angry men, angry women and not-angry women are all accorded equal status, but status conferral differs between designs.

   From the general model, the baseline-model is adjusted as follows:

$$Y_{ijklm} \sim N(\alpha_i + \eta_j, \sigma^2).$$

3. Gender Stereotyping: while angry men are accorded higher status than not-angry men, angry women are accorded lower status than not-angry women.

$$Y_{ijklm} \sim N(\alpha_i + \eta_j + x_{1jl}\beta_j + x_{2jm}\gamma_j + x_{3jlm}\theta_j, \sigma^2).$$

Effect coding is used to quantify the different conditions. The ordinal constraints based on the theoretical perspective are put on the cell means, rather than on the parameters. Specifically, the cell mean of the angry men condition needs to be higher than the not-angry

men condition and the cell mean of the not-angry women condition needs to be higher than the cell mean of the angry women condition. Cell means are calculated based using the estimated parameters; for the not-angry women condition, for instance, this results in the following: $Y_{\cdot\cdot 11} = -1/2\beta - 1/2\gamma + 1/2\theta$.

To satisfy the theoretical predictions, the following inequality constraints have to hold:

$$Y_{\cdot\cdot 11} > Y_{\cdot\cdot 12}$$

$$Y_{\cdot\cdot 22} > Y_{\cdot\cdot 21}$$

4. Status Signaling: angry men are accorded higher status than not-angry men and angry women are accorded higher status than not-angry women.

The status signaling perspective applies the same model and parameters as the gender stereotyping model. Here, the cell mean of the angry men condition needs to be higher than the not-angry men condition and the cell mean of the angry women condition needs to be higher than the cell mean of the not-angry women condition. To meet this condition, the following inequality constraints have to hold:

$$Y_{\cdot\cdot 12} > Y_{\cdot\cdot 11}$$

$$Y_{\cdot\cdot 22} > Y_{\cdot\cdot 21}$$

5. Culture Change/ Study Savviness: angry women are accorded more status than not-angry women, angry men and not-angry men. There is no effect of emotion for men.

$$Y_{ijkl} \sim N(\alpha_i + \eta_j + x_{4kl}\delta + x_{5kl}\eta, \sigma^2),$$

where parameter δ is the effect of emotion for female targets, parameter η is the effect of angry women versus men (both not-angry and angry). The indicator variables are x_{4lm} (-1/2 for not-angry female targets and 1/2 for angry female targets, and 0 for

men) and x_{5lm} (-1/3 for men, 2/3 for angry women and 0 for not-angry women).

The theoretical perspective entails that the cell mean of the angry women condition needs to be higher than the not-angry women condition and higher than the cell means of the angry men and the not-angry men conditions. To meet this condition, we put inequality constraints on the parameters:

$$\delta > 0$$

$$\eta > 0$$

6. Anger suppresses warmth: not-angry men are considered warmer than angry men and not-angry women are considered warmer than angry women.

The anger suppresses warmth model applies the same model and parameters as the status signaling model, but constrains the emotion parameters in exactly the opposite direction:

$$Y_{..12} < Y_{..11}$$

$$Y_{..22} < Y_{..21}$$

7. Cultural Differences: in high confrontation-oriented cultures, angry men and women are accorded more status than not-angry men and women, but this difference decreases or flips in harmony-oriented cultures. This model builds on the model for perspectives 2 and 3 and is extended by parameters for the main effect and interactions of culture (as a standardized continuous score):

$$Y_{ijklm} \sim N(\alpha_i + \eta_j + x_{1jl}\beta_j + x_{2jm}\gamma_j + x_{4i}\zeta + x_{5im}\upsilon, \sigma^2,$$

where we include two new parameters:

- ζ is the culture dimension main effect

- υ is the culture-by-emotion interaction effect

and

- x_{4i} is the (standardized) culture dimension score of the ith lab
- x_{5im} is the (standardized) culture dimension score of the ith lab $\times$ the indicator for target emotion ($-0.5, 0.5$ for $m = 1, 2$, for sadness/neutral and anger expressions, respectively).

The ordinal constraints are put on the culture-by-emotion parameter (υ). Specifically, the positive effect of anger vs. sadness/neutral expressions should become smaller or negative with increased cultural harmony levels. To test this prediction, the interaction coefficient needs to be negative:

$$\upsilon < 0$$

8. Unconstrained model: all effects are included, without any ordinal constraints (this is again the general model).

$$Y_{ijklm} \sim N(\alpha_i + \eta_j + x_{1jl}\beta_j + x_{2jm}\gamma_j + x_{3jlm}\theta_j, \sigma^2),$$

Note that in contrast to Study 1, we do not include the culture dimension in the unconstrained model, as culture is not relevant for the primary theoretical tests.

Additional Figures

Figure 11. Manipulation checks.

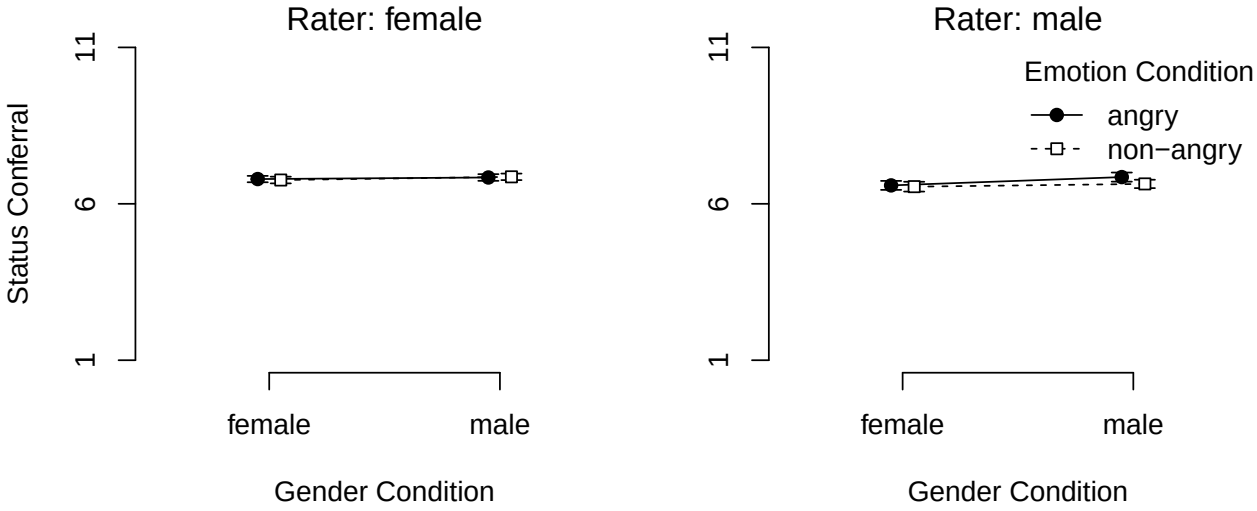


Figure 12. Status conferral per target gender, emotion, and rater gender.

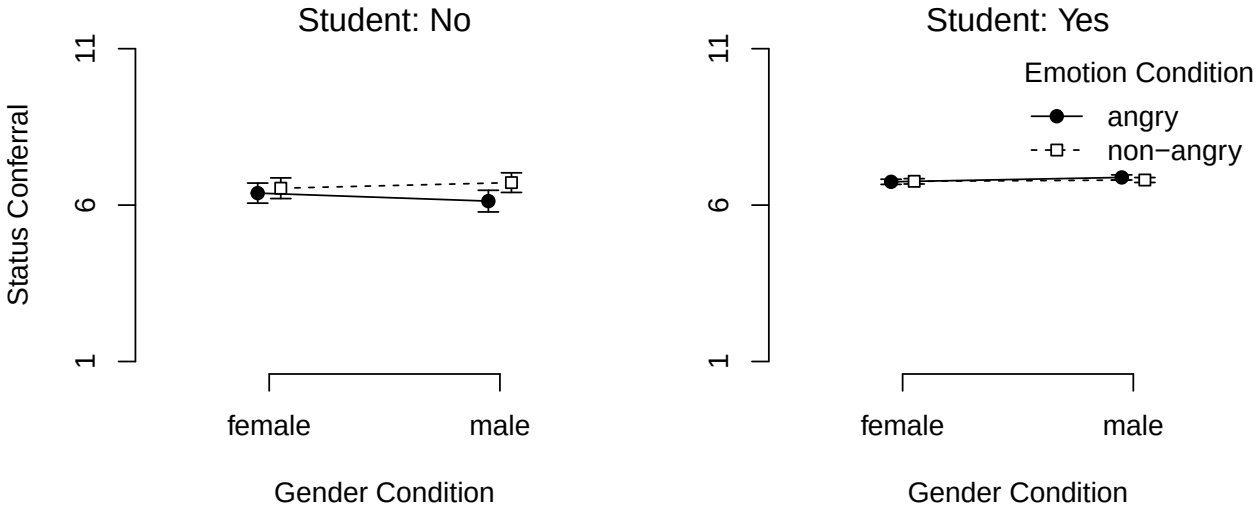


Figure 13. Status conferral per target gender, emotion, and student status.

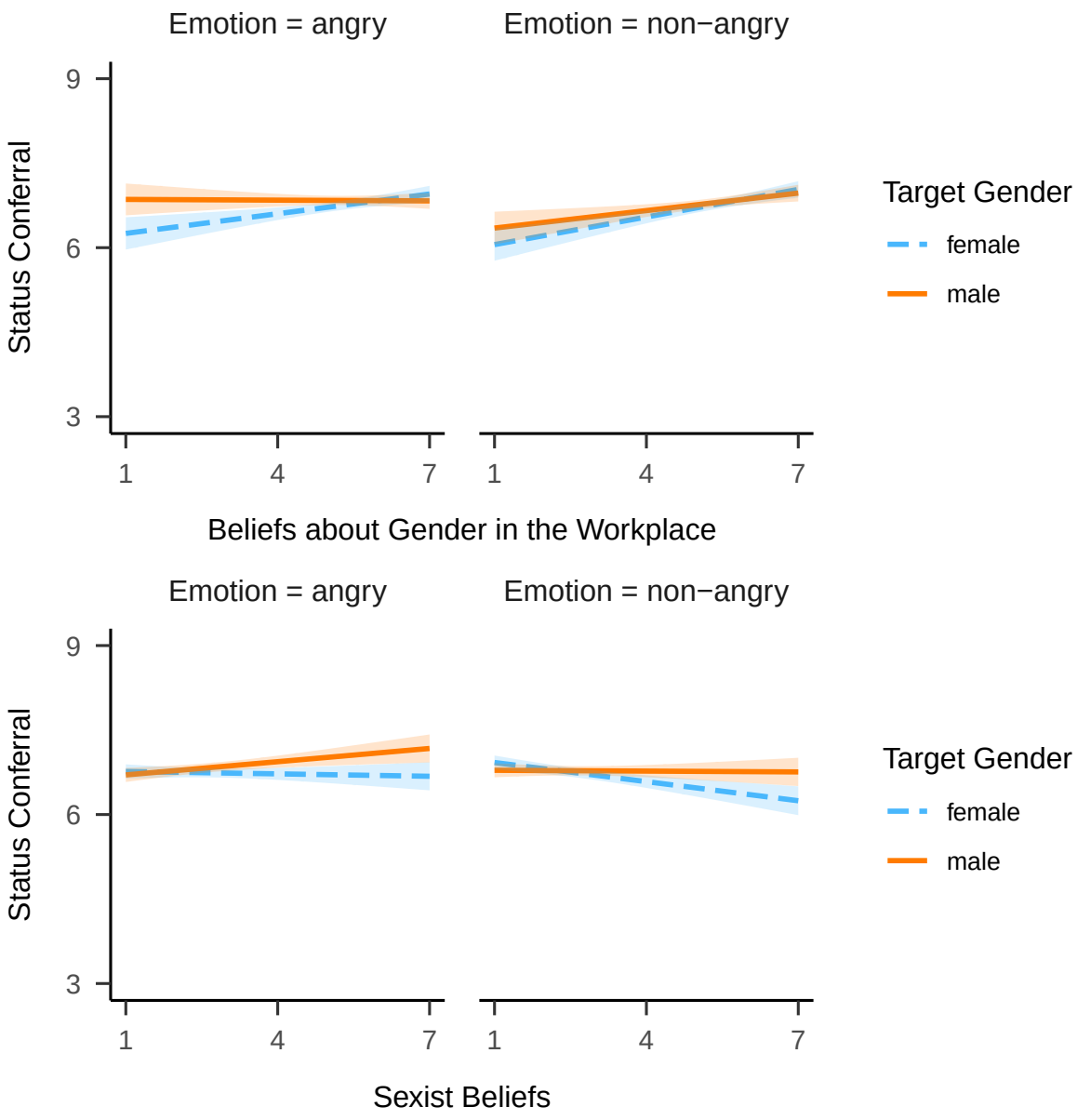


Figure 14. Moderation effects of beliefs about gender equality on status conferral.

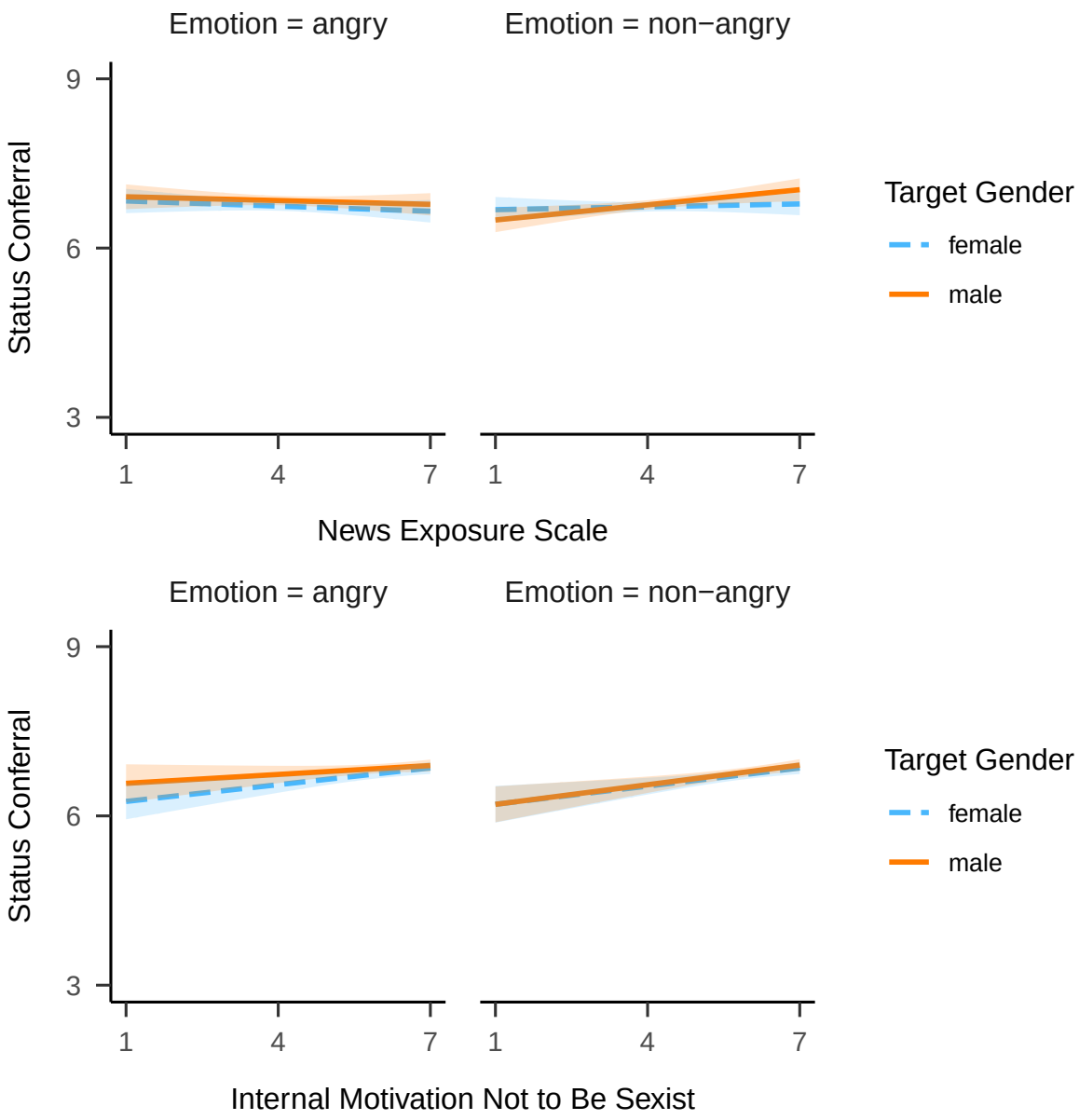


Figure 15. Moderation effects of beliefs about gender equality on status conferral.

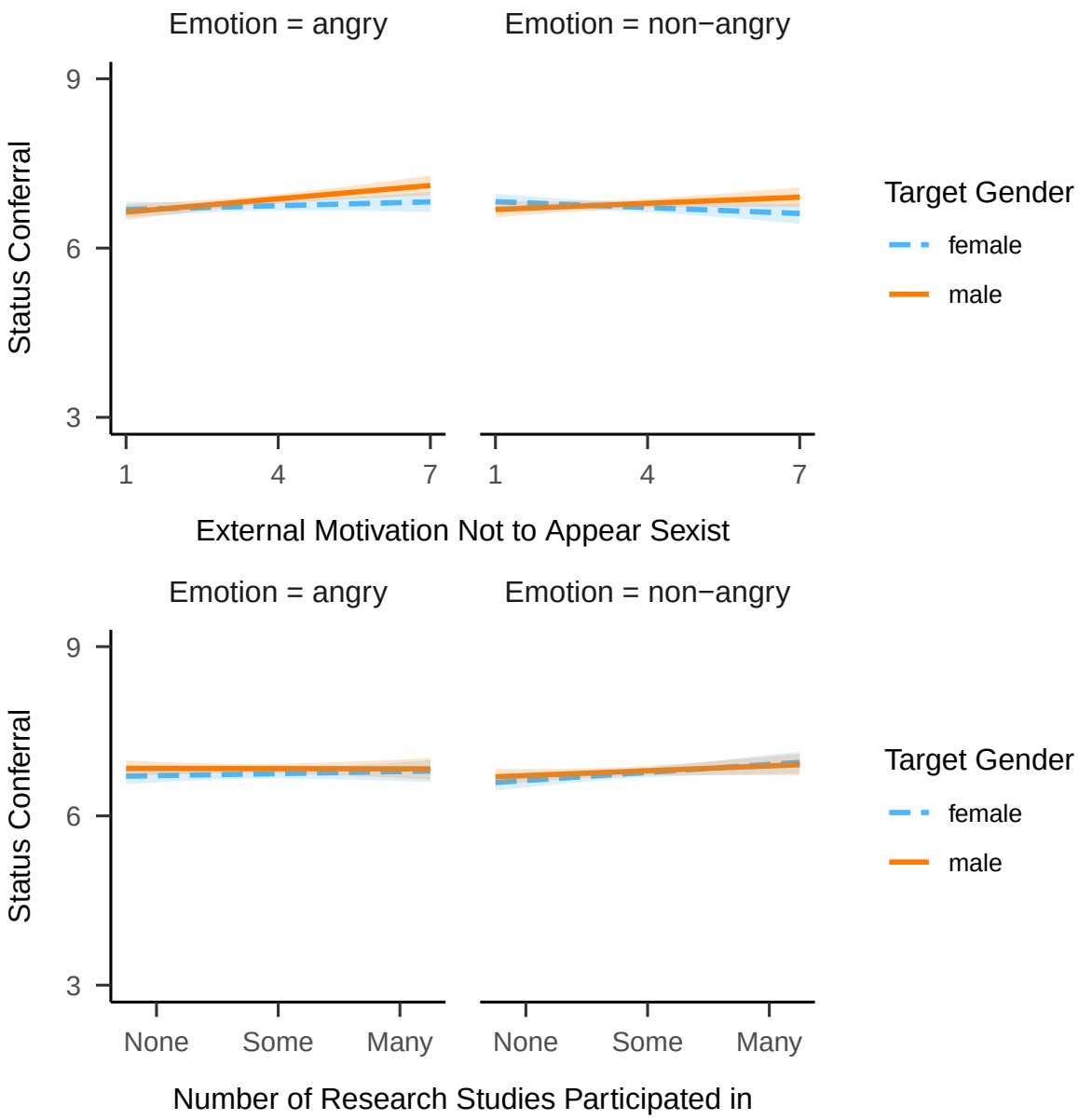


Figure 16. Moderation effects of self-presentation goals on status conferral.

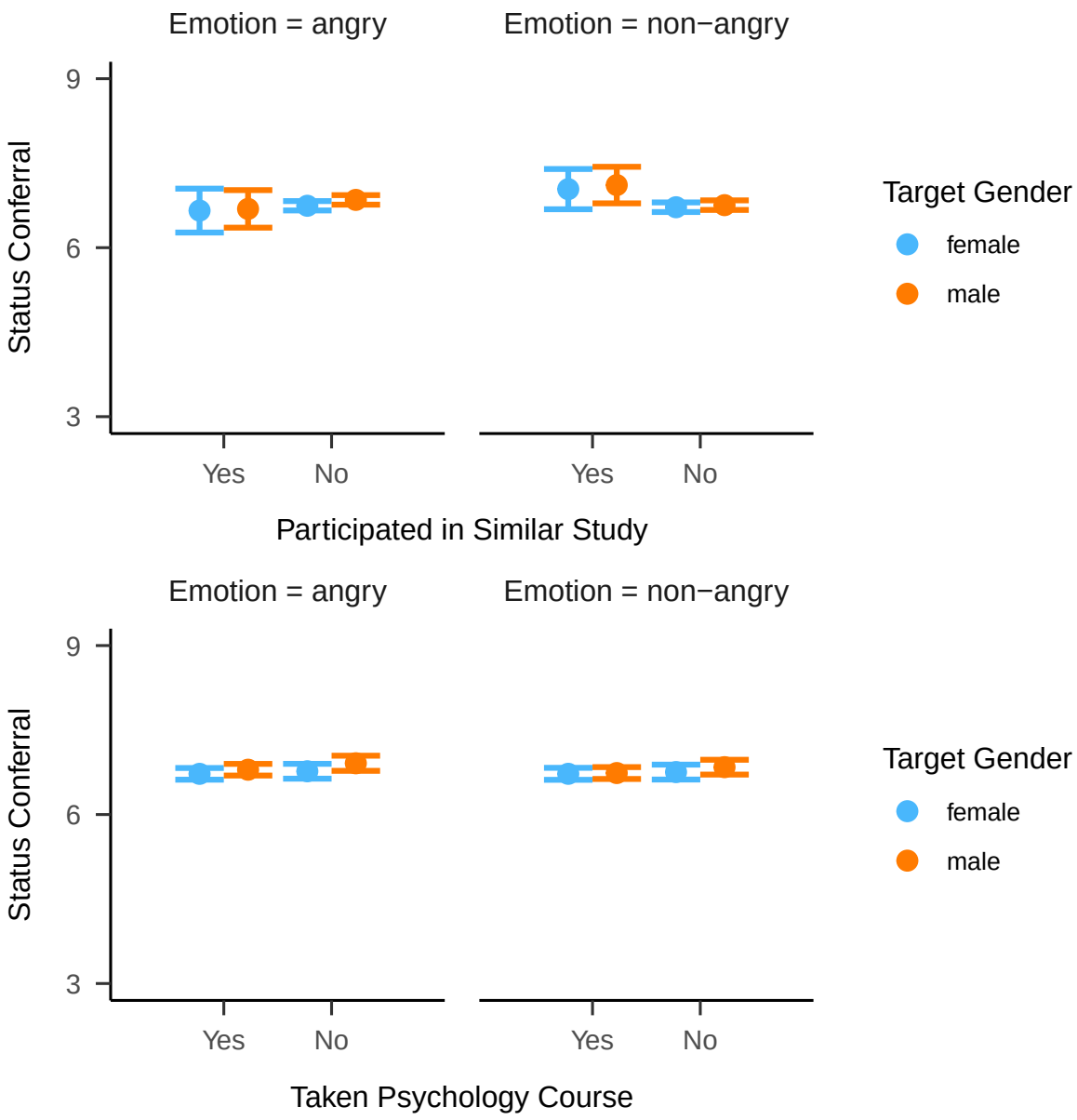


Figure 17. Moderation effects of self-presentation goals on status conferral.

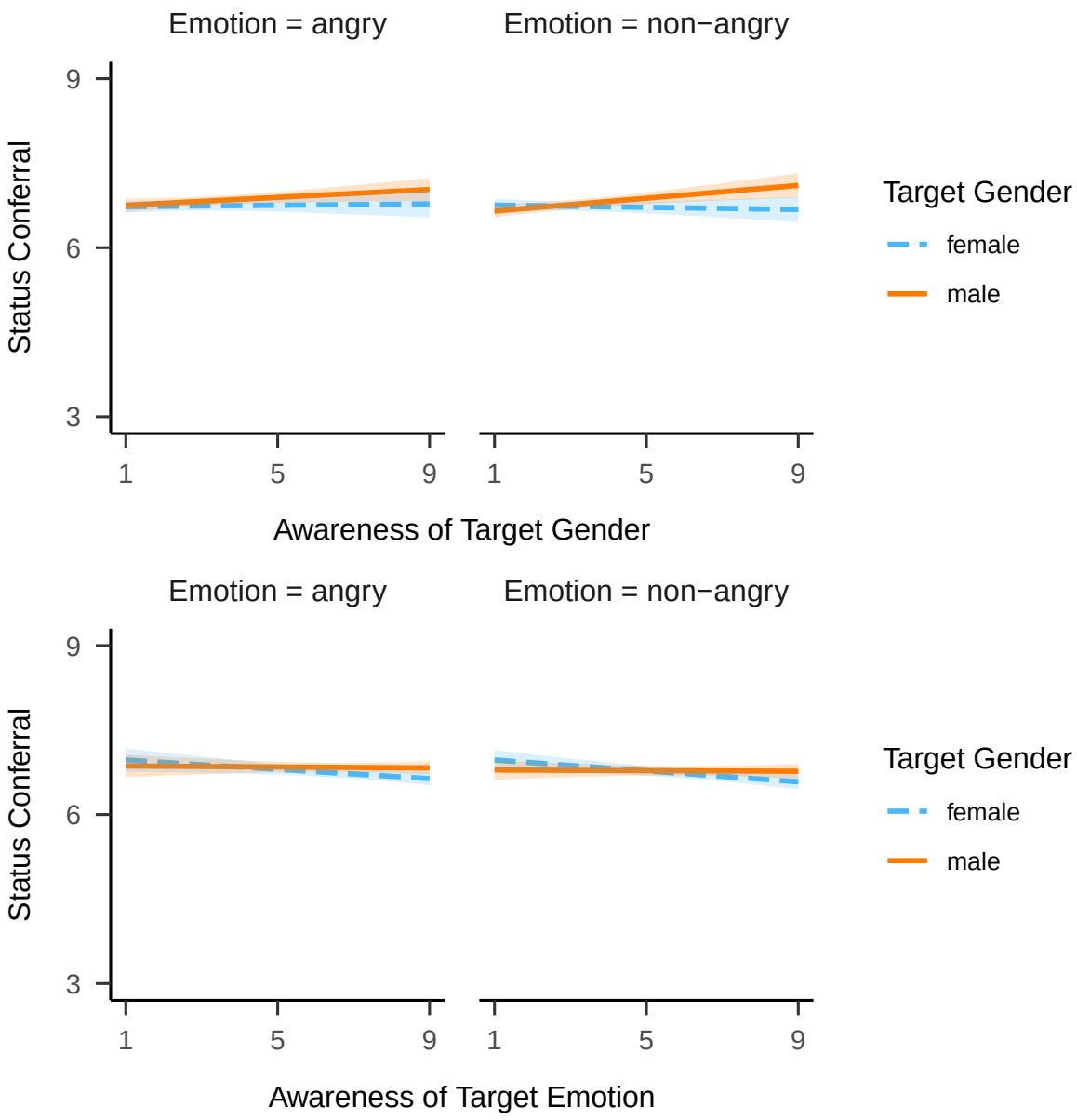


Figure 18. Moderation effects of self-presentation goals on status conferral.