The Role of Warmth and Competence Perceptions in Labor Market Outcomes

Carina Hausladen Marcos Gallo

June 25, 2023

Table of contents

1 Introduction

2 Data Structure

The PRISMA Flow Diagram, which provides a comprehensive overview of our data selection process, can be accessed at the following link.

	category	studies	levels	signals
	race	8	3	261
	gender	8	2	261
	secuality	4	5	5
	health	2	4	4
	parenthood	2	2	2
	age	2	9	9
	wealth	2	2	2
	nationality	1	2	2
	employment	1	3	3
	religion	1	8	8
	nationality	1	35	35
total	—	32.00	75.00	592.00

There is a considerable body of research on correspondence studies that specifically examines the categories of race and gender. In these studies, one of the most frequently used signals is a person's name. Because of this, the number of signals used in studies on race and gender is much larger compared to other categories, due to the higher number of levels within these subsets. However, correspondence studies also exist that focus on categories other than race and gender. In these studies, the number of signals used is equivalent to the number of levels. Nevertheless, given that there are more observations on race and gender categories, we will first analyze a dataset that includes all categories before delving into an in-depth analysis of race and gender.

2.1 Callback

The main variable of interest is the callback proportion. To quantify the callback rate, we computed the risk ratio, $p_{callback} = \frac{callback}{opportunities}$. When p is close to 0 or close to 1, the standard error is artificially compressed, which leads us to overestimate the precision of the

proportion estimate. To avoid this, the proportions are logit-transformed before they are pooled.

3 Social perception predicts callback in correspon- dence studies that vary *names*

3.1 Callback

In our sample, the callback ratio is $\theta = 0.79$ for Black names, which was significantly less than one (p = 0.07).

For the female gender compared to male, our estimated ratios are $\theta = 1.02$ (p = 0.07) in the eight studies we have. Together the data show a $20 - 30\%$ reduction in callbacks for Black names and no reduction for female names (Table S6).

3.2 warmth and competence

To measure warmth and competence, lists of names from the correspondence studies were given to participants on Prolific (787 raters total, 85.89 per name).

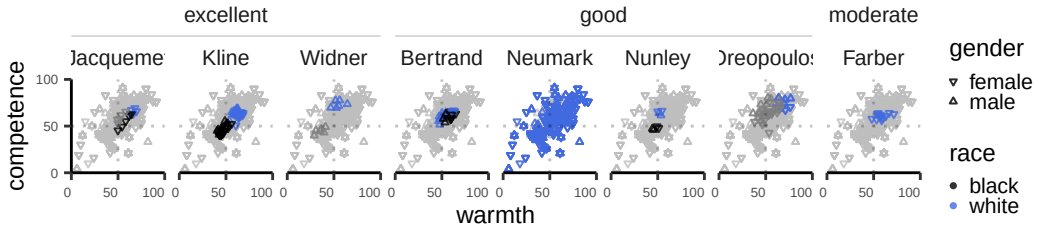
Table S2. ICC values for names

	Warmth		Competence		Mean	
	warm_ICC	warm_score	comp_ICC	comp_score	avg_ICC	avg_score
jacquemet	0.98	excellent	0.97	excellent	0.97	excellent
kline	0.95	excellent	0.96	excellent	0.95	excellent
widner	0.97	excellent	0.99	excellent	0.98	excellent
bertrand	0.83	good	0.69	moderate	0.76	good
neumark	0.94	excellent	0.81	good	0.87	good
nunley	0.65	moderate	0.93	excellent	0.79	good
oreopoulos	0.91	excellent	0.89	good	0.90	good
farber	0.86	good	0.50	poor	0.68	moderate

Average score intraclass correlations (ICCs) were used as an index of interrater reliability of warmth competence ratings. A twoway model with random effects for raters and subjects (amount of levels in category) was used. Between rater agreement was estimated. The unit of analysis was averages.

To evaluate the consistency of ratings across categories, we computed the intraclass correlation. Our results reveal that the level of agreement between raters differs across various studies, with agreement ranging from excellent to good in most studies (Figure 2A, Table S2).

3.2.1 Figure 2A



The callback rates were computed by averaging the deci- sions of multiple hiring managers. Meanwhile, the warmth and competence scores were obtained from a different sample. To ensure reliable social perception measurements, we specif- ically recruited participants residing in North America with demographics closely resembling those of the average hiring manager, and we averaged ratings across raters. This enabled us to confidently match the social perception ratings with the callback rates per name. Those warmth and competence ratings, across names in different studies, are shown in Figure 2A.

3.2.2 meta models

		95% CI			
	estimate	lower	upper	p-value	SE
competence[1]					
black_c[2]	-11.52	-23.74	0.71	0.06	3.84

female_c[3]	-3.07	-9.56	3.42	0.32	2.91
warmth[1]					
black_w[2]	-6.72	-19.19	5.76	0.19	3.92
female_w[3]	2.88	-4.39	10.16	0.40	3.27
callback[4]					
black[5]	0.79	-0.51	0.04	0.07	0.09
female[6]	1.02	-0.03	0.06	0.36	0.01

[1]The meta-analytical involves the inverse variance method and a restricted maximum-likelihood estimator for tau^2. The Q-Profile method was used to compute the confidence interval of tau^2 and tau, and a Hartung-Knapp (HK) adjustment was applied for the random effects model, with degrees of freedom set to 10.
[2]k=4 studies, o=687 observations.
[3]k=11 studies, o=816 observations.
[4]The Mantel-Haenszel method was used to calculate the overall effect size, with the Paule-Mandel estimator used to estimate the between-study variance (tau^2). A random-effects model was employed with the Hartung-Knapp (HK) adjustment to account for potential bias due to small sample sizes. The model had 1 degree of freedom (df = 1).
[5]k=4 studies, o=89872 observations.
[6]k=4 studies, o=143860 observations.

There were only minor differences in warmth or competence ratings between black and white candidates or males and females (between 2 and 7 points on the 100-point scale), except for a marginally significant difference in competence between black and white (θ = -6.72, p = 0.19, Table S6).

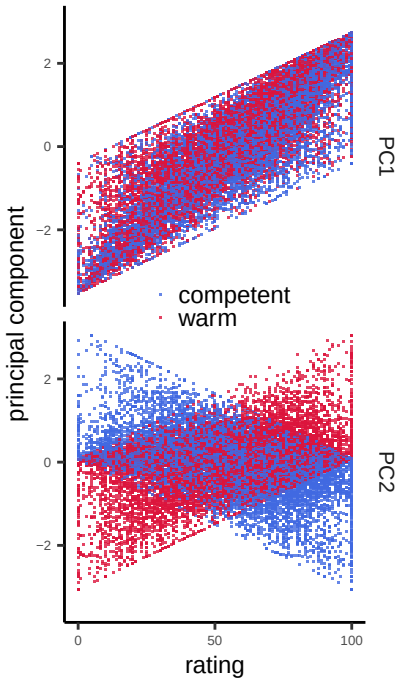
Table S4. Results of a random effects model with p(warmth, competence) for names

		95% CI			
	correlation	lower	upper	p-value	SE
by study					
bertrand	0.616	0.378	1.060	0.000	0.174
farber	0.408	-0.220	1.086	0.194	0.333
flake_leasure	0.900	1.241	1.700	0.000	0.117
gorzig	0.826	0.585	1.767	0.000	0.302
jacquemet	0.845	0.715	1.763	0.000	0.267
kline	0.900	1.241	1.700	0.000	0.117
neumark	0.631	0.565	0.923	0.000	0.091
nunley	0.740	0.073	1.826	0.034	0.447
oreopoulos	0.565	0.334	0.946	0.000	0.156

widner	0.915	0.904	2.210	0.000	0.333
pooled	0.780	0.759	1.330	0.000	0.126

Meta-analysis of k = 10 studies with o = 418 observations using inverse variance method. Random effects model with restricted maximum-likelihood estimator for tau^2 and Hartung-Knapp adjustment (df = 8). Confidence intervals for tau^2 and tau estimated using Q-Profile method. Fisher's z transformation used for correlations.

Figure 2A shows strong, reliable positive associations between warmth and competence within all eight studies, ranging from 0.41 – 0.91 (Table S4). The pooled correlation is $\hat{\rho}_{w,c} = 0.78$ (p = 0).



We, therefore, used principal component analysis (PCA) as a proxy for social perceptions. Figure 2D shows how the principal component scores (y-axis) are related to warmth and competence ratings (x-axis). The first component (PC1) reflects the positive association; it explains 79.3% of the variance. PC2 is high when only one rating is high. It accounts for only -78.3% of the total variance, indicating its less prominent role in the overall data structure. As a result, our subsequent analyses will focus on the PCs rather than the original ratings that generated them.

3.3 Correlation between callback and PC1 as an effect

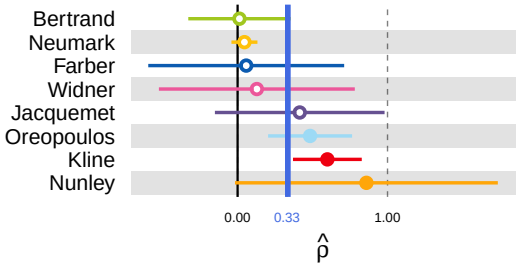


Figure 2C is a forest plot of the estimated correlations $\rho(callback \times PC1)$ and 95% confidence intervals of the eight studies. The effects across studies were pooled via a meta-analytic random effects model. The pooled correlation between the callback percentage and PC1 is $\hat{\rho} = 0.34$ (p = 0.03), indicating a moderate correlation.

To interpret the pooled effect size meaningfully, we must consider the variance of the true effect sizes distribution, τ^2, and the between-study heterogeneity, I^2.

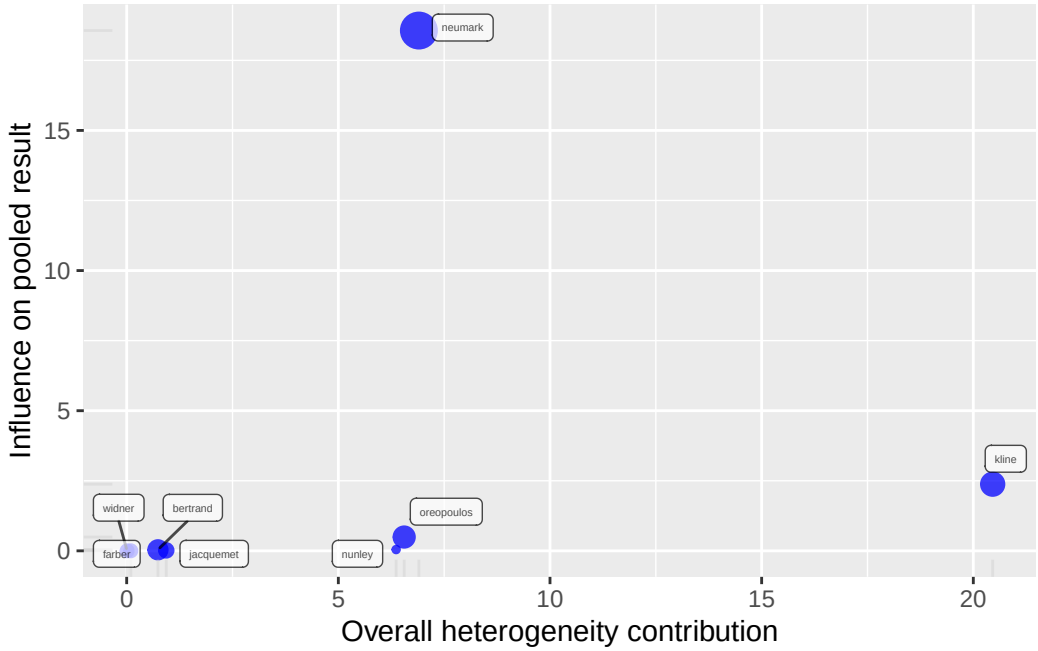
As suggested by Figure 2C, there is "substantial heterogeneity" among studies: 83 percent of the variation in effect sizes is due to between-study heterogeneity (I2 = 0.83, 95% CI [0.69 − 0.91]). Furthermore, the variance of the true effect sizes distribution is significantly greater than zero ($\tau^2 = 0.08$, 95% CI [0.03 − 0.66]).

Given the large level of heterogeneity in our analysis, we find a wide prediction interval (from -0.42 to 1.12, suggesting that future studies could potentially reveal negative correlations. Therefore, caution is warranted in interpreting the results, and further research is needed to clarify the effect of social perception on callback.

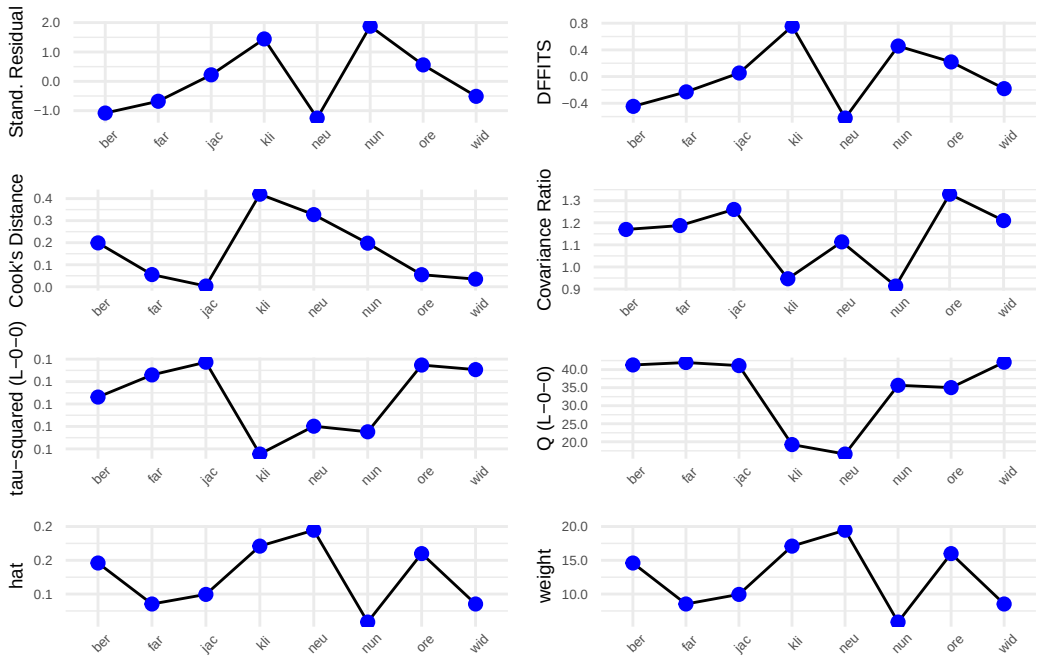
3.3.1 Between-Study Heterogeneity and Publication Bias

```
[================================================================================] DONE


GOSH Diagnostics
================================

 - Number of K-means clusters detected: 3
 - Number of DBSCAN clusters detected: 11
 - Number of GMM clusters detected: 4

 Identification of potential outliers
 --------------------------------
```

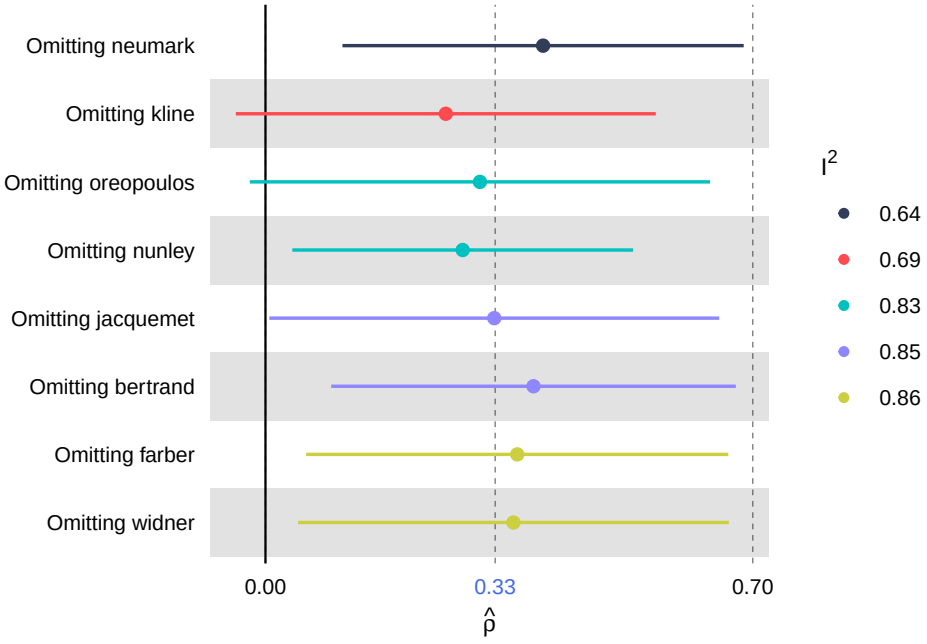
```
 - K-means: No outliers detected.
 - DBSCAN: Study 4, Study 5, Study 6, Study 3
 - Gaussian Mixture Model: Study 4, Study 5, Study 6, Study 3

[1] "jacquemet" "kline"     "neumark"   "nunley"
```



The Baujat plot is a diagnostic plot used to identify studies that disproportionately contribute to heterogeneity in a meta-analysis. The plot displays the contribution of each study to the overall heterogeneity (measured by Cochran's Q) on the x-axis and its impact on the pooled effect size (defined as the standardized squared difference between the overall estimate based on an equal-effects model with and without the ith study included in the model) on the y-axis. Our analysis found that Kline significantly influenced the heterogeneity but did not significantly affect the pooled effect size. On the other hand, Neumark contributed moderately to the overall heterogeneity but substantially impacted the pooled effect size.

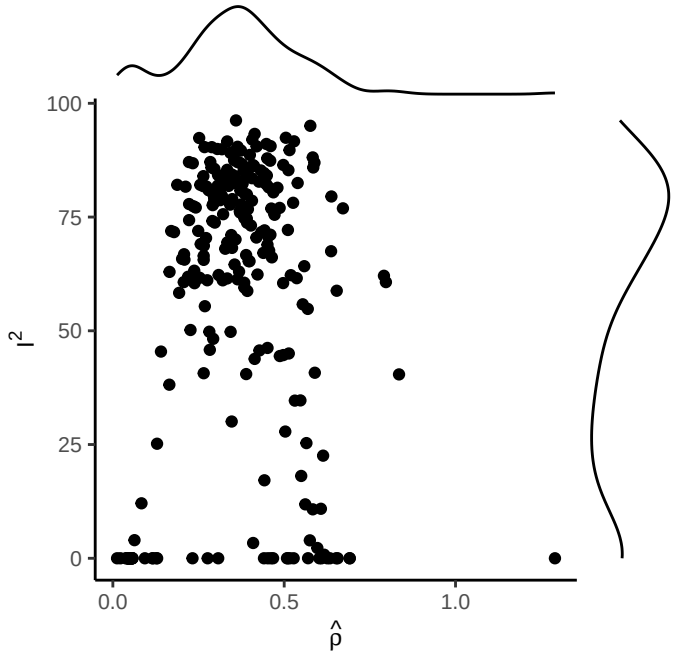


The plot displays different influence measures for each study, which help to identify potential outliers that do not fit well into the meta-analysis model. No study was detected as an outlier based on these measures.



The plot displays the overall effect and I2 heterogeneity of all meta-analyses with (callback,

PC1) as effectsize that were conducted using the leave-one-out method. The forest plot is sorted by the I2 value of the leave-one-out meta-analyses. The results show that excluding neumark leads to the largest reduction in I2, reducing it from 82% to 64%.



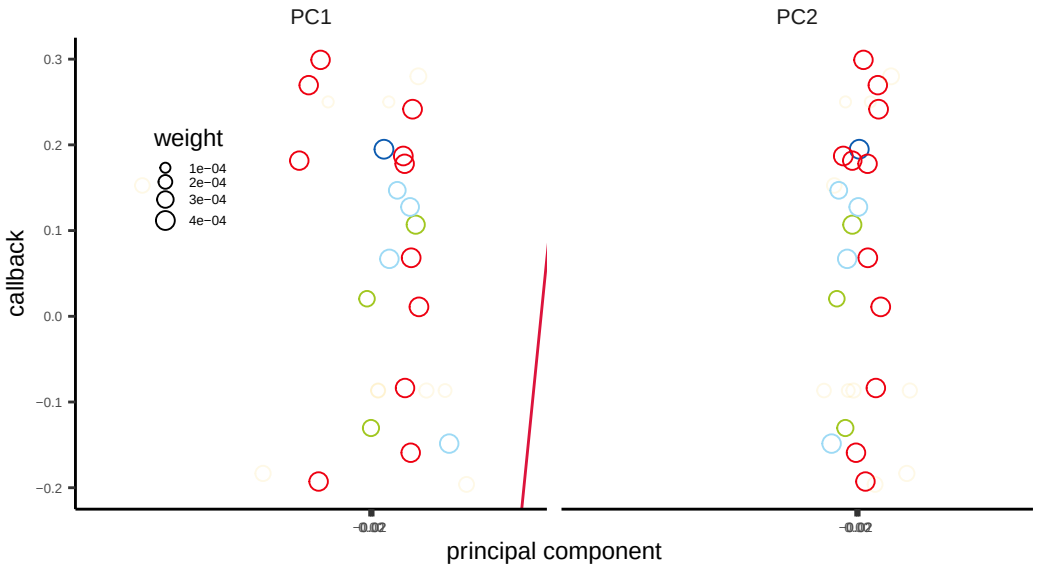
We implemented a Graphical display of heterogeneity (GOSH) plot analysis. For this analysis, we fit all possible subsets 2k−1 of our k included studies. Each subset's pool effect size $\hat{\rho}$ is plotted on the x-axis, and the between-study heterogeneity I2 on the y-axis. Three (k-means, DBSCAN, gmm) clustering algorithms are used to determine patterns in the above scatter plot. The three algorithms did not consistently identify clusters therefore, we conclude that based on this analysis, no single study needs to be excluded from estimating the meta-model.

In order to ensure the robustness of our findings and account for potential outliers, we conducted a comprehensive outlier and heterogeneity analysis. Overall, only two out of eight tests identified outliers.

When excluded, re-estimate $\hat{\rho}$ as 0.4 or 0.36 , both values remaining close to the 0.34 all-study estimate in Figure 2C.

Furthermore, Egger's regression test (Fig. S6; intercept = 1.86, 95% CI [-0.44, 4.16], t = 1.58, p = 0.16) did not indicate bias.

3.4 Meta Regression

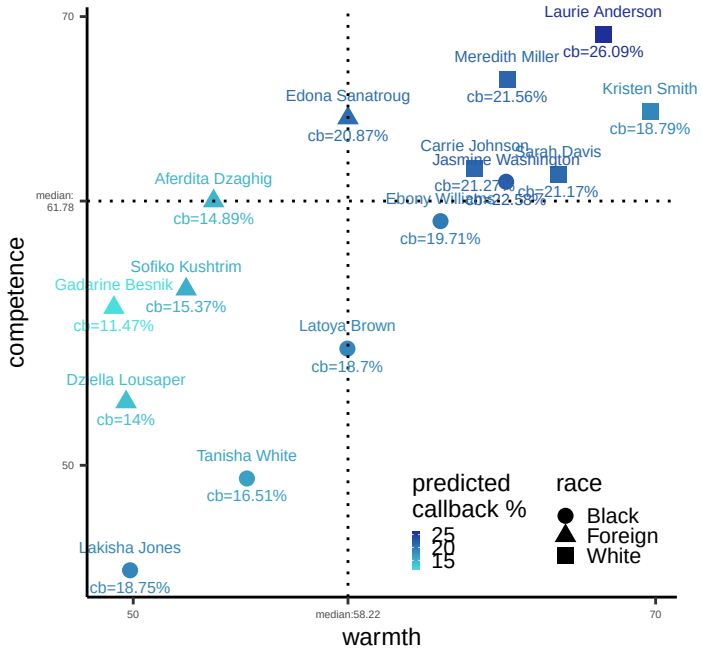


		95% CI			
	estimate	lower	upper	p-value	SE
Meta Regression for Categories[1]					
intercept_c	-0.32	-1.06	0.43	0.06	0.37
b_PC1_c	1.16	-0.28	2.59	0.13	0.72
b_PC2_c	-0.62	-3.58	2.35	0.69	1.49
Meta Regression for Names[2]					
intercept	-1.97	-2.47	-1.48	0.00	0.25
b_PC1	1.00	0.41	1.58	0.00	0.30
b_PC2	0.56	-0.83	1.96	0.43	0.71
Correlations rho(callback times variable) for names[3]					
PC1	0.34	0.03	0.66	0.03	0.13
warmth	0.34	0.08	0.64	0.02	0.12
comp	0.26	-0.06	0.58	0.09	0.14

[1]Mixed-Effects Model (k = 79; tau^2 estimator: ML)
[2]Mixed-Effects Model (k = 691; tau^2 estimator: ML)
[3]Number of studies combined: k=8; Number of observations: o=725; The meta-analytical models employed inverse variance method, restricted maximum-likelihood estimator for tau^2, Q-Profile method for confidence interval of tau^2 and tau, Hartung-Knapp adjustment for random effects model (df = 7), prediction interval based on t-distribution (df = 6), and Fisher's z transformation of correlations.

As an alternative specification to the meta-analysis with $\hat{\rho}$, Table 1 reports results of a mixed effects model of callback rates against the PCs and raw ratings. The results are visualized in Figure 2B. The coefficient for PC1 is positive $\hat{\beta}_{PC1} = 1$ and highly significant (p = 0).

The correlations for warmth ($\hat{\rho} = 0.34$, p = 0.02) and competence ($\hat{\rho} = 0.26$, p = 0.09) are similar to those observed for the first principal component (PC1).



Moreover, we tested the predictive potential of our model for names. We found that $\hat{\rho}$(callback, PC1) in (21) is closest to the pooled effect, and we, therefore, used data from this study to make predictions. Specifically, we trained a linear model on all names except one and then used this model to predict the callback for the left-out name. Fig. S7 visually represents our findings, with lighter shades of blue indicating lower callback rates. Our analysis reveals that callback rates are highest in the upper quadrant of the warmth and competence scale, while the callback rates are lowest in the lowest quadrant of this plot. Interestingly, we also observe clusters of the race that the names would signal (Black, White, or foreign-sounding).

3.5 Exploratory analysis points to differential effects of social perceptions across job types

4 Data: Categories

Prolific participants (200 raters total, 99.11 by level) rated each signal on a scale from 0 to 100 within a category (e.g., how warm/competent they think a "treasurer in gay and lesbian alliance" would be, Figure 1, and Figure 3A).

		Warmth		Competence		Mean	
	category	warm_ICC	warm_score	comp_ICC	comp_score	avg_ICC	avg_score
ameri	health	0.96	excellent	0.93	excellent	0.95	excellent
hipes	health	0.96	excellent	0.97	excellent	0.97	excellent
ishizuka	parenthood	0.97	excellent	0.92	excellent	0.95	excellent
namingit	unemployed	0.96	excellent	0.98	excellent	0.97	excellent
wrigth	religion	0.95	excellent	0.94	excellent	0.95	excellent
yemane	nationality	0.91	excellent	0.93	excellent	0.92	excellent
mishel	sexuality	0.83	good	0.94	excellent	0.89	good
farber	age	0.55	moderate	0.79	good	0.67	moderate
rivera	wealth	0.82	good	0.63	moderate	0.72	moderate
tilcsik	sexuality	0.64	moderate	0.51	moderate	0.58	moderate
bailey	sexuality	0.04	poor	0.95	excellent	0.50	poor
correll	parenthood	0.87	good	0.00	poor	0.43	poor
figinski	military	0.00	poor	0.00	poor	0.00	poor
kline	sexuality	0.00	poor	0.94	excellent	0.47	poor
neumark	age	0.08	poor	0.33	poor	0.21	poor
thomas	wealth	0.00	poor	0.97	excellent	0.49	poor

Average score intraclass correlations (ICCs) were used as an index of interrater reliability of warmth competence ratings. A twoway model with random effects for raters and subjects (amount of levels in category) was used. Between rater agreement was estimated. The unit of analysis was averages.

The intraclass correlation (ICC) (26) values vary across categories. Only two categories scored "poor", while the remaining scored either "moderate" or "excellent" (Figure 3A). Note that, for sexuality and wealth, the different signals yielded vastly different ICCs, ranging from 0 to 0.83 (Table S3).

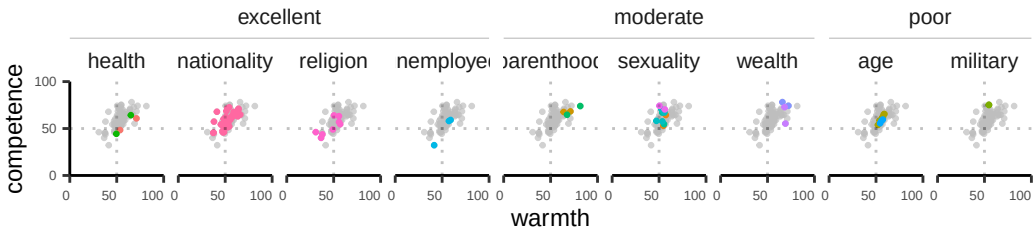


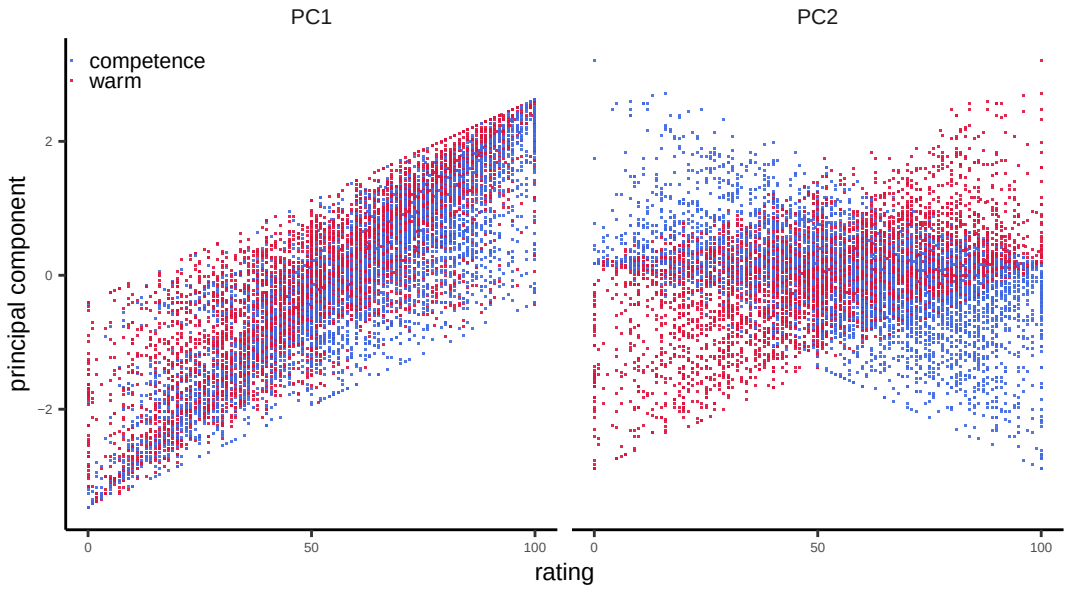
Figure 4.1: This graph displays the mean ratings of warmth and competence for each level, with the size of the dots reflecting the callback rate. The green dot represents the pooled effect of the ratings.

Table S5. Results of a random effects model with (warmth, competence) for categories

		95% CI			
	correlation	lower	upper	p-value	SE
by study					
ameri	0.574	0.514	0.794	0	0.072
bailey	0.546	0.474	0.752	0	0.071
correll	0.645	0.626	0.907	0	0.072
farber	0.602	0.616	0.778	0	0.041
figinski	0.342	0.216	0.496	0	0.072
hipes	0.699	0.724	1.005	0	0.072
ishizuka	0.710	0.749	1.026	0	0.071
kline	0.416	0.304	0.582	0	0.071
mishel	0.456	0.394	0.591	0	0.050
namingit	0.776	0.922	1.149	0	0.058
neumark	0.724	0.802	1.030	0	0.058
rivera	0.583	0.526	0.808	0	0.072
thomas	0.425	0.313	0.594	0	0.072
tilcsik	0.437	0.327	0.609	0	0.072
wrigth	0.696	0.789	0.929	0	0.036
yemane	0.639	0.724	0.791	0	0.017
pooled					
	0.595	0.579	0.792	0	0.050

Meta-analysis of k = 16 studies with o = 7830 observations using inverse variance method. Random effects model with restricted maximum-likelihood estimator for tau^2 and Hartung-Knapp adjustment (df = 15). Confidence intervals for tau^2 and tau estimated using Q-Profile method. Fisher's z transformation used for correlations.

Furthermore, we assessed the correlation between the warmth and competence ratings and found that the Pearson correlation index (ρ) was significant for most studies, with a pooled correlation of $\hat{\rho} = 0.595$ (p = 0, Table S5).



To better capture the variability in social perception of different social categories, we conducted a PCA, revealing two principle components. PC1 explained 80.73% of the variability in warmth and com- petence ratings, combining the positively correlated measures onto a single dimension. PC2 represented negative associations and accounted for -79.73% of variance.

4.1 Relating ratings to callback

```
Running 1000 iterations for an approximate permutation test.
```

In the following, we conduct a meta-regression pooling all studies to explore the extent to which the principal components predict callback. To increase the robustness of this analysis, we also perform a permutation test on our meta-regression models. The resulting estimate for the coefficient of the first principal component is 1.16, which is not statistically significant (p = 0.13). Our model explains a small portion of the heterogeneity, accounting for only 3.32% (Table 1). Figure 3D visualizes the meta-regression model.

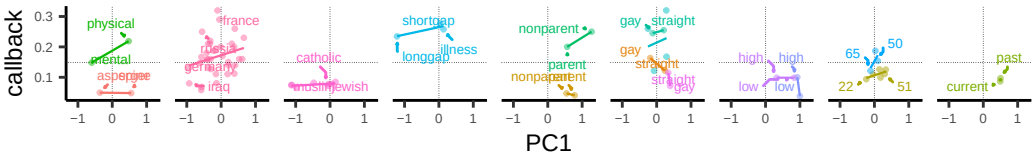


Figure 4.2: Each panel depicts distinct categories, with the first principal component plotted against callback rates. Colors represent independent studies.

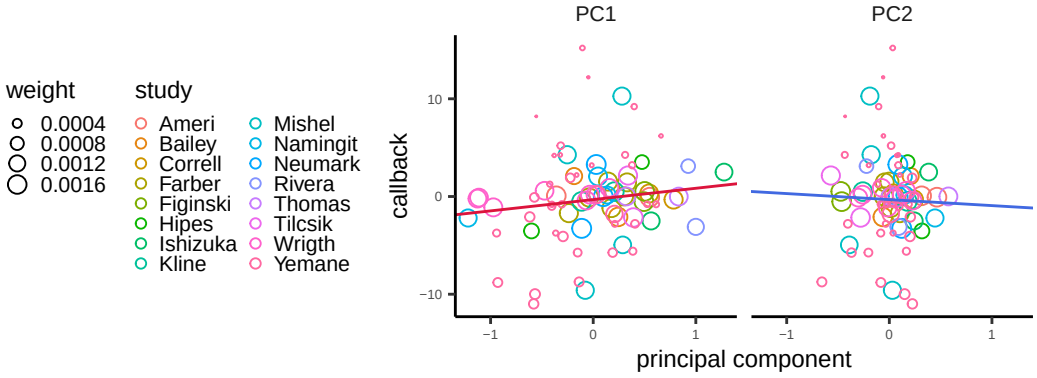


Table S7. Estimates of linear models of PC1 on callback by category

	estimate	std.error	statistic	p.value
wealth				
rivera	-0.84	NaN	NaN	NaN
thomas	0.00	NaN	NaN	NaN
unemployed				
namingit	0.02	0.01	1.74	0.33
sexuality				
bailey	-0.10	NaN	NaN	NaN
kline	0.03	NaN	NaN	NaN
mishel	0.05	0.23	0.22	0.85
tilcsik	-0.82	NaN	NaN	NaN
parenthood				
correll	-0.03	NaN	NaN	NaN
ishizuka	0.07	NaN	NaN	NaN
nationality				
yemane	0.04	0.02	1.52	0.14
military				
figinski	1.41	NaN	NaN	NaN
health				
ameri	0.00	NaN	NaN	NaN
hipes	0.07	NaN	NaN	NaN
age				
farber	0.04	0.02	1.63	0.18
neumark	0.20	0.23	0.87	0.54
wrigth	0.01	0.00	1.68	0.14

The meta-regression did not yield a significant overall effect. Therefore, we explored the relationship between PC1 and callback by category. Given the limited number of levels across categories (Figure 3B), it was not possible to calculate a meaningful effect size directly relating ratings to callback at the study level. Thus, we present a graphical representation in Figure 3B, with lines representing fitted linear models for each study (Table S7). For some categories, the relation between callback and PC1 is positive (e.g., nationality, which also samples a larger number—35—of categories). For most other categories, however, such as wealth, sexuality, and parenthood, there are both positive and negative slopes in different studies. Under the category

of sexuality, slope signs in four studies were equally split between positive and negative, which is especially striking given the large range of ICCs across signals.