Implicit Gender Bias: Evidence from 41 Countries

Ingvild Almås Jonathan de Quidt Sebastian Fest
Anna Sandberg Trolle-Lindgren

January 19, 2024

We conduct a survey experiment on representative samples from 41 countries that seeks to measure gender discrimination by asking participants to recommend wages for hypothetical job candidates. We decompose discrimination into Explicit and Implicit components, and relate them to cross-country measures of development and gender inequality.

Data collection was completed by Gallup in March 2023 but the data were embargoed until submission of this pre-analysis plan.

1 Design

1.1 Survey question

> *Suppose the [insert country-specific national health department name] in your country hires two new workers. Their responsibilities include gathering information from hospitals on patient health outcomes and writing reports.*
>
> *Imagine you are responsible for deciding these two workers' annual salaries. Annual salaries at the [ministry] can range between [15th percentile of country income distribution] [country currency] and [97th percentile] [country currency].*
>
> *Here are some details about these two workers. Both have [a 4-year college degree or country equivalent] and have 3 years of relevant work experience. Both are married with 2 children.*
>
> *The only difference is that worker A is a 32-year-old* ***[man/woman]*** *whereas worker B is a 34-year-old* ***[man/woman]****.*
>
> *What annual salary would you give worker A? Enter value in whole [currency].*
>
> *What annual salary would you give worker B? Enter value in whole [currency].*

Text in blue varies between survey country. Text in red varies within country with four treatment variations, shown in Table 1. We target 250 observations per treatment, per country, for a total of 1,000 targeted responses per country. Responses are constrained to lie between the (rounded) 15th and 97th percentiles of the country income distribution in order to minimize outliers due to typographical errors. We obtained these percentiles from the World Inequality Database `https://wid.world/`.[1]

Using $w(\mathbf{x}, \mathbf{y})$ to denote the mean wage assigned to worker $\mathbf{x}$ when their comparator is worker $\mathbf{y}$, we observe the eight values, denoted $\{w^1, ..., w^8\}$. These can be computed at the country level or by other

[1] We use values from the most recent available year, data accessed in November 2021.

Table 1: Treatments

Treatment	Label	Text	Type of comparison
1	$(m32, m34)$	32-year-old **man**, 34-year-old **man**	Same-gender
2	$(f32, f34)$	32-year-old **woman**, 34-year-old **woman**	Same-gender
3	$(m32, f34)$	32-year-old **man**, 34-year-old **woman**	Different-gender
4	$(f32, m34)$	32-year-old **woman**, 34-year-old **man**	Different-gender

subgroups (e.g. by respondent gender).

$$w^1 = \overline{w(m32, m34)} \qquad w^5 = \overline{w(f32, f34)}$$
$$w^2 = \overline{w(m34, m32)} \qquad w^6 = \overline{w(f34, f32)}$$
$$w^3 = \overline{w(m32, f34)} \qquad w^7 = \overline{w(f32, m34)}$$
$$w^4 = \overline{w(m34, f32)} \qquad w^8 = \overline{w(f34, m32)}$$

We can rank all observed wage responses within a country from smallest to largest, and assign to each wage w its percentile value $p(w) \in [0, 1]$ within the full distribution of wage responses that we observe in that country. Then denote by $\{p^1, ..., p^8\}$ the means of these percentiles for each worker-comparator pair.

$$p^1 = \overline{p(w(m32, m34))} \qquad p^5 = \overline{p(w(f32, f34))}$$
$$p^2 = \overline{p(w(m34, m32))} \qquad p^6 = \overline{p(w(f34, f32))}$$
$$p^3 = \overline{p(w(m32, f34))} \qquad p^7 = \overline{p(w(f32, m34))}$$
$$p^4 = \overline{p(w(m34, f32))} \qquad p^8 = \overline{p(w(f34, m32))}$$

1.2 Discrimination measures

We propose to measure discrimination in two ways. Our primary measure is based on differences in (proposed) **wage percentiles** between men and women. Our secondary measure is based on a measure of **normalized differences in raw wages**. Both are designed to be expressed in comparable units across countries. We prioritize the percentile measure as primary because we expect it to be less noisy. The normalized wage measure is quite closely related to the Gini coefficient, see Appendix B for more discussion.

1.2.1 Percentile measure

For the percentile measure, we define **Explicit** discrimination as the average difference between male and female wage percentiles in different-gender comparisons (i.e., when a man and woman are compared side-by side):

$$\text{Explicit} := \frac{p^3 + p^4}{2} - \frac{p^7 + p^8}{2} \tag{1}$$

We will refer to the average difference between male and female wage percentiles in same-gender comparisons (i.e., when the respondent never directly compares a man to a woman) as **Indirect** discrimination:

$$\text{Indirect} := \frac{p^1 + p^2}{2} - \frac{p^5 + p^6}{2} \tag{2}$$

We expect indirect discrimination to exceed explicit discrimination in most cases. We define the difference between the two as **Implicit** discrimination. It is the part of indirect discrimination that is not expressed explicitly.[2]

$$\text{Implicit} := \text{Indirect} - \text{Explicit} \tag{3}$$

Finally, we define a fourth measure, **Experienced** discrimination. This is the amount of discrimination a female candidate experiences in our experiment on average, where she has a 50% chance of matching with a man.

$$\text{Experienced} := 0.5 \times \text{Explicit} + 0.5 \times \text{Indirect} = \text{Explicit} + 0.5 \times \text{Implicit} \tag{4}$$

Note that all our measures could be positive or negative. Indeed, we might expect that in some countries people Explicitly discriminate against men while Implicitly discriminating against women (this would be perfectly consistent with the theory of Cunningham and de Quidt (2023)).

1.2.2 Normalized wage measure

For the normalized wage measure, we use the following.

$$\begin{aligned}
\text{Explicit} &:= \frac{1}{2\bar{w}}\left[\frac{w^3+w^4}{2} - \frac{w^7+w^8}{2}\right] \\
\text{Indirect} &:= \frac{1}{2\bar{w}}\left[\frac{w^1+w^2}{2} - \frac{w^5+w^6}{2}\right] \\
\text{Implicit} &:= \text{Indirect} - \text{Explicit} \\
\text{Experienced} &:= \text{Explicit} + 0.5 \times \text{Implicit}
\end{aligned}$$

See appendix B.4 for explanation of the connection between our measure and the Gini index and why we normalize by $2\bar{w}$.

Our primary analysis will construct these measures at the country level and in the full sample. For the analysis in section 2.3 we will construct them in different subgroups.

1.3 Country selection

Our questions were included as a module in a larger Gallup survey to be conducted in 41 countries. According to our agreement with Gallup, we could include 31 already pre-selected countries to be surveyed online. In addition, we could include 10 countries that were to be surveyed through face-to-face interviews.[3]

The ten additional countries were selected according to the following procedure. First, we classified all countries according to their decile rank in the 2019 UN Gender Equality Index. The pool of countries to

[2]Note that our labels slightly differ from the terminology of Cunningham and de Quidt (2023); we think this difference makes sense for simple intuition. In that theory, implicit discrimination is expressed *more strongly* in the same-gender comparisons than the different-gender comparisons. Thus our treatments identify variation that is driven by implicit discrimination. But a) there could exist comparisons where implicit discrimination is *even stronger*, e.g. if gender was further "diluted," i.e., we do not necessarily observe the full extent of implicit discrimination. And b) some of the discrimination that we call "explicit" could in fact still be implicit, according to the definition of that theory (e.g., because even in the different-gender comparisons, gender is also diluted by age variation). See Appendix B.3 for formal discussion.

[3]The countries available to choose from were: Bangladesh, Bolivia, Cambodia, Cameroon, Colombia, Egypt, Ethiopia, Ghana, India, Indonesia, Iraq, Jordan, Kenya, Morocco, Nigeria, Pakistan, Peru, Philippines, Russia, Senegal, South Africa, Sri Lanka, Tanzania, Thailand, Turkey, Uganda, Ukraine, Venezuela, Vietnam, Zambia, and Zimbabwe. We decided to exclude Russia, Ukraine and Ethiopia from the pool of eligible countries due to ongoing conflict.

choose from all belonged to five distinct deciles (deciles 2-6). For each decile, we picked the two largest countries. To obtain a geographical spread, we resampled if the chosen country was from the same continent as an already sampled country in that decile. In addition, we decided to include a maximum of three countries per decile (including the already pre-selected countries of the online survey).

Table 2 lists the 41 selected countries, along with the survey mode (web/face-to-face interview). For each country, we imposed a survey constraint preventing wage responses outside a (wide) interval corresponding to the rounded 15th to 97th percentiles of the country pre-tax income distribution for the most recently available year (up to 2021). These were imposed to avoid order-of-magnitude errors when typing wages, without otherwise materially influencing the responses. We took them from the World Inequality Database `https://wid.world`.

1.4 Statistical details

1.4.1 Weights

In all our analyses we plan to use the poststratification weights supplied to us by Gallup.

1.4.2 Attrition/missing data

Respondents are permitted to skip our questions, which form part of a larger survey. Our primary analysis will treat missing responses as missing at random. We will regress an individual-level attrition dummy on treatment dummies and report F-tests for selective attrition by treatment arm (Treatments 1–4 in Table 1). If any countries have significant (at the 5% level) selective attrition by treatment, we will report robustness analyses that exclude those countries.

1.4.3 Outliers

As discussed above, we put loose bounds on allowable responses to minimize order-of-magnitude errors in the survey questions. However the possible range is still quite wide and outliers are a possible concern. We expect the percentile measure to be robust to outliers in wage responses. If we observe clear patterns of outlier influences in the normalized wage measures, we will also report versions based on winsorized data.

2 Analysis

The planned analysis is primarily descriptive, seeking to describe patterns of discrimination in the full sample and across countries. As such, we do not propose a large number of statistical tests nor multiple-testing corrections.

2.1 Describing patterns in explicit and implicit discrimination

We will produce a table showing our estimates of Explicit, Implicit, Indirect, and Experienced discrimination by country, with standard errors clustered at the respondent level.[4] We will also report the cross-country average, and test for overall presence of discrimination.

- We will conduct a simple two-sided t-test (from a single regression that pools data from all countries) that tests whether Experienced discrimination is significantly different from zero in the full sample.

[4]Since we have two wage observations per respondent, and exploit both within- and between-respondent variation, we need to account for clustering in the analysis. See appendix B.1 for details of the regression specification.

- We will also report two-sided t-tests testing whether its two components (Explicit and Implicit) are significantly different from zero in the full sample.

We will produce bar graphs that order countries by Experienced discrimination (from smallest to largest) and decompose it into Explicit and Implicit discrimination. We will also produce the graph decomposing Indirect discrimination into Explicit and Implicit which allows us to more easily describe how these two components relate to one another in magnitude. We also plan to make a colored world map shading countries according to our discrimination measures.

2.1.1 Relationship between implicit and explicit discrimination

It is interesting to explore how implicit and explicit discrimination are related to one another. Measuring this relationship is complicated by the fact that a) we measure implicit discrimination by subtracting explicit discrimination from indirect discrimination (equation (3)), and b) all terms are measured with error since they are based on sample averages.

Denote countries by $c \in \{1, ..., C\}$. Assume the following "structural" relationship, which assumes Implicit_c depends linearly on Explicit_c and an independent component ν_c.[5]

$$\text{Implicit}_c = \gamma_0 + \gamma_1 \text{Explicit}_c + \nu_c \tag{5}$$

$$\implies \text{Indirect}_c = \gamma_0 + (1 + \gamma_1)\text{Explicit}_c + \nu_c \tag{6}$$

where (6) follows from $\text{Indirect}_c := \text{Explicit}_c + \text{Implicit}_c$. We are interested in identifying γ_1. If we observed all terms we could simply regress Indirect_c on Explicit_c and subtract one from the slope coefficient. However, we do not observe Indirect_c or Explicit_c but our estimates:

$$\widehat{\text{Indirect}}_c = \text{Indirect}_c + \rho_c$$
$$\widehat{\text{Explicit}}_c = \text{Explicit}_c + \iota_c$$

where ρ_c, ι_c represent the sampling errors that arise because we do not observe the full population in any of our countries (these would naturally vanish as the number of respondents per country becomes sufficiently large).

Appendix B.2 shows that "naively" regressing $\widehat{\text{Indirect}}_c$ on $\widehat{\text{Explicit}}_c$ will in general result in a biased estimate of γ_1 (the underlying mechanism can be thought of as a form of attenuation bias). Our particular concern is that the bias could lead us to make a sign error: we could mistakenly conclude that $\gamma_1 < 0$ when in fact it is positive. We show in the Appendix how to construct a consistent estimate, using our estimate of the variance of ι_c.

Therefore we report two estimates of γ_1, first from the "naive" regression of Indirect_c or Explicit_c, and second, the debiased estimate, which will tend to be larger. Our primary estimate will be the debiased one.

Using the debiased measure we will formally test if γ_1 is significantly different from zero. A positive coefficient means that implicit discrimination is greater in countries with more explicit discrimination, suggesting that both types of discrimination go hand in hand. A negative coefficient means the reverse, which could be interpreted as showing that when explicit discrimination becomes socially unacceptable it finds its outlet in increased implicit discrimination.

[5] If Implicit_c and Explicit_c are completely independent, we would find $\gamma_1 = 0$. If they are positively related (more Explicit discrimination is associated with more Implicit discrimination) we would find $\gamma_1 > 0$, and the converse if they are negatively related.

2.2 Relating discrimination to measures of economic development and gender inequality

Next, we study the relationship between discrimination and other factors. To this end, we regress each discrimination measure (Implicit, Explicit and Experienced), estimated at the national level, on country-level indicators of economic development and gender inequality, respectively, to explore how these variables co-evolve. We will show results with and without regional fixed effects and controls for mode of data collection (online versus face-to-face). We will also estimate a joint specification with all correlates on the right-hand side.

We will measure economic development by gross national product per capita (logged), but we also look at the human development index and gross national income (logged). We will measure gender inequality by the Global Gender Gap Index (GGGI), but we will also look at the UNDP Gender Inequality Index (GII), the female-to-male labor force participation rate given by the ILO in the World Bank database, and the Gender Social Norms Index (GSNI).[6]

2.3 Measuring discrimination in demographic subgroups

We will pool data across countries and look at discrimination for the following subgroups:

1. Subgroup 1: Gender of respondent
2. Subgroup 2: WEIRD vs. non-WEIRD respondent (WEIRD are defined as the people with income and education above the median of their own country, and living in urban areas)
3. Subgroup 3: Age of respondent (split into terciles)

2.4 Simulating alternative labor markets

Our data allow us to simulate labor markets in which men and women meet with probabilities that differ from 50%. This analysis only makes sense to perform with the normalized wage measure because changing this probability would affect the wage distribution and hence alter the percentile measure.

Our experiment simulates a labor market with equal male and female labor force participation and random matching of comparators. In a fully segregated labor market, women would always match with women and men always match with men, meaning discrimination would always equal Indirect. A policy that required every candidate to be matched with an opposite-gender comparator would mean discrimination would always equal Explicit. We can explore alternatives by varying q_f (the probability that a woman matches with a woman) in this equation:

$$\text{Simulated}^1 := \text{Explicit} + q_f \times \text{Implicit}$$

Alternatively we could simulate variation in labor force participation, which also affects the probability a woman matches a woman (and the probability a man matches with a man). Normalizing the male participation rate to 1, assume female labor force participation equals $l_f \leq 1$ (and workers match randomly),[7] then the average worker expects to match with a man with probability $\frac{1}{1+l_f}$ and a woman with probability $\frac{l_f}{1+l_f}$. Plugging in our experimental estimates for the wages a man or woman expects

[6] In exploratory analyses, we include several additional correlates. For example, we look at each sub-component of the GII, the GGGI and the GSNI. We also use the data from Gallup to construct country-level indicators of participants' characteristics and beliefs.

[7] This assumes that the male labor force participation rate is higher than the female one.

to earn in each scenario, we obtain that the gap between male and female earnings will be:

$$\text{Simulated}^2 := \frac{1}{2\bar{w}} \left[\frac{w^1 + w^2 + l_f(w^3 + w^4)}{2(1 + l_f)} - \frac{w^7 + w^8 + l_f(w^5 + w^6)}{2(1 + l_f)} \right]$$

We keep the denominator unchanged to aid comparison with the primary measures of discrimination. We could further amend this formula to account for non-equal labor force participation combined with more or less segregated labor markets.

We propose to investigate in exploratory analyses how these simulated measures differ from the primary ones. We will again use the same measure of labor market participation from ILO/The World Bank as mentioned above, plus proxies for segregation in the labor market, such as those used by Bettio et al. (2009) to shed light on the role of participation and segregation for empirically measured discrimination in societies across the world.

References

Bettio, F., A. Verashchagina, and F. Camilleri-Cassar (2009). Gender segregation in the labour market: Root causes, implications and policy responses in the EU.

Cunningham, T. and J. de Quidt (2023). Implicit preferences. *mimeo*.

Sethi, R. (2021). Notes on the gini coefficient(s). *mimeo*.

A Tables

Table 2: Countries in sample

	Country	Mode	Currency	Lower bound	Upper bound
1	Argentina	Web	ARS	230,000	3,900,000
2	Australia	Web	AUD	18,000	210,000
3	Austria	Web	EUR	13,000	140,000
4	Belgium	Web	EUR	12,000	140,000
5	Brazil	Web	BRL	3,700	210,000
6	Bulgaria	Web	BGN	4,100	63,000
7	Canada	Web	CAD	15,000	220,000
8	Chile	Web	CLP	1,500,000	68,000,000
9	China	Web	CNY	15,000	330,000
10	Colombia	Interview	COP	2,200,000	120,000,000
11	Denmark	Web	DKK	140,000	1,300,000
12	Egypt	Interview	EGP	20,000	400,000
13	Finland	Web	EUR	14,000	140,000
14	France	Web	EUR	12,000	110,000
15	Germany	Web	EUR	10,000	140,000
16	Greece	Web	EUR	4,000	49,000
17	India	Interview	INR	38,000	1,100,000
18	Ireland	Web	EUR	16,000	180,000
19	Italy	Web	EUR	7,200	85,000
20	Japan	Web	JPY	1,000,000	18,000,000
21	Kenya	Interview	KES	59,000	1,700,000
22	Malaysia	Web	MYR	14,000	200,000
23	Mexico	Web	MXN	26,000	1,000,000
24	Netherlands	Web	EUR	15,000	150,000
25	New Zealand	Web	NZD	20,000	240,000
26	Norway	Web	NOK	280,000	2,100,000
27	Pakistan	Interview	PKR	84,000	1,300,000
28	Philippines	Interview	PHP	51,000	1,100,000
29	Poland	Web	PLN	18,000	210,000
30	Portugal	Web	EUR	5,300	69,000
31	South Africa	Interview	ZAR	6,900	810,000
32	South Korea	Web	KRW	8,500,000	160,000,000
33	Spain	Web	EUR	7,700	78,000
34	Sri Lanka	Interview	LKR	190,000	4,100,000
35	Sweden	Web	SEK	180,000	1,500,000
36	Switzerland	Web	CHF	24,000	220,000
37	Taiwan	Web	TWD	280,000	3,200,000
38	Tanzania	Interview	TZS	870,000	25,000,000
39	United Kingdom	Web	GBP	9,000	100,000
40	United States	Web	USD	12,000	280,000
41	Vietnam	Interview	VND	15,000,000	320,000,000

B Appendix

B.1 Regression specification

Respondent i in country c answers two questions $q \in \{a, b\}$, each corresponding to one of eight comparisons (e.g., assigning a wage for a 34-year old man compared to a 32-year old woman, etc). Their responses are determined as follows:

$$p_{iq}^t = \sum_{t=1}^{8} w^t + \nu_i + \epsilon_{iq}$$

In words, there is a true population parameter p^t, an independent individual-specific shock ν_i, and an individual-by-question specific shock ϵ_{iq}. The ν_is affect the covariance structure of the data, which we deal with by clustering at respondent level.

Then for country c, consider the following regression without a constant, where p_{iqc}^t are observed responses and p_c^t are the parameters on treatment dummies $T_{iqc}^t, t = 1, ..., 8$::

$$p_{iqc}^t = \sum_{t=1}^{8} p_c^t T_{iqc}^t + \nu_{ic} + \epsilon_{iqc} \tag{7}$$

From this we recover estimates $\hat{p}_c^t, t = 1, ..., 8$ which are nothing other than the comparison-specific sample means of the wage percentiles.

Now, by the definitions of our measures:

$$\text{Explicit}_c := \frac{p_c^3 + p_c^4}{2} - \frac{p_c^7 + p_c^8}{2}$$
$$\text{Indirect}_c := \frac{p_c^1 + p_c^2}{2} - \frac{p_c^5 + p_c^6}{2}$$

Rearranging, we get:

$$p_c^4 := 2\text{Explicit}_c - p_c^3 + p_c^7 + p_c^8$$
$$p_c^2 := 2\text{Indirect}_c - p_c^1 + p_c^5 + p_c^6$$

Substitute into (7) and rearrange:

$$\begin{aligned}
p_{iqc}^t = {} & p_c^1 T_{iqc}^1 + (2\text{Indirect}_c - p_c^1 + p_c^5 + p_c^6) T_{iqc}^2 + \\
& p_c^3 T_{iqc}^3 + (2\text{Explicit}_c - p_c^3 + p_c^7 + p_c^8) T_{iqc}^4 + \\
& p_c^5 T_{iqc}^5 + p_c^6 T_{iqc}^6 + p_c^7 T_{iqc}^7 + p_c^8 T_{iqc}^8 + \nu_{ic} + \epsilon_{iqc} \\
= {} & \text{Indirect}_c (2T_{iqc}^2) + \text{Explicit}_c (2T_{iqc}^4) + \\
& p_c^1 (T_{iqc}^1 - T_{iqc}^2) + p_c^3 (T_{iqc}^3 - T_{iqc}^4) + p_c^5 (T_{iqc}^5 + T_{iqc}^2) + \\
& p_c^6 (T_{iqc}^6 + T_{iqc}^2) + p_c^7 (T_{iqc}^7 + T_{iqc}^4) + p_c^8 (T_{iqc}^8 + T_{iqc}^4) + \nu_{ic} + \epsilon_{iqc}
\end{aligned}$$

So, if we define

$$\begin{aligned}
T^{1*}_{iqc} &:= T^{1}_{iqc} - T^{2}_{iqc} \\
T^{2*}_{iqc} &:= 2T^{2}_{iqc} \\
T^{3*}_{iqc} &:= T^{3}_{iqc} - T^{4}_{iqc} \\
T^{4*}_{iqc} &:= 2T^{4}_{iqc} \\
T^{5*}_{iqc} &:= T^{5}_{iqc} - T^{2}_{iqc} \\
T^{6*}_{iqc} &:= T^{6}_{iqc} - T^{2}_{iqc} \\
T^{7*}_{iqc} &:= T^{7}_{iqc} - T^{4}_{iqc} \\
T^{8*}_{iqc} &:= T^{7}_{iqc} - T^{4}_{iqc}
\end{aligned}$$

and run the regression

$$p^{t}_{iqc} = \sum_{t=1}^{8} \beta^{t}_{c} T^{t*}_{iqc} + \nu_{ic} + \epsilon_{iqc}$$

the regression coefficients on T^{2*}_{iqc} and T^{4*}_{iqc} correspond to our estimates of Indirect and Explicit discrimination, respectively. From those we can easily compute the other parameters of interest as linear combinations.

We can either estimate these parameters at the country level, or, by pooling all data estimate the overall average in our sample. We apply poststratification weights throughout.

B.2 Identifying the relationship between Implicit and Explicit discrimination

Assume the following "structural" relationship, which assumes Implicit_c depends linearly on Explicit_c and an independent component ν_c:

$$\text{Implicit}_c = \gamma_0 + \gamma_1 \text{Explicit}_c + \nu_c \tag{8}$$

$$\implies \text{Indirect}_c = \gamma_0 + (1+\gamma_1)\text{Explicit}_c + \nu_c \tag{9}$$

where (9) follows from the definition of $\text{Indirect}_c = \text{Explicit}_c + \text{Implicit}_c$. We are interested in identifying γ_1. If we observed all terms we could simply regress Indirect_c on Explicit_c and subtract one from the slope coefficient. However, we do not observe Indirect_c or Explicit_c but our estimates:

$$\begin{aligned}
\widehat{\text{Indirect}}_c &= \text{Indirect}_c + \rho_c \\
\widehat{\text{Explicit}}_c &= \text{Explicit}_c + \iota_c
\end{aligned}$$

where ρ_c and ι_c are the country-specific sampling errors in estimation of Indirect_c and Explicit_c, which by randomization, independent sampling are independent and approximately Normal with estimated variances $\hat{\sigma}^2_{\rho,c}$ and $\hat{\sigma}^2_{\iota,c}$.

Substituting into (9), we obtain:

$$\begin{aligned}
\widehat{\text{Indirect}}_c - \rho_c &= \gamma_0 + (1+\gamma_1)(\widehat{\text{Explicit}}_c - \iota_c) + \nu_c \\
\widehat{\text{Indirect}}_c &= \gamma_0 + (1+\gamma_1)\widehat{\text{Explicit}}_c + (\nu_c + \rho_c - (1+\gamma_1)\iota_c)
\end{aligned}$$

We have C observations of $\widehat{\text{Indirect}}_c$ and $\widehat{\text{Explicit}}_c$. If we then regress $\widehat{\text{Indirect}}_c$ on $\widehat{\text{Explicit}}_c$, we obtain

regression slope coefficient:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\widehat{Cov}(\widehat{\text{Explicit}}_c, \widehat{\text{Indirect}}_c)}{\widehat{Var}(\widehat{\text{Explicit}}_c)} \\ &= \frac{\widehat{Cov}(\widehat{\text{Explicit}}_c, \gamma_0 + (1+\gamma_1)\widehat{\text{Explicit}}_c + (\nu_c + \rho_c - (1+\gamma_1)\iota_c)}{\widehat{Var}(\widehat{\text{Explicit}}_c)} \\ &\xrightarrow[C\to\infty]{} (1+\gamma_1)\left(1 - \frac{Var(\iota_c)}{Var(\widehat{\text{Explicit}})}\right)\end{aligned}$$

Where the second line follows from substitution for $\widehat{\text{Implicit}}_c$ and the third follows from $\widehat{\text{Explicit}}_c = \text{Explicit}_c + \iota_c$ and the assumption that all disturbance terms are independent of one another.

We are interested in the parameter γ_1 (which captures the linear relationship between Explicit and Implicit). The problem is attenuation bias (since $\widehat{\text{Explicit}}_c$ is estimated with noise). Attenuation bias might mean that $\hat{\beta}_1 - 1 < 0$ even when $\hat{\gamma}_1 > 0$. Fortunately, we have estimates of the parameters of the attenuation bias term, and so can construct a consistent estimate as follows:

$$\hat{\gamma}_1 = \frac{\hat{\beta}_1}{\left(1 - \frac{\sum_{c=1}^{C} \frac{\hat{\sigma}^2_{\iota,c}}{C}}{\widehat{Var}(\widehat{\text{Explicit}})}\right)} - 1 \tag{10}$$

We plan to compute standard errors via a parametric bootstrap that resamples with replacement both the individuals within countries and the countries themselves.

B.3 Relationship to Cunningham and de Quidt (2023)

To save on notational clutter, we will use p throughout this section to denote the outcome for a given candidate, which in our experiment is either a wage percentile or a raw wage value. We then assume that this outcome of interest is determined by a linear separable function of explicit and implicit values, following the theory in Cunningham and de Quidt (2023). That allows us to interpret our treatments and analysis through the lens of that theory.

Represent candidates by vectors of binary attributes $\mathbf{x} \in \mathcal{X} = \{-1, 1\}^n$, attributes indexed by i. In our experiment attributes are gender, age, and a "background" attribute. We can write the wage percentile given to candidate $\mathbf{x}$ with comparator candidate $\mathbf{y}$ as:

$$p(\mathbf{x}, \mathbf{y}) = v(\mathbf{x}) + \sum_{i=1}^{n} x_i \kappa_i \theta_i(\mathbf{x}, \mathbf{y}),$$

where v is the "explicit value" of candidate $\mathbf{x}$ (the wage percentile they would get in the absence of any implicit discrimination), $\kappa_i \in \{-1, 0, 1\}$ is the implicit preference attached to attribute i (e.g., if $x_1 = 1$ denotes male gender, $\kappa_1 = 1$ would be an implicit preference favoring men).

θ_i is the "influence" of implicit preferences in this comparison. Each θ_i takes on one of two values depending on whether the candidates have the same or different gender. We simply write out the terms here, see Cunningham and de Quidt (2023) for the basis behind these claims. Summary in words:

- Implicit influence of gender is high when gender is shared (by "dominance-k" – see the paper)
- Implicit influence of age is high when gender is non-shared (by "dilution" – see the paper)
- Implicit influence of the background attribute is high when gender is shared (by "dilution")

For each attribute in turn:

- x_1: Gender (male = 1). $\theta_1(same) = \theta_1^H, \theta_1(diff) := \theta_1^L$
- x_2: Age (34 years = 1). $\theta_2(same) = \theta_2^L, \theta_2(diff) := \theta_2^H$.
  - Note that age is confounded with first/second presentation of the candidates, so we don't cleanly identify its implicit preference if any.
- x_3: Background (always = 1). $\theta_3(same) = \theta_3^H, \theta_2(diff) := \theta_2^L$.
  - This captures everything else that is held constant in all comparisons.

We can write out each comparison (m34 is a 34-year old male, etc):

$$\begin{array}{lllllll}
p^1 = p(m32, m34) = v(m32) & +\kappa_1\theta_1^H & -\kappa_2\theta_2^L & +\kappa_3\theta_3^H \\
p^2 = p(m34, m32) = v(m34) & +\kappa_1\theta_1^H & +\kappa_2\theta_2^L & +\kappa_3\theta_3^H \\
p^3 = p(m32, f34) = v(m32) & +\kappa_1\theta_1^L & -\kappa_2\theta_2^H & +\kappa_3\theta_3^L \\
p^4 = p(m34, f32) = v(m34) & +\kappa_1\theta_1^L & +\kappa_2\theta_2^H & +\kappa_3\theta_3^L \\
p^5 = p(f32, f34) = v(f32) & -\kappa_1\theta_1^H & -\kappa_2\theta_2^L & +\kappa_3\theta_3^H \\
p^6 = p(f34, f32) = v(f34) & -\kappa_1\theta_1^H & +\kappa_2\theta_2^L & +\kappa_3\theta_3^H \\
p^7 = p(f32, m34) = v(f32) & -\kappa_1\theta_1^L & -\kappa_2\theta_2^H & +\kappa_3\theta_3^L \\
p^8 = p(f34, m32) = v(f34) & -\kappa_1\theta_1^L & +\kappa_2\theta_2^H & +\kappa_3\theta_3^L
\end{array}$$

We can then express **explicit** discrimination (ignoring the normalization in the case of the normalized wage measure) as follows:

$$\begin{aligned}
\text{Explicit} &= \frac{p^3 + p^4}{2} - \frac{p^7 + p^8}{2} \\
&= \overline{v(m)} - \overline{v(f)} + 2\kappa_1\theta_1^L
\end{aligned} \tag{11}$$

We see that our explicit measure contains the difference in explicit values $\overline{v(m)} - \overline{v(f)}$, but also is partly influenced by implicit preferences through the term $2\kappa_1\theta_1^L$.

We express **implicit discrimination** as:

$$\begin{aligned}
\text{Implicit} &= \frac{p^1 + p^2}{2} - \frac{p^5 + p^6}{2} - \text{Explicit} \\
&= 2\kappa_1(\theta_1^H - \theta_1^L)
\end{aligned} \tag{12}$$

We see that our implicit measure is completely independent of explicit values, and gives us the implicit preference κ_1 weighted by the change in influence $(\theta_1^H - \theta_1^L)$, which is expected to be positive as outlined above.

Our measure of Indirect discrimination equals Explicit + Implicit, or

$$\text{Indirect} = \overline{v(m)} - \overline{v(f)} + 2\kappa_1\theta_1^H$$

where $2\kappa_1\theta_1^H$ is the full contribution of implicit preferences in the indirect comparisons (since the man with $x_1 = 1$ gets $+\kappa_1\theta_1^H$ while the woman gets $-\kappa_1\theta_1^H$.

Our measure of Experienced discrimination equals Explicit + 0.5Implicit, or

$$\text{Experienced} = \overline{v(m)} - \overline{v(f)} + \kappa_1\theta_1^L + \kappa_1\theta_1^H$$

where $\kappa_1\theta_1^L + \kappa_1\theta_1^H$ is the average influence of implicit preferences across comparisons.

B.4 Relationship to Gini index

As a benchmark, consider the Gini index for wage inequality between groups 1–8 in our sample. Intuitively, this compares the difference in wages for all pairwise comparisons to the average wage in the economy. For n groups this is:[8]

$$G := \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|w^i - w^j|}{2(n-1)\sum_{i=1}^{n} w^i}.$$

We have eight groups, so $G = \frac{\sum_{i=1}^{8}\sum_{j=1}^{8}|w^i - w^j|}{14\sum_{i=1}^{8} w^i}$. We propose two changes:

1. Only compare male wages to female wages. Note this changes the normalization:

$$G' := \frac{\sum_{i=1}^{4}\sum_{j=5}^{8}|w^i - w^j|}{4\sum_{k=1}^{8} w^k}$$

   G' equals 1 whenever all income goes to one gender, however divided within that gender. However, it doesn't distinguish between male-favoring and female-favoring inequality.

2. Replace absolute differences with simple differences. This would give the same result as the absolute difference version if all male groups earn more on average than all female groups. The simple-difference Gini is bounded between -1 and 1, and we get negative values if mean female income is greater than mean male income. In general, it is larger the more men earn relative to women. Intuitively, it is proportional to the average wage difference between men and women, divided by the average wage.

$$G^* := \frac{\sum_{i=1}^{4}\sum_{j=5}^{8}(w^i - w^j)}{4\sum_{k=1}^{8} w^k}$$

A little algebra shows that

$$\begin{aligned} G^* &= \frac{1}{4\bar{w}}\left[\frac{w^3 + w^4}{2} - \frac{w^7 + w^8}{2}\right] + \frac{1}{4\bar{w}}\left[\frac{w^1 + w^2}{2} - \frac{w^5 + w^6}{2}\right] \\ &= \frac{1}{2}(\text{Explicit} + \text{Indirect}) \\ &= \text{Explicit} + 0.5 \times \text{Implicit} = \text{Experienced} \end{aligned}$$

[8] As discussed by e.g. Sethi (2021), there are two popular forms of the Gini, which differ in whether the normalization is by n or $n-1$. These satisfy different axioms, in particular "population symmetry" which concerns how they behave when the population is duplicated.