Pre-Analysis Plan Amendment: Visibility as Vulnerability: Transgender Passing Privilege and Hiring Discrimination

Date: 05/01/2026
Investigators: Taryn Eames (University of Toronto)

Amendment

Minimal details were provided in the original Pre-Analysis Plan relating to Survey Experiment Two, which was originally conceived as a pure image rating exercise. Since then, I have changed the design of the survey to also involve a profile component, where respondents rate professional profiles (which include a headshot). As such, the survey experiment now includes the same three treatment arms used in the field experiment.

In this amendment, I provide additional information on the survey design and planned analyses. The two sections of italicized text below include all references to Survey Experiment 2 in the original Pre-Analysis Plan. I also ran a small pilot (N=50 in Germany and the United States) to inform power analysis and the target sample size.

Survey Design

A. Survey Goals

The following was provided in the original Pre-Analysis Plan:

In Survey Experiment Two, the goal is to measure additional attributes associated with each image—for example, how approachable/competent/easy-going/confident/etc. an image seems on a 0-100 scale. Again, "triplet" images will be included in this experiment.

This can be used to consider the difference in attribute measures in general between presumably transgender woman and their cisgender counterparts—what attributes have the largest difference between transgender and cisgender ratings? It can also be used to consider the extent to which different attributes contribute to discrimination.

In addition to the above, the survey will be fielded on Prolific, in both Germany (700 respondents) and the United States (700 respondents).

B. The Survey

Aside from survey consent, instructions, and debrief, the survey proceeds as follows:

1. Preliminary Demographic Questions
    a. What is your age (in years)?

b. Which lottery do you prefer? This is an attention check. To confirm you are paying attention, please select Lottery B.
c. Which state do you live in?
d. Which best describes the area where you live? [Downtown → Rural]

2. Brief Professional Profile Ratings
    a. Respondents view seven fictitious, brief professional profiles and rate each across a set of nine attributes. They are instructed to review each profile for approximately 30 seconds and to rate profiles based on their initial impression. Each profile is someone currently working in one of nine occupations; at the top of the page, the respondent sees:
        - Interested in a **[ROLE]** role ([brief description of role])—for example: *Interested in an **Office Clerk** role (office coordination, scheduling, data entry, customer service)*
    b. Attributes are rated on a 1-7 Likert scale [Not at all → Extremely]. They are presented in a random order, except for attribute ix. below, which always appears last. The order is randomized once for each respondent and then remains fixed throughout the survey. The attributes are:
        - How **competent** is this person?
        - How **trustworthy** is this person?
        - How **physically attractive** is this person?
        - How **professionally** does this person **present themselves**?
        - How **emotionally stable** does this person seem?
        - How **comfortable** would most **customers feel** interacting with this person?
        - How **comfortable** would most **coworkers feel** working closely with this person on a team?
        - How likely is this person to **raise concerns** at work about comments or behavior they view as **insensitive/exclusionary**?
            - This is meant to measure "woke"-ness, through a neutral workplace-relevant prompt
        - Compared to typical **[ROLE]** applicants, how likely would you be to **offer this person an interview**?
    c. Of the 18 total profile bases, 2 additionally contain an attention check with answers presented on the same 1-7 Likert scale *[This is an attention check. Please select [X] to confirm you are paying attention.]*

3. Additional Demographic Questions
    a. Please select the highest level of education you have completed.
    b. What is your annual household before-tax income?
    ~~c. Which of the following describes your primary job? (If you are not currently working, select your most recent job) [Occupation Categories]~~

d. In your job (or most recent job), how often do you interact with customers/clients/patients as a core part of your work? *[Daily → Never]*
e. In your occupation (or most recent occupation), in general, would you describe the workforce as… *[Mostly Men → Mostly Women]*
f. Have you ever been involved in hiring decisions (e.g., screening applications, interviewing, deciding who to hire)? *[Yes; in the past 2 years → Never]*
g. This is an attention check. To confirm you are paying attention, please select "Prefer not to say" below.
~~h. Overall, how religious are you? *[1-7 Likert, Not at all → Extremely]*~~
i. How would you describe your political views in each of the following areas?
    - Economic Issues *[1-7 Likert, Very Progressive → Very Conservative]*
    - Social Issues *[1-7 Likert, Very Progressive → Very Conservative]*
j. Which political party do you support most?
k. How often, to your knowledge, do you have direct contact (in person, by phone, online) with transgender people who are close to you, such as family members or friends? *[Daily or almost daily → Never]*
l. How often, to your knowledge, do you have direct contact (in person, by phone, online) with transgender people who are not close to you, such as coworkers or acquaintances? *[Daily or almost daily → Never]*
m. Overall, how positively or negatively do you feel toward the following groups?
    - Transgender people *[1-7 Likert, Very Negatively → Very Positively]*
    - Gay/lesbian people *[1-7 Likert, Very Negatively → Very Positively]*

4. Treatment Recollection Question
    n. While you were rating the profiles, did you think one of the profiles might be a transgender woman? *[Yes: I felt certain about this at the time → No: I did not think of that at the time, and even now I cannot think of any one profile this could apply to]*

Note: Question n. above was added after the pilot was fielded, at the suggestion of a colleague who raised the possibility that respondents' ability to infer that a profile described a transgender woman may affect the size of observed penalties. Because the question is asked only after all profile ratings are completed, it should not affect the profile rating outcomes. Questions c. and h. (crossed out) above were removed after the pilot was fielded, to reduce time spent answering questions and after power calculations determined how many hypotheses are feasible to test.

C. Profile Construction

Profiles are designed to resemble LinkedIn profiles or CVs and consist of five components: (i) a base profile (see Section D for details), (ii) a headshot, (iii) a name, (iv) a location, and (v) listed hobbies. Each respondent rates seven profiles; thus, seven base profiles are drawn from a pool of 18 and presented in random order. For each profile, location and hobbies are assigned at random.

Headshots and names are assigned jointly, not independently, to match the profile treatment or control arm. The same treatment arms are considered as in the main field experiment:

Group	Headshot	Name Change
Treatment 1	Pass Poorly	Disclosed (MtF)
Treatment 2	Pass Well	Disclosed (MtF)
Treatment 3	Pass Poorly	Not Disclosed

Control profiles are intended to represent cisgender men (presumably cisgender man headshot; male name) and cisgender women (presumably cisgender woman headshot; female name).

Each respondent sees exactly one treated profile, three cisgender women control profiles, and three cisgender men control profiles. Including only one treated profile is done to reduce social desirability bias and demand effects by obscuring the survey's specific purpose. Rather than appearing solely focused on transgender versus cisgender profiles, the survey is framed more generally as a study of how respondents form impressions of professional profiles. To make male-to-female name change disclosure appear less unusual and more consistent with the broader profile design, some cisgender women profiles also disclose a last name change.

An example profile is provided below:

Name: Daniel Reynolds
Location: Greenville, SC
Education: High School (2018)

Professional Summary

Friendly restaurant employee with a background in food service. Team-oriented, reliable, and service-focused. Confident with guests, always careful and responsible.

Work Experience (Last 2 Jobs Only)

Waiter, *The Lantern*
2024/08–Present
- Took reservations by phone and in person
- Warmly welcomed guests and provided courteous service
- Regularly checked the work area and kept it clean and organized

Waiter, *Maple Street Café*
2023/01–2024/08
- Served and assisted guests
- Worked as part of a team and followed hygiene standards reliably

Hobbies

Hosting friends for movie nights

D. Base Profile Generation

Base profiles were constructed for the field experiment occupations and generated as follows:

1. For each occupation, one base profile was written in English and one in German (with the help of ChatGPT), inspired by real resumes from each country. Company names listed are generally fictitious—if real, it is a large company that exists in much of the country.

2. With the help of ChatGPT, English profiles were translated to German. During translation, the following was prioritized:
    a. The German profile reads like it was written by a native German-born worker who has lived and worked in Germany all their lives
        i. This may require swapping United States education, credentials, or regulations with approximately equivalent German versions
    b. The German profile has the same "voice" as the original, given country norms (e.g., if the English profile is "flashy" or "sloppy" relative to United States norms, the German version is similarly "flashy" or "sloppy" relative to German norms)
    c. Both describe the same work tasks in the job description
3. Step 2 is repeated, translating from German to English
4. I hired a professional translator to review these translations (and the core survey questions) to:
    a. Review the profile translations and confirm
        i. There are no red flags or oddities in the original German profiles
        ii. The translations meet the above criteria
    b. Translate core survey questions from English to German

Profile education indicates that the fictitious individuals are between 24 and 30 years old.

E. Names

Current names listed on the profiles are randomly selected from a set of 20 female-sounding names and a set of 20 male-sounding names. If the treated profile discloses a male-to-female name change, the replacement first name is drawn from a separate and distinct set of 20 male-sounding first names. If the presumably cisgender woman discloses a change in last name, the replacement last name is drawn from a separate and distinct set of 20 last names.

The 20 female first names and 40 male first names were randomly selected from lists of the top 200 baby names of the 1990s in the relevant country (BabyMed, n.d.; Social Security Administration, 2025). Each name was then checked using genderize.io (n.d.) to ensure that it served as a strong gender signal. In the United States, only names classified by genderize.io as male or female with 100% probability were retained. In Germany, the same 100% threshold was used for male names, while a 99% threshold was used for female names, since very few German female names were classified as female with 100% probability.

The 40 last names were randomly selected from country-specific lists of common surnames. In the United States, surnames were drawn from the 1,000 most common surnames, restricting to those for which more than 50% of individuals with that last name were White (US Census Bureau, 2021). In Germany, surnames were drawn from the 100 most common surnames (Forebears, n.d.). Surnames that could also plausibly function as first names (e.g., Paul) were excluded to avoid gender ambiguity. Very similar surnames were filtered out, keeping the first-

drawn instance (e.g., Schmidt, Schmitt, Schmitz)—this was done to ensure respondents didn't find it odd if they were randomly assigned multiple profiles with very similar names.

F. Headshots

Headshots are randomly selected from the set of 30 "triplets" (i.e., 90 total headshot images) and assigned to match the profile condition. Presumably cisgender men are assigned the man image from the triplet; presumably cisgender women and passing transgender women are assigned the presumably cisgender woman image; and non-passing transgender women are assigned the non-passing transgender woman image.

G. Locations

The set of locations that can be listed on profiles are:

United States		**Germany**	
Location	Signal	Location	Signal
New York, NY	Left	Berlin, Berlin	Left
Los Angeles, CA	Left	Hamburg, Hamburg	Left
Minneapolis, MN	Left	Münster, NRW	Left
Portland, OR	Left	Köln, NRW	Left
Baltimore, MD	Left		
Seattle, WA	Left		
Riverside, CA	n/a	Mülheim an der Ruhr, NRW	n/a
Overland Park, KS	n/a	Krefeld, NRW	n/a
Reno, NV	n/a	Braunschweig, Niedersachsen	n/a
Omaha, NE	n/a	Osnabrück, Niedersachsen	n/a
Phoenix, AZ	n/a	Leverkusen, NRW	n/a
Indianapolis, IN	n/a	Hildesheim, Niedersachsen	n/a
Boise, ID	n/a		
Mesa, AZ	n/a		
Las Vegas, NV	n/a		
Fort Myers, FL	Right	Chemnitz, Sachsen	Right
Tulsa, OK	Right	Dresden, Sachsen	Right
Greenville, SC	Right	Augsburg, Bayern	Right
Wichita, KS	Right	Regensburg, Bayern	Right
Lubbock, TX	Right		
Colorado Springs, CO	Right		

The location listed on each profile is randomly selected from this country-specific list and is intended to signal left- or right-leaning politics were selected based on federal election results and on which cities people in each country would likely perceive as politically left or right. Locations without a political signal were randomly drawn from population weighted city lists in each country, where cities that might convey political signals were manually excluded.

H. Hobbies

Hobbies listed on the profiles are randomly selected from the following set:

Hobby	Signal
Running (I am training for a half-marathon)	Fitness
Swimming	Fitness
Hiking	Fitness
Long-distance cycling	Fitness
Going to the gym regularly	Fitness
Photography and photo/video editing	Productive
Languages (right now I'm learning Italian)	Productive
Personal coding projects	Productive
Learning new software programs	Productive
Chess and logic puzzles	Productive
Watching TV and movies	Passive
Walking around the neighborhood	Passive
Listening to podcasts	Passive
I follow sports – mainly basketball	Passive
Watching documentaries	Passive
Hosting friends for movie nights	Social
Hosting board game nights with friends	Social
Planning weekend get-togethers with friends	Social
Helping out at a local food bank	Prosocial
Mentoring students through a community program	Prosocial
Helping organize local donation drives	Prosocial
Volunteering with anti-discrimination groups	"Woke"

Hobbies are assigned randomly across the seven profiles, subject to three constraints: (i) exactly one profile is assigned the "woke" hobby, (ii) at most one profile is assigned the social hobby, (iii) at most one profile is assigned the prosocial hobby.

Analysis

A. Estimation Strategy

The following was provided in the original Pre-Analysis Plan:

Per image, I plan to measure attribute ratings as the average across all respondents.

I plan to measure the attribute "transgender gap" as follows:

$$a_{itnr} = \rho NP_i + \omega M_i + \theta_t + \sigma_n + \eta_r + \varepsilon_{itr}$$

where a_{itr} is the attribute value given to image applicant i in triplet t with name n by respondent r; NP_i equals 1 if image i is a presumably transgender woman (i.e., she does not pass); M_i equals 1 if image i is a presumably cisgender man; and $\theta_t, \sigma_n, \eta_r$ are triplet, name, and respondent fixed effects.

To consider heterogeneity in ratings, I can add interaction terms to the above based on respondent demographics (age, sex, education level, income, state of residence, political party).

This estimation strategy is modified and expanded below.

Treatment Effects on Attribute Ratings

For each mechanism-related attribute A, the "long" specification which identifies the impact of both treatments independently as well as their interaction is as follows:

$$(1) \quad A_i = \delta DT_i + \gamma NP_i + \lambda[DT_i \times NP_i] + X_i'\beta + \eta_{r(i)} + \tau_{b(i)} + \theta_{t(i)} + \kappa_{o(i)} + \varepsilon_i$$

where A_i is the outcome for observation i (and i indexes respondent-profile rating observations), DT_i equals 1 if the profile indirectly discloses her transgender identity via a male-to-female name change; NP_i equals 1 if the profile includes a non-passing headshot, X_i is a vector of controls, and $\eta_{r(i)}, \tau_{b(i)}, \theta_{t(i)}, \kappa_{o(i)}$ are respondent, base profile, image triplet, and profile order fixed effects.

Because the interaction term may be imprecisely estimated, I also consider the following, more parsimonious specification that omits the $DT_i \times NP_i$ interaction:

$$(2) \quad A_i = \delta DT_i + \gamma NP_i + X_i'\beta + \eta_{r(i)} + \tau_{b(i)} + \theta_{t(i)} + \kappa_{o(i)} + \varepsilon_i$$

In this "short" specification, δ and γ are composite effects: each coefficient averages the effect of one treatment component over the randomized distribution of the other treatment component.

To test for treatment effect heterogeneity, I interact the treatment indicators with moderators M_i:

$$(3) \quad A_i = \delta DT_i + \gamma NP_i + \rho M_i + \omega_1[M_i \times DT_i] + \omega_2[M_i \times NP_i] + X_i'\beta + \eta_{r(i)} + \tau_{b(i)} + \theta_{t(i)} + \kappa_{o(i)} + \varepsilon_i$$

The coefficients ω_1 and ω_2 test whether the effects of transgender identity disclosure and non-passing appearance differ by moderator M_i, respectively. When the main effect of $M_i$is absorbed by the fixed effects, it is omitted from the estimated equation. For power reasons, I use the short specification as the baseline for these heterogeneity analyses.

Compared to the original pre-analysis plan, the primary specification above omits name fixed effects. Names are randomly assigned from a large pool, so name-specific variation is orthogonal to treatment by design. In addition, since some profiles disclose a name change, name fixed effects are not entirely straightforward to implement. Because I do not expect large name effects, and because including a large number of name indicators is unlikely to materially improve precision (and may instead reduce it), I omit these fixed effects.

In the set of profile-specific control variables in X_i, I plan to include: (i) an indicator for whether a presumably cisgender profile discloses a last name change; (ii) an indicator for whether the profile is a presumably cisgender man; (iii) indicators for whether the listed hobby contains a fitness, productive, social, prosocial, or "woke" signal; and (iv) indicators for whether the listed location signals a left-leaning or right-leaning political geography. In regressions (1), (2), and (3) standard errors will be clustered at the respondent level.

Attribute Importance in Interview Offer Rating

Finally, I consider the extent to which each mechanism-related attribute is associated with the respondent's *interview offer* rating:

$$(4) \quad I_i = \sigma Z_i + \delta DT_i + \gamma NP_i + \omega_1[Z_i \times DT_i] + \omega_2[Z_i \times NP_i] + X_i'\beta + \tau_{b(i)} + \theta_{t(i)} + \eta_{r(i)} + \kappa_{o(i)} + \varepsilon_i$$

where I_i is the *interview offer* rating and Z_i is the standardized rating of the attribute of interest A. To produce Z_i, the attribute rating is standardized within respondent country and profile gender using only control profiles:

$$Z_i \equiv \frac{A_i - \bar{A}_{0,c(i),g(i)}}{s_{A,0,c(i),g(i)}}$$

where $c(i)$ denotes the respondent's country, $g(i)$ denotes the profile's gender identity (man or woman), and the subscript 0 refers to control profiles (presumably cisgender men and women). Thus, $\bar{A}_{0,c(i),g(i)}$ and $s_{A,0,c(i),g(i)}$ are the mean and standard deviation of A_i among control profiles rated by respondents in the same country as respondent i and with the same gender as profile i.

Treatment Effects on Transgender Signal Recognition

In addition to the attribute rating outcomes above, I examine whether the treatment cue shown to the respondent affects whether they recognize or suspect that one of the evaluated profiles may have been a transgender woman. Because this outcome is measured at the respondent level and every respondent views exactly one treated profile, this analysis does not contain a respondent-level no-cue control condition. I therefore estimate differences across the three treated-profile cue combinations, rather than estimating separate main effects of DT_r, NP_r, and their interaction.

Let N_r equal 1 if respondent r viewed a treated profile with a non-passing headshot and no male-to-female name-change disclosure and let B_r equal 1 if respondent r viewed a treated profile with both a non-passing headshot and a male-to-female name-change disclosure. The omitted

category is respondents who viewed a treated profile with a passing headshot and a male-to-female name-change disclosure. I estimate:

$$(5) \quad R_r = \gamma N_r + \lambda B_r + X_r'\beta + \tau_{b(r)} + \theta_{t(r)} + \Gamma_{c(r)} + \kappa_{o(r)} + \varepsilon_r$$

where r indexes respondents; R_r equals 1 if the respondent reports that, while rating the profiles, they recognized or suspected that one profile might be a transgender woman;[1] X_r includes controls corresponding to the treated profile viewed by the respondent; and $\tau_{b(r)}$, $\theta_{t(r)}$, $\Gamma_{c(r)}$, $\kappa_{o(r)}$are treated base profile, image triplet, country, and treated profile order fixed effects. Because this analysis is at the respondent level, I use heteroskedasticity-robust standard errors rather than clustering by respondent.

I also estimate the corresponding heterogeneity specification:

$$(6) \quad R_r = \gamma N_r + \lambda B_r + \rho M_r + \omega_1[M_r \times N_r] + \omega_2[M_r \times B_r] + X_r'\beta + \tau_{b(r)} + \theta_{t(r)} + \Gamma_{c(r)} + \kappa_{o(r)} + \varepsilon_r$$

where $M_r$is the moderator of interest.

Because this analysis is at the respondent level, I use heteroskedasticity-robust standard errors rather than clustering by respondent.

B. Hypotheses

I have organized hypotheses into six families for clarity. All six families will use pooled data from both Germany and the United States unless otherwise stated. If a variable is an outcome or key moderator, I plan to code "prefer not to answer" as missing and omit those observations from that specific analysis. Hypothesis families are below:

1. **Family 1**: via equation (1), tests whether the two main treatment components (indirect transgender identity disclosure and non-passing appearance), and their interaction affect ratings on each of the eight mechanism-related attributes, excluding the *interview offer* attribute. This family contains 24 tests in total: for each attribute, one test of H_0: $\delta = 0$, one test of H_0: $\gamma = 0$, and one test of H_0: $\lambda = 0$.

2. **Family 2:** via equation (2), tests whether the estimated effects of the two treatment components differ from one another on each of the same eight attributes. This family contains 8 tests in total, one per attribute, corresponding to H_0: $\delta - \gamma = 0$.

[1] Specifically, R_r equals 1 if the respondent selected one of the following answers to question n. ("While you were rating the profiles, did you think one of the profiles might be a transgender woman?"): "Yes: I felt certain about this at the time," "Yes: I suspected this at the time," or "Yes: I felt very unsure, but this crossed my mind at the time." R_r equals zero if the respondent selected either: "No: I did not think of that at the time, but now I can think of one profile this could apply to" or "No: I did not think of that at the time, and even now I cannot think of any one profile this could apply to."

3. **Family 3**: via equation (3), tests whether the two main treatment components differ across respondent and profile characteristics for the *interview offer* attribute. This family contains 28 tests in total: for each moderator, one test of H_0: $\omega_1 = 0$ and one test of H_0: $\omega_2 = 0$. See Section E for the list of 14 planned moderators.

4. **Family 4**: via equation (3), tests whether treatment effects in all nine attributes differ between Germany and the United States. This family contains 18 tests in total: for each attribute, one test of H_0: $\omega_1 = 0$ and one test of H_0: $\omega_2 = 0$.

5. **Family 5**: via equation (4), tests which of the eight mechanism-related attributes are most strongly associated with *interview offer* ratings, and whether the association between each attribute's rating and the *interview offer* rating differs for profiles containing either of the two main treatment components (relative to control profiles). This family contains 24 tests in total: for each mechanism-related attribute, one test of H_0: $\sigma = 0$, one test of H_0: $\omega_1 = 0$, and one test of H_0: $\omega_2 = 0$.

6. **Family 6**: via equation (5), tests whether the treatment cues affect whether the respondent recognized or suspected the person in the treated profile as a transgender woman; and via equation (6), tests whether those effects differ across respondent characteristics.
    a. This family contains 3 tests regarding baseline treatment effects in equation (5): one test of H_0: $\gamma = 0$, one test of H_0: $\lambda = 0$, and one test of H_0: $\lambda - \gamma = 0$.
    b. This family contains 4 tests regarding heterogeneity in treatment effects in equation (6): for each of the two moderators, one test of H_0: $\omega_1 = 0$ and one test of H_0: $\omega_2 = 0$. The moderators are: (i) distance to downtown and (ii) contact with transgender people who are not personally close to the respondent.

    Note: I include this family because recognition may be an important first stage in discrimination against transgender people. If respondents do not recognize or suspect that a profile represents a transgender woman, treatment effects are less likely to reflect transgender-related bias. I test whether recognition effects are larger among respondents who may have had more exposure to transgender people: German respondents in more urban areas,[2] and respondents who report contact with transgender people who are not personally close to them. I focus on non-close contact because it may increase respondents' ability to infer that someone is transgender without generating the warmth or empathy associated with close contact. If so, these respondents may have greater scope to discriminate on that basis when assigning ratings.

[2] This is driven by German field experiment evidence which indicates that discrimination is larger the closer a job is located to the city center.

To account for multiple hypothesis testing, I will control the False Discovery Rate using the Benjamini-Hochberg step-up procedure across the full set of 109 hypotheses (Benjamini and Hochberg, 1995).[3] The adjustment will be applied jointly across all hypotheses, rather than separately within hypothesis families.

C. Power Analysis: MDE Calculation Approach

Following Rainey (2026), I used a pilot study to inform the target sample size and power analysis. The pilot was conducted in March 2026 and included $G = 50$ respondents in Germany and $G = 50$ respondents in the United States, where G denotes the number of respondent clusters. The final survey and data collection process will follow the pilot design (except that question n. was added at the end of the survey, after respondents completed the experimental profile ratings). Using the principle that variance inputs used in power calculations should correspond to the planned analysis, I estimated the hypotheses described above using the pilot data and treated the resulting respondent-clustered standard errors for the coefficients of interest as estimates of precision. I then projected these standard errors to various sample sizes using inverse square root scaling in the number of respondent clusters:

$$\widehat{SE}_{study} = \widehat{SE}_{pilot} \times \sqrt{\frac{G_{pilot}}{G_{study}}}$$

Because the primary analysis controls the false discovery rate at 5 percent using the Benjamini-Hochberg step-up procedure across the full set of 109 hypotheses, the relevant significance cutoff depends on the rejection order. Let:

$$\alpha_H = \frac{H}{109} \times 0.05$$

denote the Benjamini-Hochberg critical value associated with the H^{th} ordered rejection. For each H, I compute the corresponding benchmark minimum detectable effect as:

$$MDE_H = \left(z_{1-\alpha_H/2} + z_{1-\beta}\right) \times \widehat{SE}_{\text{study}}$$

where $1 - \beta = 0.80$ is the target power.

These are approximate MDE benchmarks specific to rejection order, not exact power calculations for the full Benjamini-Hochberg procedure. They indicate the minimum effect size that would be detectable if a coefficient had to clear the Benjamini-Hochberg threshold associated with the H^{th} rejection. I report these benchmarks for several values of H $(1, 10, 20, 30, 40)$ to show how detectability changes under different assumptions about how many hypotheses are ultimately rejected.

[3] These 109 hypotheses include all hypotheses in the six families above, but exclude linked field experiment analyses, supplementary analyses, and robustness checks described in Sections F, G, and H

D. Power Analysis: Target Sample and Resulting MDE ($G = 1{,}400$)

After conducting the power analysis, I selected a target sample size of $G = 1{,}400$, equally split between Germany and the United States. I believe this sample provides adequate power for the main analyses for three reasons.

1. In **Family 1**, the MDE_{10} for the two main treatment components in equation (1) is approximately 0.16 to 0.26 points on the 1–7 Likert scale across the eight mechanism-related attributes. I would characterize these as generally small effects, and I expect many true effects to be at least this large.

2. In **Family 5**, the MDE_{10} for the coefficients on the standardized mechanism-related attributes in equation (4) is approximately 0.06 to 0.08 points in the *interview offer* rating per one standard deviation increase in the mechanism-related attribute. I would also characterize these as small effects, and I expect many true attribute effects on *interview offer* ratings to exceed this threshold.

3. Detectable effects for comparisons and interaction terms are larger but remain small enough that their investigation remains substantively meaningful. In **Family 1**, the MDE_{20} for the interaction term in equation (1) is approximately 0.28 to 0.36 points on the 1-7 Likert scale. In **Family 2**, the MDE_{20} for tests of whether the effects of transgender identity disclosure and non-passing appearance differ from one another is approximately 0.24 to 0.33 Likert points. In **Family 3**, the MDE_{20} for moderator interactions in equation (3), restricting attention to indicator moderators, is approximately 0.34 to 0.48 Likert points. In **Family 4**, the MDE_{20} for the U.S. moderator in equation (3) is approximately 0.26 to 0.38 Likert points. I would characterize these as moderate effects: such effects would represent meaningful departures from equality or additivity, and I expect at least some true effects in this set to exceed these thresholds. I also expect that more than 20 of the 109 hypotheses will satisfy the FDR criterion. If so, the relevant Benjamini–Hochberg cutoff will be less stringent than the $H = 20$ benchmark, and effective MDEs will thus be somewhat smaller.

These targets also seem broadly reasonable given prior evidence. Van Borm and Baert (2018), using the same 1–7 Likert response scale, report treatment effects when respondents rate transgender women compared to cisgender women applicants ranging in absolute value from 0.181 to 1.629 points, with most statistically significant estimates falling between 0.371 and 0.610 points. Although that study considers different attributes and a different respondent population and does not consider moderators or the role of not passing as cisgender, these estimates suggest that effects in the range detectable here are not implausibly large.

Family 1

Results for Family 1 are below. For reference: δ is the coefficient on DT_i (equals 1 if the profile indirectly discloses her transgender identity via a male-to-female name change), γ is the

coefficient on NP_i (equals 1 if the profile includes a non-passing headshot), and λ is the coefficient on their interaction $[DT_i \times NP_i]$—all specified from equation (1).

Attribute	Null Hypothesis	$\widehat{SE}_{pilot}$	MDE_1	MDE_{10}	MDE_{20}	MDE_{40}
Competent	$H_0: \delta = 0$	0.214	0.249	0.211	0.198	0.183
Trustworthy	$H_0: \delta = 0$	0.195	0.226	0.191	0.179	0.166
Physically Attractive	$H_0: \delta = 0$	0.176	0.205	0.173	0.162	0.151
Professionally Presented	$H_0: \delta = 0$	0.193	0.225	0.190	0.178	0.165
Emotionally Stable	$H_0: \delta = 0$	0.234	0.271	0.229	0.215	0.200
Customer Comfort	$H_0: \delta = 0$	0.250	0.290	0.245	0.230	0.214
Coworker Comfort	$H_0: \delta = 0$	0.213	0.248	0.210	0.196	0.182
Raise Concerns	$H_0: \delta = 0$	0.213	0.247	0.209	0.196	0.182
Competent	$H_0: \gamma = 0$	0.176	0.205	0.173	0.163	0.151
Trustworthy	$H_0: \gamma = 0$	0.171	0.198	0.168	0.157	0.146
Physically Attractive	$H_0: \gamma = 0$	0.206	0.239	0.202	0.190	0.176
Professionally Presented	$H_0: \gamma = 0$	0.249	0.289	0.245	0.230	0.213
Emotionally Stable	$H_0: \gamma = 0$	0.171	0.199	0.168	0.158	0.147
Customer Comfort	$H_0: \gamma = 0$	0.161	0.186	0.158	0.148	0.137
Coworker Comfort	$H_0: \gamma = 0$	0.175	0.203	0.171	0.161	0.149
Raise Concerns	$H_0: \gamma = 0$	0.267	0.309	0.262	0.246	0.228
Competent	$H_0: \lambda = 0$	0.328	0.381	0.322	0.302	0.280
Trustworthy	$H_0: \lambda = 0$	0.307	0.357	0.302	0.283	0.263
Physically Attractive	$H_0: \lambda = 0$	0.370	0.430	0.364	0.341	0.317
Professionally Presented	$H_0: \lambda = 0$	0.393	0.457	0.387	0.362	0.336
Emotionally Stable	$H_0: \lambda = 0$	0.371	0.430	0.364	0.341	0.317
Customer Comfort	$H_0: \lambda = 0$	0.372	0.432	0.366	0.343	0.318
Coworker Comfort	$H_0: \lambda = 0$	0.394	0.458	0.388	0.363	0.337
Raise Concerns	$H_0: \lambda = 0$	0.393	0.456	0.386	0.362	0.336

Family 2

Results for Family 2 are below. For reference: δ is the coefficient on DT_i and γ is the coefficient on NP_i in equation (2).

Attribute	Null Hypothesis	$\widehat{SE}_{pilot}$	MDE_1	MDE_{10}	MDE_{20}	MDE_{40}
Competent	$H_0: \delta - \gamma = 0$	0.292	0.340	0.287	0.269	0.250
Trustworthy	$H_0: \delta - \gamma = 0$	0.260	0.302	0.255	0.239	0.222
Physically Attractive	$H_0: \delta - \gamma = 0$	0.277	0.322	0.272	0.255	0.237
Professionally Presented	$H_0: \delta - \gamma = 0$	0.314	0.365	0.309	0.290	0.269
Emotionally Stable	$H_0: \delta - \gamma = 0$	0.291	0.338	0.286	0.268	0.249
Customer Comfort	$H_0: \delta - \gamma = 0$	0.303	0.352	0.298	0.279	0.259
Coworker Comfort	$H_0: \delta - \gamma = 0$	0.267	0.310	0.262	0.246	0.228
Raise Concerns	$H_0: \delta - \gamma = 0$	0.359	0.417	0.353	0.331	0.307

Family 3

Results for Family 3 are below. For reference: ω_1 is the coefficient on $[M_i \times DT_i]$ and ω_2 is the coefficient on $[M_i \times NP_i]$ in equation (3)—see Section E below for details. Below are results for respondent moderators (i.e., characteristics about the respondent) and the *interview offer* attribute as the only outcome

Moderator	Type	Null Hypothesis	$\widehat{SE}_{pilot}$	MDE_1	MDE_{10}	MDE_{20}	MDE_{40}
Age (years)	Integer	$H_0: \omega_1 = 0$	0.015	0.018	0.015	0.014	0.013
Male	Indicator	$H_0: \omega_1 = 0$	0.459	0.533	0.451	0.423	0.392
Male Dominated Occupation	Indicator	$H_0: \omega_1 = 0$	0.465	0.540	0.457	0.428	0.398
Customer Service Occupation	Indicator	$H_0: \omega_1 = 0$	0.419	0.486	0.411	0.386	0.358
Socially Conservative	1-7 Likert	$H_0: \omega_1 = 0$	0.142	0.164	0.139	0.130	0.121
Negative Transgender Feelings	Adj. Likert	$H_0: \omega_1 = 0$	0.403	0.468	0.396	0.371	0.344
Transgender Contact (Close)	Indicator	$H_0: \omega_1 = 0$	0.403	0.468	0.396	0.371	0.345
Transgender Contact (Not Close)	Indicator	$H_0: \omega_1 = 0$	0.486	0.564	0.477	0.448	0.415
Distance to Downtown*	Integer	$H_0: \omega_1 = 0$	0.274	0.318	0.269	0.252	0.234
Post Secondary Education	Indicator	$H_0: \omega_1 = 0$	0.462	0.536	0.454	0.425	0.395
High Income	Indicator	$H_0: \omega_1 = 0$	0.385	0.447	0.378	0.355	0.329
Age (years)	Integer	$H_0: \omega_2 = 0$	0.015	0.017	0.015	0.014	0.013
Male	Indicator	$H_0: \omega_2 = 0$	0.410	0.477	0.403	0.378	0.351
Male Dominated Occupation	Indicator	$H_0: \omega_2 = 0$	0.510	0.593	0.501	0.470	0.436
Customer Service Occupation	Indicator	$H_0: \omega_2 = 0$	0.404	0.469	0.397	0.372	0.346
Socially Conservative	Likert Scale	$H_0: \omega_2 = 0$	0.128	0.149	0.126	0.118	0.110
Negative Transgender Feelings	Adj. Likert	$H_0: \omega_2 = 0$	0.252	0.293	0.247	0.232	0.215
Transgender Contact (Close)	Indicator	$H_0: \omega_2 = 0$	0.370	0.430	0.364	0.341	0.317
Transgender Contact (Not Close)	Indicator	$H_0: \omega_2 = 0$	0.472	0.548	0.463	0.434	0.403
Distance to Downtown*	Integer	$H_0: \omega_2 = 0$	0.314	0.365	0.309	0.290	0.269
Post Secondary Education	Indicator	$H_0: \omega_2 = 0$	0.397	0.462	0.390	0.366	0.339
High Income	Indicator	$H_0: \omega_2 = 0$	0.403	0.468	0.396	0.371	0.345

** Includes German respondents only*

Below are results for profile moderators (i.e., characteristics about the profile) and the *interview offer* attribute as the only outcome

Moderator	Type	Null Hypothesis	$\widehat{SE}_{pilot}$	MDE_1	MDE_{10}	MDE_{20}	MDE_{40}
Male Dominated Occupation	Indicator	$H_0: \omega_1 = 0$	0.433	0.502	0.425	0.399	0.370
Customer Service Occupation	Indicator	$H_0: \omega_1 = 0$	0.521	0.605	0.512	0.480	0.445
Politically "Right" Location	Indicator	$H_0: \omega_1 = 0$	0.381	0.443	0.374	0.351	0.326
Male Dominated Occupation	Indicator	$H_0: \omega_2 = 0$	0.415	0.482	0.408	0.382	0.355
Customer Service Occupation	Indicator	$H_0: \omega_2 = 0$	0.476	0.553	0.468	0.439	0.407
Politically "Right" Location	Indicator	$H_0: \omega_2 = 0$	0.385	0.447	0.379	0.355	0.330

Family 4

Results for Family 4 are below. For reference: ω_1 is the coefficient on $[US_i \times DT_i]$ (US_i equals 1 if the respondent evaluating profile i lives in the United States) and ω_2 is the coefficient on $[US_i \times NP_i]$ in equation (3).

Attribute	Null Hypothesis	$\widehat{SE}_{pilot}$	MDE_1	MDE_{10}	MDE_{20}	MDE_{40}
Competent	$H_0: \omega_1 = 0$	0.303	0.351	0.297	0.279	0.259
Trustworthy	$H_0: \omega_1 = 0$	0.298	0.346	0.293	0.274	0.255
Physically Attractive	$H_0: \omega_1 = 0$	0.311	0.361	0.305	0.286	0.266
Professionally Presented	$H_0: \omega_1 = 0$	0.324	0.376	0.318	0.298	0.277
Emotionally Stable	$H_0: \omega_1 = 0$	0.334	0.388	0.328	0.308	0.286
Customer Comfort	$H_0: \omega_1 = 0$	0.368	0.428	0.362	0.339	0.315
Coworker Comfort	$H_0: \omega_1 = 0$	0.365	0.423	0.358	0.336	0.312
Raise Concerns	$H_0: \omega_1 = 0$	0.355	0.413	0.349	0.327	0.304
Interview Offer	$H_0: \omega_1 = 0$	0.417	0.484	0.409	0.384	0.356
Competent	$H_0: \omega_2 = 0$	0.293	0.340	0.288	0.270	0.251
Trustworthy	$H_0: \omega_2 = 0$	0.284	0.330	0.279	0.261	0.243
Physically Attractive	$H_0: \omega_2 = 0$	0.340	0.394	0.334	0.313	0.290
Professionally Presented	$H_0: \omega_2 = 0$	0.371	0.431	0.365	0.342	0.317
Emotionally Stable	$H_0: \omega_2 = 0$	0.312	0.362	0.306	0.287	0.266
Customer Comfort	$H_0: \omega_2 = 0$	0.342	0.397	0.336	0.315	0.292
Coworker Comfort	$H_0: \omega_2 = 0$	0.347	0.402	0.340	0.319	0.296
Raise Concerns	$H_0: \omega_2 = 0$	0.404	0.469	0.397	0.372	0.345
Interview Offer	$H_0: \omega_2 = 0$	0.414	0.481	0.407	0.381	0.354

Family 5

Results for Family 5 are below. For reference: σ is the coefficient on A_i (A_i is the standardized rating of the attribute of interest A, standardized using control profiles only), ω_1 is the coefficient on $[A_i \times DT_i]$ and ω_2 is the coefficient on $[A_i \times NP_i]$ in equation (4).

Attribute	Null Hypothesis	$\widehat{SE}_{pilot}$	MDE_1	MDE_{10}	MDE_{20}	MDE_{40}
Competent	$H_0: \sigma = 0$	0.057	0.066	0.056	0.052	0.048
Trustworthy	$H_0: \sigma = 0$	0.068	0.079	0.067	0.063	0.059
Physically Attractive	$H_0: \sigma = 0$	0.071	0.082	0.070	0.065	0.061
Professionally Presented	$H_0: \sigma = 0$	0.062	0.072	0.061	0.057	0.053
Emotionally Stable	$H_0: \sigma = 0$	0.062	0.072	0.061	0.057	0.053
Customer Comfort	$H_0: \sigma = 0$	0.076	0.088	0.074	0.070	0.065
Coworker Comfort	$H_0: \sigma = 0$	0.073	0.085	0.072	0.068	0.063
Raise Concerns	$H_0: \sigma = 0$	0.084	0.097	0.082	0.077	0.072
Competent	$H_0: \omega_1 = 0$	0.146	0.170	0.144	0.135	0.125
Trustworthy	$H_0: \omega_1 = 0$	0.179	0.208	0.176	0.165	0.153
Physically Attractive	$H_0: \omega_1 = 0$	0.142	0.165	0.139	0.131	0.121
Professionally Presented	$H_0: \omega_1 = 0$	0.168	0.195	0.165	0.154	0.143
Emotionally Stable	$H_0: \omega_1 = 0$	0.141	0.164	0.138	0.130	0.120
Customer Comfort	$H_0: \omega_1 = 0$	0.134	0.155	0.131	0.123	0.114
Coworker Comfort	$H_0: \omega_1 = 0$	0.152	0.176	0.149	0.140	0.130
Raise Concerns	$H_0: \omega_1 = 0$	0.208	0.241	0.204	0.192	0.178
Competent	$H_0: \omega_2 = 0$	0.135	0.156	0.132	0.124	0.115
Trustworthy	$H_0: \omega_2 = 0$	0.123	0.143	0.121	0.113	0.105
Physically Attractive	$H_0: \omega_2 = 0$	0.129	0.149	0.126	0.118	0.110
Professionally Presented	$H_0: \omega_2 = 0$	0.116	0.134	0.114	0.107	0.099
Emotionally Stable	$H_0: \omega_2 = 0$	0.116	0.135	0.114	0.107	0.099
Customer Comfort	$H_0: \omega_2 = 0$	0.120	0.139	0.118	0.110	0.102
Coworker Comfort	$H_0: \omega_2 = 0$	0.136	0.158	0.134	0.126	0.117
Raise Concerns	$H_0: \omega_2 = 0$	0.171	0.198	0.168	0.157	0.146

Family 6

Because question n. was added after the pilot was fielded, it is not possible to conduct pilot-based power analysis for Family 6.

E. Moderators in Family 3

I plan to explore the following moderators in Family 3. I do not have strong priors about whether heterogeneity will operate primarily through male-to-female name-change disclosure, non-passing appearance, or both. To my knowledge, prior work has not examined moderators of

discrimination using separate experimental variation in transgender disclosure and passing status. Because the survey is exploratory and mechanism-oriented, I therefore pre-specify a broad set of moderators. The priors discussed below focus on the expected overall direction of heterogeneity rather than on which treatment component will drive it.

Several moderators are motivated by patterns observed in the complementary field experiment. First, the field experiment suggests that discrimination is larger for jobs located closer to the city center. Because this result is somewhat counterintuitive, I use the survey experiment to test whether treatment penalties vary with respondents' own distance from downtown (and explore potential reasons). Second, the field experiment leaves some ambiguity about occupation-based heterogeneity: larger penalties for non-passing applicants may reflect stronger discrimination in customer-facing occupations, weaker penalties in male-dominated occupations, or both. The survey experiment helps separate these possibilities by examining heterogeneity both by respondents' occupational context and by the occupation listed on the evaluated profile.

The planned moderators are:

1. Respondent Moderator: Age (years)
    a. Via: What is your age (in years)?
    b. I plan to use age answers directly
    c. My prior is that older respondents will exhibit larger treatment penalties—evidence is mixed (Norton and Herek, 2013), but some suggests that younger people are generally more accepting of transgender people (Parker et al., 2022)
2. Respondent Moderator: Male
    a. Via: demographic information provided by Prolific
    b. I plan to code an indicator equal to 1 if the respondent is Male
    c. My prior is that male respondents exhibit larger treatment penalties—there is evidence that women consistently hold more favorable attitudes and less prejudice toward transgender people (Billard, 2018)
3. Respondent Moderator: Socially Conservative
    a. Via: How would you describe your political views in each of the following areas?
    b. I plan to use respondent 'Social Issues' ratings directly
    c. My prior is that socially conservative respondents exhibit larger treatment penalties—there is evidence that conservative individuals report higher prejudice against transgender people (Napier, 2025)
4. Respondent Moderator: Negative Transgender Feelings
    a. Via: Overall, how positively or negatively do you feel toward the following groups? [focus on response to Transgender people]

   b. I plan to code a variable representing negative feelings as follows: 0 if the respondent selected 4 or higher, 1 if they selected 3, 2 if they selected 2, and 3 if they selected 1[4]
   c. My prior is that respondents who report negative feelings toward transgender people exhibit larger treatment penalties
5. Respondent Moderator: Transgender Contact (Close)
   a. Via: How often, to your knowledge, do you have direct contact (in person, by phone, online) with transgender people who are close to you, such as family members or friends?
   b. I plan to code an indicator equal to 1 if the respondent selected anything other than 'Never' when answering this question[5]
   c. My prior is that individuals who have direct contact with transgender people who are close to them exhibit smaller treatment penalties
6. Respondent Moderator: Transgender Contact (Not Close)
   a. Via: How often, to your knowledge, do you have direct contact (in person, by phone, online) with transgender people who are not close to you, such as coworkers or acquaintances?
   b. I plan to code an indicator equal to 1 if the respondent selected anything other than 'Never' when answering this question *and* answered 'Never' to the question about contact with transgender people who are close to them
   c. I do not have a strong prior for this moderator: on the one hand, contact with transgender people may foster understanding and acceptance; on the other hand, it may increase the likelihood that individuals recognize individuals as transgender, thereby enabling identity-based discrimination
7. Respondent Moderator: Distance to Downtown (for German respondents only)
   a. Via: Which best describes the area where you live? [Downtown → Rural]
   b. I plan to code a variable representing distance to downtown, which increases for every [Downtown → Rural] value: Downtown (city centre) [-0.5], City (urban, but not downtown) [0], Suburban (outside the city) [1], Rural (small town or countryside) [2]
   c. Given results from the field experiment, my prior is that individuals located closer to the city center exhibit larger treatment penalties
8. Respondent Moderator: Postsecondary Education
   a. Via: Please select the highest level of education you have completed.

[4] Responses in the upper half of the scale may not be comparable across respondents. For example, one respondent who accepts transgender people but does not feel *especially* positively toward this group may select 4; another person with the same attitude may select 7.

[5] That includes responses: "A few times a year," "About once a month," "About once a week," "Daily or almost daily"

    b. I plan to code an indicator equal to 1 if the respondent answered 'Bachelor's degree' or 'Graduate degree (Master's, Doctorate, or professional degree)'
    a. My weak prior is that respondents with postsecondary education will exhibit smaller treatment penalties, although I primarily include this moderator because education is a common and substantively relevant dimension of heterogeneity across many analyses— there is evidence that higher education level is associated with more positive views toward transgender people (Norton and Herek, 2013), though evidence is mixed (Luhur et al., 2019)
9. Respondent Moderator: High Income
    a. Via: What is your annual household before tax income?
    b. I plan to code an indicator equal to 1 if the respondent answered '$100,000-$150,000' or above (in the United States); '€5,500-€7,300' or above (in Germany)
        - Because it is more common to ask about after-tax monthly income in Germany, that question is used in the German survey. Income cutoffs were aligned based on country-specific income distribution
    c. I have no strong priors about the relationship between income and treatment penalty magnitude. I primarily include this moderator because income is a common and substantively relevant dimension of heterogeneity across many analyses—evidence is mixed with some small positive (Flores et al., 2016) or null associations (Luhur et al., 2019)
10. Respondent Moderator: Male Dominated Occupation
    a. Via: In your occupation (or most recent occupation), in general, would you describe the workforce as… [Mostly Men → Mostly Women]
    b. I plan to code an indicator equal to 1 if the respondent answered 'Mostly Men (for example, ~7 out of 10 are men)'
    c. My prior is that respondents who work in male-dominated occupations exhibit larger treatment penalties—given evidence from Granberg et al. (2020) about transgender discrimination in Sweden, and evidence that conformity to gendered norms is higher in male-dominated occupations (Milner et al., 2018)
11. Respondent Moderator: Customer Service Occupation
    a. Via: In your job (or most recent job), how often do you interact with customers/clients/patients as a core part of your work?
    b. I plan to code an indicator equal to 1 if the respondent answered 'Daily'
    c. I have no strong priors about whether respondents who work in customer service occupations exhibit smaller or larger treatment penalties
12. Profile Moderator: Male Dominated Occupation
    a. Via: base profile
    b. I plan to code an indicator equal to 1 if the profile is in a male dominated occupation: Warehouse Associate, Commercial Truck Driver, Industrial

Maintenance (in the United States) or Fachlagerist/-in, Berufskraftfahrer/-in, Mechatroniker/-in (in Germany)

c. My prior is that respondents will exhibit smaller treatment penalties when workers apply in male-dominated occupations—respondents may perceive transgender women, especially those who do not pass, as possessing traits stereotypically associated with male-dominated work, such as physical strength or mechanical aptitude

13. Profile Moderator: Customer Service Occupation
    a. Via: base profile
    b. I plan to code an indicator equal to 1 if the profile is in a customer service occupation: Office Clerk, Nurse, Sales Associate, Server (in the United States) or Bürokaufmann/-frau, Pflegefachkraft, Verkäufer/-in, Servicekraft (in Germany)
    c. My prior is that respondents will exhibit larger treatment penalties when workers apply in customer service occupations—given complementary field experiment results, and evidence that physical appearance is more strongly rewarded in jobs involving substantial interpersonal interaction (Stinebrickner et al., 2018)
14. Profile Moderator: Politically "Right" Location
    a. Via: profile location
    b. I plan to code an indicator equal to 1 if the profile lists a politically right-leaning geography
    c. My weak prior is that respondents will exhibit smaller treatment penalties when workers are located in right-leaning locations— conditional on having the same work experience, a transgender woman working in a politically right-leaning area may be perceived as especially resilient, capable, or professionally successful because respondents may view that environment as less accepting of transgender people in general

F. Linked Field Experiment Analyses

First, in the original pre-analysis plan, I assigned party-level "trans-progressiveness" scores to German political parties as follows: Green = 10, Die Linke = 9, SPD = 8, Union = 3, and AfD = 0. I then used these scores to construct geography-specific trans-progressiveness measures by taking the weighted average of party scores based on local party support in the 2025 federal election. Because the original party scores were rooted in subjective judgement rather than empirically estimated, I will use Survey Experiment 2 to construct an alternative party-level measure based on respondents' self reported social conservatism. Specifically, I will calculate the average response to the 1–7 *socially conservative* item among respondents who report supporting that party most. This will produce a party-level score, where higher values indicate that supporters of that party are, on average, more socially conservative. I will use this measure as a broader, empirically grounded alternative to the original progressiveness coding. Note: to avoid

constructing party scores from very small cells, I will calculate these party-level social conservatism scores only for parties supported by at least 30 German respondents. Parties with fewer than 30 supporters in the survey will be excluded from the geography-specific measure (and local vote share will be re-scaled such that it sums to 1).

Second, because this survey and the field experiment use the same images, I plan to link image-level *physically attractive* and *professionally presented* measures estimated from German survey respondents to the field experiment data. This will allow me to assess how much of the estimated penalty associated with non-passing appearance is statistically associated with these survey-rated appearance measures. For both Method 1 and Method 2 below, I will add the two image-level measures to the primary field experiment specification compare the estimated coefficients on NP_i and $DT_i \times NP_i$ before and after adding them. I will use the Gelbach (2016) decomposition to allocate the resulting attenuation across *physical attractiveness* and *professional presentation*. The extent to which each appearance measure attenuates the estimated coefficients will be informative about how much of the estimated penalty is associated with differences in these two appearance measures.

Method 1: Presumably cisgender man versus cisgender woman

This method uses the difference in perceived *physical attractiveness* and *professional presentation* between the presumably cisgender man and presumably cisgender woman images in the same triplet. It is intended to capture differences in appearance between the male and female images within the same triplet, while excluding rating penalties associated with the gender expression of non-passing transgender women.

For each field experiment image, I will assign image-level fixed effects for *physical attractiveness* and *professional presentation* estimated in the survey experiment using the following specification and including German respondents only:

$$(7) \quad A_i = X_i'\beta + \eta_{r(i)} + \tau_{b(i)} + \vartheta_{h(i)} + \kappa_{o(i)} + \pi_{a(i)} + \varepsilon_i$$

where A_i denotes either the *physically attractive* rating or the *professionally presented* rating; I estimate (7) separately for each attribute. Compared to specification (4), specification (7) differs in two ways. First, the regression is estimated using control observations only; hence, indicators DT_i and NP_i are excluded. Second, the triplet fixed effect $\theta_{t(i)}$ is replaced with an image-specific fixed effect $\vartheta_{h(i)}$. That is, specification (4) controls for triplet-specific fixed effects, where each triplet contains a presumably cisgender woman image, a presumably cisgender man image, and a non-passing transgender woman image. By contrast, specification (7) estimates separate fixed effects for each image.

For passing images, I will use the fixed effect for the presumably cisgender woman image. For non-passing images, I will use the fixed effect for the presumably cisgender man image from the same triplet. Because the field experiment regressions include twin fixed effects, this specification identifies the relevant within-triplet attribute contrasts directly.

Method 2: Non-passing transgender woman versus presumably cisgender woman

This method is analogous to Method 1, except that it uses the difference in perceived *physical attractiveness* and *professional presentation* between the non-passing transgender woman and presumably cisgender woman images within each triplet. It therefore captures both differences in appearance between the images and any rating differences associated with the gender expression of transgender women who do not pass.

For each field experiment image, I will assign image-level fixed effects for *physical attractiveness* and professional presentation estimated in the survey experiment using the following specification and including German respondents only:

$$(8) \quad A_i = \delta DT_i + X_i'\beta + \eta_{r(i)} + \tau_{b(i)} + \vartheta_{h(i)} + \kappa_{o(i)} + \pi_{a(i)} + \varepsilon_i$$

where A_i denotes either the *physical attractiveness* rating or the *professional presentation* rating; I estimate (8) separately for each attribute. Compared to specification (7), specification (8) is equivalent except for the following modifications. First, the regression is estimated using control observations and all observations where $NP_i = 1$. Second, the disclosure indicator DT_i is included so that the estimated image-specific fixed effects for non-passing images are not conflated with the effect of male-to-female name-change disclosure. The indicator NP_i is excluded because the relevant non-passing variation is captured directly by the image-specific fixed effects.

For passing images, I will use the fixed effect for the presumably cisgender woman image. For non-passing images, I will use the fixed effect for the non-passing transgender woman image from the same triplet. Because the field experiment regressions include twin fixed effects, this specification identifies the relevant within-triplet attribute contrasts directly.

G. Supplemental Analyses

I plan to report descriptive summary statistics for key respondent characteristics, including how respondents who report negative feelings toward transgender people differ from those who do not and how respondents who have direct contact with transgender people (both "close" and "not close" to them) differ from those who do not.

As a validation and benchmarking exercise, I will compare the direction and relative magnitude of survey-based *interview offer* ratings across treatment arms to the corresponding field experiment estimates via specification (1). The field experiment provides the more credible estimate of actual hiring discrimination, so this comparison can provide insight on the extent to which survey results may be affected by social desirability bias and the artificial rating context.

As a supplemental check on the pooled results, I will report country-specific estimates for the primary results emphasized in the paper, estimating the same specifications separately for respondents in Germany and respondents in the United States. These estimates are intended to

assess whether the pooled results are directionally similar across countries or whether they are driven primarily by one country.

In addition, I will assess whether the "woke" hobby signal increases ratings on the *raise concerns* attribute. This serves as a validation check for whether the *raise concerns* attribute captures perceptions related to workplace-relevant "woke"-ness and whether the survey design generated meaningful variation in this perceived attribute.

Finally, following Muralidharan et al. (2025), whenever I report estimates from the "short" models, equations (2) and (3), I plan to also report estimates from the corresponding "long" models. The main difference between these short and long models is whether the interaction term $[DT_i \times NP_i]$ is included. Importantly, omitting this interaction term changes the interpretation of the coefficients on DT_i and NP_i. In the long models, these are conditional effects: the coefficient on DT_i is the effect of male-to-female name change disclosure among *passing* profiles, while the coefficient on NP_i is the effect of non-passing appearance when transgender identity is *not* disclosed. In the short models, by contrast, these coefficients are composite effects. The coefficient on DT_i averages the effect of disclosure across passing and non-passing profiles, while the coefficient on NP_i averages the effect of non-passing appearance across disclosed and undisclosed profiles. When the interaction is imprecisely estimated, I view these weighted average effects as preferable, because disclosure (or similarly, someone's transgender identity being known) and non-passing appearance occur jointly in practice. If I cannot estimate all three components precisely, I would rather focus on relevant weighted averages than focus only on conditional effects.

H. Robustness Checks

The primary analysis will exclude respondents who do not complete the survey or who answer incorrectly on two or more attention check questions. Each respondent answers between two and four attention check questions; under the randomization logic, approximately 36%, 50%, and 14% of respondents are expected to answer two, three, and four attention check questions respectively. As robustness checks, I will re-estimate the primary results emphasized in the paper using: (i) all completed responses regardless of attention check performance, and (ii) a stricter sample excluding any respondent who answers incorrectly on at least one attention check question.

As an additional robustness check, I will re-estimate the primary results emphasized in the paper among respondents with hiring experience. This assesses whether the main findings are similar among respondents who have previously been involved in hiring-related decisions. I will also re-estimate the primary results emphasized in the paper using cisgender women controls only (i.e., omitting cisgender men).

I also plan to re-estimate equation (4) including all eight mechanism-related attributes simultaneously, rather than estimating a separate regression for each attribute. In this specification, the model will include the standardized rating for each attribute and its interactions

with DT_i and NP_i. I view this as a robustness check rather than the preferred specification because many of the attributes are likely correlated with one another, which may influence standard errors and make individual coefficient estimates unstable. However, this specification is still informative because it shows which attributes remain associated with *interview offer* ratings after conditioning on the other measured attributes.

References

BabyMed (n.d.). *Top German Baby Names in the 1990s.* Retrieved February 20, 2026 from https://www.babymed.com/baby-baby-names/top-german-baby-names-1990s

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Billard, T. J. (2018) Attitudes Toward Transgender Men and Women: Development and Validation of a New Measure. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00387

Flores, A. R., Brown, T. N. T., & Park, A. S. (2016, December). *Public support for transgender rights: A twenty-three country survey*. The Williams Institute. https://williamsinstitute.law.ucla.edu/wp-content/uploads/Public-Opinion-Trans-23-Countries-Dec-2016.pdf

Forebears (n.d.). *Most Common Last Names in Germany*. Retrieved February 20, 2026 from https://forebears.io/germany/surnames

Gelbach, J. B. (2016). When do covariates matter? And which ones, and how much? *Journal of Labor Economics, 34*(2), 509–543. https://doi.org/10.1086/683668

genderize.io (n.d.). *Check the Gender of a Name*. Retrieved February 20, 2026 from https://genderize.io/

Granberg, M., Andersson, P. A., & Ahmed, A. (2020). Hiring discrimination against transgender people: Evidence from a field experiment. *Labour Economics, 65*, Article 101860. https://doi.org/10.1016/j.labeco.2020.101860

Luhur, W., Brown, T. N. T., & Flores, A. R. (2019, August). *Public opinion of transgender rights in the United States: 2017 Ipsos international survey series*. The Williams Institute, UCLA School of Law. https://williamsinstitute.law.ucla.edu/wp-content/uploads/Public-Opinion-Trans-US-Aug-2019.pdf

Milner, A., Kavanagh, A., King, T., & Currier, D. (2018). The influence of masculine norms and occupational factors on mental health: Evidence from the baseline of the Australian longitudinal study on male health. *American Journal of Men's Health, 12*(4), 696–705. https://doi.org/10.1177/1557988317752607

Muralidharan, K., Romero, M., Wüthrich, K. (2025). Factorial designs, model selection, and (incorrect) inference in randomized experiments. *The Review of Economics and Statistics; 107 (3): 589–604*. https://doi.org/10.1162/rest_a_01317

Napier, J. L. (2025). A Cross-Cultural Investigation of Prejudice Against Transgender People. *Social Psychological and Personality Science*, *16*(6), 599-609. https://doi.org/10.1177/19485506241289638

Norton, A. T., & Herek, G. M. (2013). Heterosexuals' attitudes toward transgender people: Findings from a national probability sample of U.S. adults. *Sex Roles, 68*(11–12), 738–753. https://doi.org/10.1007/s11199-011-0110-6

Parker, K., Horowitz, J. M., & Brown, A. (2022, June 28). *Americans' complex views on gender identity and transgender issues*. Pew Research Center. https://www.pewresearch.org/social-trends/2022/06/28/americans-complex-views-on-gender-identity-and-transgender-issues/

Rainey, C. (2026, January 31). *Power rules: Practical advice for computing power (and automating with pilot data)*. https://www.carlislerainey.com/power-rules/power-rules.pdf

Social Security Administration (2025). *Popular names of the 1990s.* Retrieved February 20, 2026 from https://www.ssa.gov/oact/babynames/decades/names1990s.html

Stinebrickner, T. R., Stinebrickner, R., & Sullivan, P. J. (2018). *Beauty, job tasks, and wages: A new conclusion about employer taste-based discrimination* (NBER Working Paper No. 24479). National Bureau of Economic Research. https://doi.org/10.3386/w24479

US Census Bureau (2021). *Frequently occurring surnames from the 2010 census*. Retrieved April 5, 2023. https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

Van Borm, H. & Baert, S. (2018). What drives hiring discrimination against transgenders? *International Journal of Manpower, 39(4)*, 581-599. https://doi.org/10.1108/IJM-09-2017-0233