Discrimination by Elected Officials against Disadvantaged Group: A meta-analysis

Abstract

Responsiveness to citizens constitutes a fundamental feature of representative democracy. A prominent design for assessing potential biases in responsiveness are correspondence experiments in which (putative) citizens task legislators to answer to their requests. Differences in the response rate are then attributed to a particular characteristic randomized in the design. This meta-analysis focuses on two sources of discrimination: gender and ethnicity of a constituent. After screening over 2,500 studies, the analysis includes 33 audits entailing 57 experiments in 23 countries. Two substantive findings stand out: the response rate is 1.5 higher for female than male constituents, and constituents from ethnic minorities had a 4.4 percentage points lower response rate than those from a majority. Additional tests relating to design features, such as message content, level of representation, study quality, are largely inconclusive. Our study concludes with a critical appraisal of the aggregate insights from correspondence studies for research on political representation.

1 Introduction

> Equal activity is crucial for equal consideration since political activity is the means by which citizens make their needs and preferences known to the governing elites and induce them to be responsive.
>
> Verba (2003, p. 663)

Through political activities, such as voting, donating, and participating in public discourse, citizens make their voices heard and have their interests considered by decision-makers. Citizen action and parliamentary reaction influence whose interests are aggregated and in which way. Voting as political activity is aggregated by clear rules, ensuring - at least in principle - that each vote has the same weight. In other words, gender, wealth, or occupation of a constituent do not matter, as all votes count the same. With other activities, such as donating or writing to your representative, equal consideration is not or less formalized. Thus, the impact of political activity both absolute and in comparison with other people is far less clear. As elections are carried out in long intervals, the aggregation of citizens' interests through other means between elections are a crucial feature of democratic politics.

Robert Dahl postulates that ideal democracies should be responsive to citizens and treat them as political equals (R. A. Dahl 1998). This research focuses on *service responsiveness* as the component of responsiveness concerning "non-legislative services that a representative actually performs for individuals or groups in his district" (Eulau and Karps 1977, p.243). Service responsiveness comprises for instance responding to letters from constituents or attending personal meetings with citizens. A common concern about these activities is: Do politicians discriminate against groups of constituents by being less responsive to them? And if so, are politicians with a shared identity of these groups less biased?

Over the past decade, experimental designs became increasingly popular in political sciences (Druckman and Green 2021) in general and in addressing this question of unequal responsiveness in particular. 'Audit studies' or 'correspondence studies' are the most common tool for

researching the response of political elites. They are a specific type of field experiment in which correspondence, in most cases an email, from a (putative) citizen is sent to a legislator. In the correspondence, a certain aspects or features is manipulated and randomized. For example, studies randomize the name of the sender, signalling differences in gender or ethnicity of the putative citizen. Other studies alter the content of the correspondence, signalling party membership, educational background or asking about different issues. Given the design, researchers then attribute differences in reactions, such as in the response rate or time, to the randomized treatment. An extensive discussion (Campbell and Bolet 2021; Allegretti 2021; Naurin and Öhberg 2021) centers around organizational and ethical concerns regarding audit studies with politicians.

This paper focuses on audit studies with elected officials, such as Members of the Parliament (MP), introduced to political sciences in the seminal paper of Butler and Broockman (2011), based on the approach taken first by Putnam, Leonardi, and Nanetti 1993. For two reasons, we do not just conduct another audit study to answer the research question, but instead systematically summarize relevant existing studies in a meta-analysis. First, a sufficiently large number of audit studies with politicians exist and a subset of them focus on discrimination (e.g. Butler and Broockman 2011; Butler 2014). However, only two attempts (Costa 2017; Peters 2018) have been undertaken to aggregate the findings of these studies. Our goal is to provide a comparative assessment of many audits for two particular types of discrimination—gender and ethnicity—and provide a quantitative estimate of its degree. Second, increasing ethical concerns (Desposato 2015; John and Foos 2022; Bischof et al. 2022; Loewen and Rubenson 2022) about the reasonableness of deception and administrative burden in the political context have been voiced. Given these valid objections, a collective assessment of audit experiments provides an assessment of how much discrimination one could expect. This estimate can then be used in cost-benefits calculations for future research. Overall, aggregating existing findings before conducting further experiments appears to be both more relevant and more appropriate.

This review focuses on two relevant dimensions of discrimination: gender and ethnicity. On both dimensions, discrimination has been documented and tested in a sufficient amount of origi-

nal research for a systematic review. For example, research has shown that men are more likely to be recruited by party leaders (Crowder-Meyer 2013), that political parties are more responsive to women (Homola 2019), and that voters penalize female candidates (Teele, Kalla, and Rosenbluth 2018). Investigating gender-based discrimination in service responsiveness thus follows the reasonable ground that women and men are not treated equally by their representatives. Similarly, it is an established finding that ethnic minorities suffer from discrimination in various spheres of life, from the job market to welfare benefits and public housing (Kaas and Manger 2012; Hemker and Rink 2017; Einstein and Glick 2017). Case studies, including audit experiments, have found political representation to suffer from ethnicity-based discrimination, but no review has aggregated these findings systematically.

The contribution of this research endeavor is threefold. First, a meta-study of audit experiments aggregates well-designed individuals studies and offers external validity to studies on responsiveness. The systematic review identifies common findings and explains variances in results, e.g. between designs and contexts. The meta-analysis quantifies potential discrimination with higher certainty by combining the statistical power of single studies. Second, this review provides an systematic summary of the scientific state of the field on gender- and ethnicity-based discrimination in service responsiveness. In contrast to most literature reviews, the literature search is systematic, seeking to identify and evaluating all relevant studies. We follow the guidelines of the Cochrane Handbook for Systematic Reviews of Interventions both in procedure and reporting (Higgins et al. 2022). Third, the review informs politicians, bureaucrats, and the civil society about the extent of discrimination.

2 Responsiveness and discrimination in the literature

In her seminal work, Hanna Pitkin defines *accountability* as one dimension of formalistic representation. It describes the *responsiveness of the representative* to the constituents, as well as the ability to punish the representative for not acting in the interest of the constituents (Pitkin 1967,

p.55). In this view, the key responsibility and quality of a representative is to aggregate and act on behalf of the interests of the constituents, in other words to 'look after his constituents, or do what they want' (Pitkin 1967, p.57). To be able to 'do what they want', a representative must listen to those being represented. Legislators should be able to, willing to, and effectively do responsively communicate with their constituents. The responsiveness to inquiries of constituents thus constitutes a key part of formalistic representation. This formalistic view is also reflected in legal frameworks (e.g. German Grundgesetz Art. 38). Most representative democracies prominently manifest the idea that legislators listen to, understand, and represent the will of the entire people, not just parts of the electorate.

Discrimination in this communication would constitute a major threat to functioning representation. Here, the importance of responsiveness relates to descriptive representation (Mansbridge 2015) where descriptive features such as the physical appearance, lived experience, or background of a legislator can influence the trust of groups sharing or not sharing these features. Given that parliaments in most liberal democracies have low levels of descriptive representation (Stockemer 2015; Casellas and Wallace 2015), biases in responsiveness can exacerbate descriptive inequalities.

The research question whether, to what extent, and in which contexts representatives display bias in responsiveness to constituents is relevant to theoretical concepts of representation in two ways. First, biased responsiveness is a threat to the *accountability* of representatives towards groups of constituents, limiting a key aspect of representation. Second, such bias is an indicator and source of *communicative mistrust*, leading to biased representation of the interests of certain subgroups. This systemic review focuses on two descriptively underrepresented groups across parliaments in liberal democracies: women and ethnic minorities (Stockemer 2015; Casellas and Wallace 2015).

The review concentrates on correspondence (or audit) studies, which have been used for assessing discrimination in various settings since the 1980s. Since around 2010 (e.g. Butler, Karpowitz, and Pope 2012), these audits have been employed for studying politicians' biases in re-

sponsiveness to their constituencies. The central merit of audits is their ability to detect the effects of constituent characteristics on responsiveness without concern about variations in the willingness and capacity to participate in politics. Audit experiments randomly assign particular characteristics. Thereby, at least that is the hope, they offer design-based assurances of unbiased assessments of discrimination.

While this design certainly has limitations, it lends itself to a systemic review because these studies follow an almost canonical research design. In particular, response rate – the extent to which elected representatives answer constituency requests – are common outcome measures. A meta-analysis allows us to assess and provide a general assessment of discrimination across multiple studies and contexts. More importantly, we also explore heterogeneity in effect sizes and how they relate to important contextual or design features.

Only two attempts to systematically review and summarize the findings of audit studies on elected officials exist so far. A meta-analysis by Costa (2017) and a qualitative overview by Peters (2018). We briefly summarize their findings and name some limitations.

Costa (2017) aims to analyze all experiments on elite responsiveness and constituent communication, including survey and audit experiments. The main focus is on responsiveness per se, and ethnic discrimination is only one aspect across all reviewed studies. Relevant for this review, the response rate to requests by a member of an ethnic minority is 10% (significantly) lower than for white people. The review mostly focuses on works in American politics, assesses various types of treatments and designs, and does not follow guidelines or best practices for systematic reviews (e.g. Higgins et al. (2022) or Shea et al. (2017)). As such, a comparative assessment of one canonical design focusing on gender and ethnicity becomes warranted.

Peters (2018) provides a summary and research agenda on representation, distinguishing between differential representation by gender, ethnicity, and income. Her focus is on a systematic categorization of representation studies more generally, considering responsiveness only as one aspect out of many. Peters finds some connection between the gender of MPs and their ideologies. Substantial variance between studies, from no effects to mixed effects to strongly gendered

effects, exist, which makes it challenging to gauge the overall effect and its size. For ethnicity, Peters (2018) finds that preferences of whites are better represented than those of minorities. Because only five studies on this subject have been identified at the time, the generalizability is limited. Peters (ibid.) specifically calls on future research to conduct meta-analyses to estimate the size of differential or unequal representation. Our goal is to expand and complement the existing reviews by providing a systematic summary of the state of research using meta-analysis.

3 Research Design

This section describes in detail how we identified relevant experiments, which data was extracted and which tools were employed for the analysis. A number of relevant studies included several experiments. In these cases, the inclusion or exclusion was decided for each experiment separately.

3.1 Selection Process

The study selection follows the PICO-Framework by the Cochrane Collaboration (McKenzie et al. 2021) and consists of four steps: identification, screening, eligibility, and inclusion. Studies matching the following criteria were included: (1) the subject of the experiment are elected officials, (2) the intervention constitutes actual correspondence to the legislator altering either the gender, ethnicity or both features of the putative constituent, (3) the design is experimental with at least one treatment and control group, and (4) the study measures at least the response rate and possibly other response criteria (e.g. response time, quality of response). Even though audit studies in representation research became common only over the last decade, there was no time or geographic limit set for the search. Studies from all countries are included, independent of the level of democracy, as long as the subjects were elected officials.

The search strategy was guided by the idea to capture published as well as non-peer-reviewed experimental designs. With this goal in mind, 16 databases, listed in Appendix A were searched.

All of them, with the exception of WoS, were accessed collectively via ProQuest. A twofold approach made it possible to identify search terms with both high sensitivity and precision: a backward search based on relevant keywords and a forward search based on early, seminal papers. The exact search terms and results can be found in the Appendix. The final results from the databases were retrieved on February 5th 2022. After de-duplicating, 2,480 studies remained. In addition to the systematic search, a general search using the citation network of seminal studies yielded an additional 59 experiments.

Through several steps with structured screening and assessment, the relevant studies are selected from all studies identified in the search. Following the PRISMA guidelines, Figure 1 displays the schematic overview of the entire selection process (Moher et al. 2009). In the screening step, the title and abstract of all 2,480 studies were inspected. 111 studies remained in our sample. For these cases, the full text was assessed with regard to the inclusion criteria, documenting the reason for exclusion.[1] Most excluded studies did not report an audit study at all or one with a different target, different goal, or design. Re-analyses of already published experiments were excluded, as well as studies in languages other than English, German, and French. In the end, 33 studies were eligible for inclusion. They contained 66 experiments (29 randomizing gender, 37 randomizing ethnicity), with some studies reporting several experiments and some experiments randomizing both gender and ethnicity. Excluding experiments for which the relevant data was not available, the meta-analysis on gender includes 27 experiments, the meta-analysis on ethnicity 32 experiments.

3.2 Information retrieval and measurement

For the subsequent meta-analysis, several pieces of information were extracted from the studies for each experiment. While most studies report discussed data directly, some were retrieved from model summaries or computed independently based on replication material.[2] As background in-

[1] The documentation of the reasons can be found in the supplementary materials.

[2] For one study targeting political candidates and elected officials, raw data only for the elected officials was used (Driscoll et al. 2018).

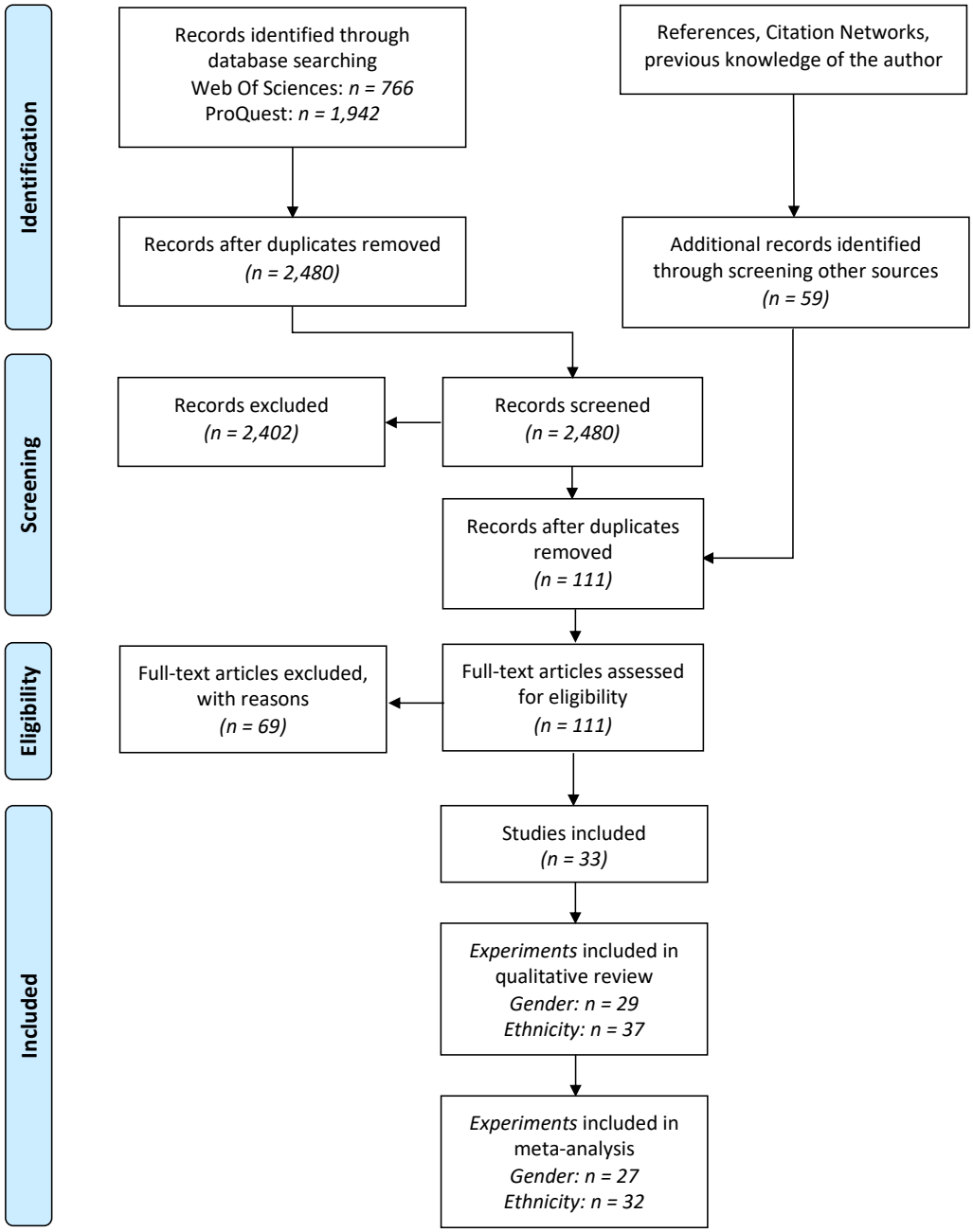

Figure 1: FLOW-Chart of Study Selection Process

formation, we obtained: the total number of legislators contacted, the absolute number of emails sent per treatment group and either the absolute number of responses received or the relative response rate in the treatment and control group.

The outcome variable and the treatment are the two key pieces of information from each experiment. A reasonable outcome measure used in a meta-analysis should fulfil two criteria: it should (1) measure the concept and (2) bed used in all experiments. The *response rate* does that and is computed as the quota of replies received from legislators in a given time and the total number of emails sent. Across experiments, there is some variation along these dimensions, such as cut-off time and definition of response. Due to lacking comparability, further outcome measures are not included in this meta-analysis here but discussed later.

Some differences regarding the treatment groups are important to mention. For gender, all studies randomize simply between male and female aliases. No study uses gender-neutral names or conceals the gender of the putative constituent. For ethnicity, comparison is more difficult since the treatments in different experiments signal different ethnicity of the putative constituent, varying between countries and experiments. For instance, an experiment in the US randomized between names perceived as Latino and White, a study in Germany between Turkish and German names (Alizade and Ellger 2021; Janusz and Lajevardi 2016). Treatments signalling ethnicity are re-coded as minority or majority before the analysis. we refrain from using the labels autochthonous (majority, native) and allochthonous (minority, non-native), as in some cases, the autochthonous name stems from people now constituting a minority, for instance indigenous Brazilians. All but one of the ethnicity treatments were name-based, which, given the debate on this issue (Landgrave and Weller 2022), we return to in the discussion.

Five contextual variables were extracted for the meta-analysis regression: message content, level of government, type and quality of design, and publication status. For message content, we distinguish between *policy, service and career* requests. Policy requests contain a question about the position or activities of an MP towards a specific policy. Service requests seek general information, e.g. on voter registration, and career requests ask for advice for going into politics,

Type	Example	Reference
Policy	'[...] I know the federal budget has a lot of money for transit infrastructure but I want to know what you are doing to specifically improve options in [Constituency]. Is this an issue that you are interested in? Who should I speak with about my concerns? [...]'	Hancock 2018
Service	'[...] I lost my job and I don't know what to do. What should I do to get unemployment benefits? No one will tell me where to go. [...]'	Magni and Leon 2021
Career	'[...] My name is [...] and I am a college sophomore. I'm interviewing politicians for a class project to learn about how they entered their field and what advice they might have for students interested in politics. As someone who really cares about my community, one day I hope to be a politician. What advice would you give to me? [...]'	Kalla, Rosenbluth, and Teele 2018

Table 1: Examples of requests by type

e.g. an internship or general advice. Table 1 displays examples.

The level of government of the sample was measured on a three-level scale: local, state, and federal. Some experiments were fielded on more than one level. The research design distinguishes 'between'-designs, in which treatment allocation differs between legislators. Some studies sent several emails to each legislator with different treatment conditions ('within'-design). Few studies use a combination of both to increase power in certain subgroups (e.g. Loewen and MacKenzie 2019).

As a measure of quality, encompassing various aspects related to internal and external validity, we used the second version of the Cochrane Risk-of-Bias tool (RoB) in order to assess the risk of bias in five different dimensions: the randomization process, deviations from the intended intervention, missing outcome data, measurement of the outcome, and selection of the reported results (J. A. C. Sterne et al. 2019). The evaluated distinguishes between low risk of bias, some concerns, and high risk of bias. Finally, we recorded if the study was published or not.

3.3 Meta analysis

Our main analytical strategy comprises a separate meta-analysis of response rates for treatments on gender and ethnicity, which we follow with a discussion on co-gender and co-ethnicity effects. The meta-analysis contains only studies for which results could be retrieved in the necessary format. At least the size of each treatment group and either the relative response rate or the absolute number of responses in each group are necessary to compute a comparable effect size with uncertainty measures. Since studies report different effect measures, e.g. logistic regression coefficients, we reconstructed the effect size for each experiment as the difference in response rates between treatment groups:

$$\Delta rr = \frac{responses_{female}}{n_{female}} - \frac{responses_{male}}{n_{male}} = rr_{female} - rr_{male} \tag{1}$$

where for each treatment group, $responses$ denotes the absolute number of responses, and n the absolute number of emails successfully sent with the respective alias. The calculation for experiments on ethnicity follows the same logic. The absolute number of the successfully sent emails is used as the denominator in order to capture the *per-protocol effect*. The main difference to the intention-to-treat effect is that observations in which the treatment failed are excluded from the analysis. A common failure of treatment is non-functioning emails. Less commonly, MPs detect the experiment or spillover problems occur when MPs share staff. Limitations of the per-protocol-effect are considered in the discussion.

The unit of analysis of the meta-analysis are experiments, not studies. If more than two treatment groups with different ethnic signals are used, each pair of majority-minority treatments is considered as a separate observation. For instance, Landgrave and Weller (2022) use White, Black, and Hispanic aliases, which are considered as two separate observations here, one between White and Black aliases, and one between White and Hispanic aliases. The fact that these observations are not independent of each other is discussed below. The uncertainty of the effect is calculated for each experiment using the standard error formula for proportions:

$$SE_{rr_{female}} = \sqrt{\frac{rr_{female} \times (1 - rr_{female})}{n_{female}}} \tag{2}$$

$$SE_{\Delta rr} = \sqrt{SE^2_{rr_{female}} + SE^2_{rr_{male}}} \tag{3}$$

$$CI_{\Delta rr} = \Delta rr \pm 1.96 \times SE_{\Delta rr} \tag{4}$$

where the standard error of the difference in response rate $SE_{\Delta rr}$ is the combined standard error of the response rate in each treatment group $SE_{rr_{female}}$ and $SE_{rr_{male}}$. It would be more appropriate to calculate clustered standard errors, e.g. on alias-level. However, the raw data is not available for most studies. For the sake of inclusion and comparability, regular standard errors are calculated for all experiments.

The combined effect is calculated using a random-effects model, and not a fixed-effect models (Hedges and Vevea 1998; Borenstein et al. 2010). A random-effects model is more appropriate because it is plausible that the true underlying gender and ethnicity bias varies between countries and context factors of the studies. The consequences compared with a fixed-effect model are that (1) the confidence intervals of the overall effect are wider, and (2) that the study weights are more similar (Borenstein et al. 2010). The estimator of our model is calculated using a weighted mean of the difference in response rates of each experiment. The weight is based on the inverse of the overall *study error variance* and depends on two sources of variance:

$$\zeta_i = \theta_i - \mu \tag{5}$$

$$\epsilon_i = \Delta rr_i - \theta_i \tag{6}$$

where the first source of variance, ζ_i, is the difference between the true effect for a study θ_i and the true combined effect for all studies μ. The variance of ζ_i around θ_i is defined as τ^2 and one key metric estimated in the meta-regression later. The second source of variance is ϵ_i, the

difference between the observed effect (the observed difference in response rate Δrr_i) and the true effect for the study θ_i. While ϵ_i decreases with increasing sample size, ζ_i is independent from the experiment parameters and estimated as the excess variation of study-to-study variation that cannot be explained by the expected variation if the true effect was the same in all studies (see p.106 Borenstein et al. 2010). Thus, the weight of a study i is calculated as:

$$W_i = \frac{1}{V_i + T^2} \tag{7}$$

where V_i is the within-study error variance (i.e., the variance of the distribution around θ_i, depending on the sample size n_i), and T^2 is the between-study variance (i.e., the variance of the distribution around μ, which is similar for all studies) (ibid.).

To sum up, the combined effect M is computed as the weighted mean of the individual effects. The standard error of the combined effect SE_M is the square root of the meta-analysis variation, which is the inverse sum of weights.

$$M = \frac{\sum_{i=1}^{k} W_i * \Delta rr_i}{\sum_{i=1}^{k} W_i} \tag{8}$$

$$SE_M = \sqrt{\cdot \frac{1}{\sum_{i=1}^{k} W_i}} \tag{9}$$

The random-effects model allows for variation of the true effect sizes of different experiments, but does not explain this variation (Thomsen and Sanders 2020, p.1571). A further meta-regression incorporates the mentioned context factors of the experiments, such as message content or publication status, in order to investigate the variance of effect sizes. All computations is done in R (RStudio Team 2022; R Core Team 2021) using the package *metafor*. The Restricted Maximum Likelihood estimator is used for the variance of the true effect sizes τ^2 because it is more accurate (Viechtbauer 2010).

4 Results

A descriptive assessment of all studies offers some initial insights about the type, location, and design of correspondence studies. In total, 33 studies reporting 57 experiments were included in the review. Most studies only report one experiment (23), a few entail two (4) or three (5). Magni and Leon (2021) are an exception, reporting experiments on gender bias in 11 countries. Of the studies, 28 are published in journals or books, three are published as thesis (Hancock 2018; Nisser 2017; Kemper 2018) and two are available as working papers (Schakel et al. 2021; Janusz and Lajevardi 2016).

The meta-analysis only includes 50 experiments, for which the relevant quantitative data could be retrieved. With some experiments using a factorial design randomizing both gender and ethnicity, the meta-analysis on gender bias includes 27 experiments, the analysis on ethnicity includes 31 experiments. Almost a third of the experiments (21) were conducted in the US. Other countries with more than one experiment include Canada (6), Germany (5), the UK (3), Brazil (3), Bosnia and Herzegovina (2), and the Netherlands (2). The review includes one experiment each from Argentina, Chile, China, Colombia, Denmark, France, India, Ireland, Italy, Mexico, New Zealand, Sweden, Switzerland, Taiwan, and Uruguay.

Considerable variation in the design of correspondence studies exist. In the majority of experiments, putative citizens asked for information about services (41). Much fewer studies used policy enquiries (9) or career advice (5). The first experiment was conducted in 2008 on US legislators by Butler and Broockman (2011). Before Butler and Broockmann's publication in 2011, six further experiments were already fielded. Almost half of the experiments (27) were conducted in 2017 and 2018, showing the growing popularity of correspondence studies.

Only 26 experiments followed a pre-registered plan, which specified hypotheses, randomization, treatment, outcomes, and analysis in advance. For the remaining 31 experiments, conscious or unconscious decisions of the researchers in the process could have been informed by (preliminary) findings and show other weaknesses in the design. One domain of concerns are baseline differences between the treatment groups that are not investigated. Many studies do not block

on relevant co-variables, such as the party of a legislator, in the randomization process. The second domain of concern are deviations from the intended intervention. Many studies report non-functioning e-mails and not retrieving e-mail IDs for all targeted MPs, but do not investigate adherence problems. Other common concerns include: MPs uncovering the experiment, unreasonable exclusion of observations, and model specifications that are not preregistered or justified in the text.

4.1 Gender

The systematic search identified a total of 29 experiments randomizing the gender of the putative citizen.[3] Figure 2 displays the forest plot from the meta-analysis, whereby positive effects mean a higher response rate for female aliases. The last column contains the difference in response rate and the 95% confidence intervals. The column on the left shows the study in which the experiment was published. Some studies reported results from several experiments. Columns three to four contain information on the country in which the experiment was conducted, the sample size and the type of request as described in section 3.2.

The general trend in the results is clear: Most experiments find that requests from women have a *higher* response rate than requests from men. The difference in response rate ranges from -8.9 to 7 percentage points, with most experiments finding positive effects for women. In the majority of experiments, the effect in favor of women is not statistically significant. Only in four experiments, the difference in response rate is significant at common levels (Kalla, Rosenbluth, and Teele 2018; Crawfurd and Ramli 2021; Lloren 2017; Dhima 2020). All experiments with a significant difference have large sample sizes greater than 1,500 and were conducted in western democracies (US, UK, Switzerland and Canada). Two of the four significant results stem from studies asking for career advice, the other two used a service request.

There are eight experiments showing higher response rates for male than for female senders, with differences as great as -8.9 percentage points (Magni and Leon 2021, see results for Uruguay).

[3]The forest plot includes only the 27 experiments, for which the difference in response rates was available or could be computed.

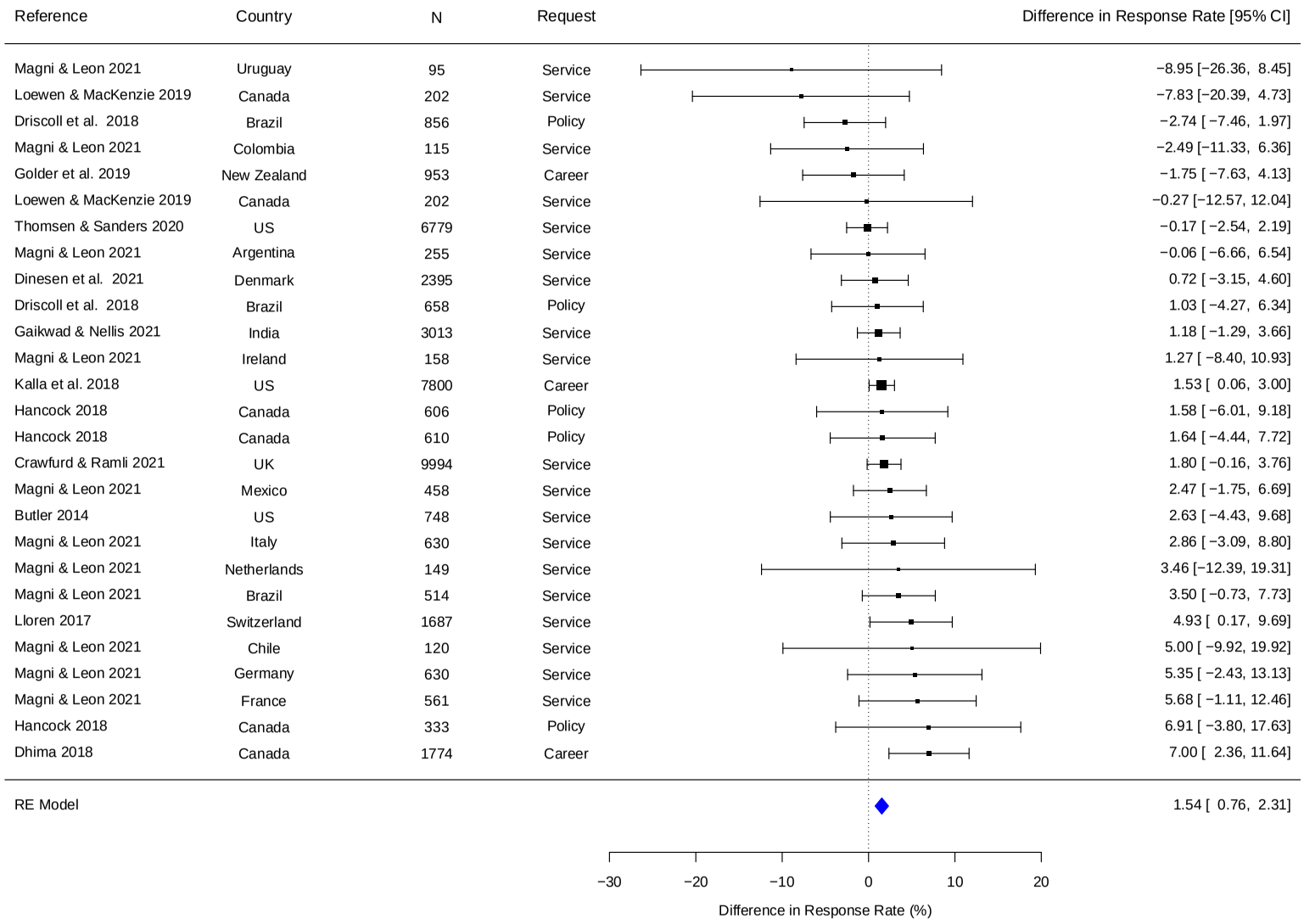

Figure 2: Meta-Analysis Results of Experiments with Gender Treatment

Differences in response rate are given in percentage points. Positive numbers mean that emails sent with female aliases got a higher response rate than emails sent with male aliases. In the forest plot, the square symbols the point estimate, the brackets the 95% confidence interval. The size of the square depicts the weight of the experiment in the random-effects model.

Three of these experiments only find a marginal difference smaller than 0.5 percentage points. None of the findings in favor of men are significant, even the two studies with differences greater than 5 percentage points (Loewen and MacKenzie 2019; Magni and Leon 2021). This could be a statistical power problem, as all studies with results in favor of men have a sample size below 1,000.

In the bottom of the figure, a random-effects model aggregating all studies yields a 1.54 percentage points higher response rate for female than for male senders. The effect is significant at the 0.1%-level. The uncertainty of this effect is rather small: the 95% confidence interval for the effect ranges from 0.76 to 2.31 percentage points. In other words, the results suggest an overall bias in favor of female citizens.

We use a meta-regression with some moderator variables for exploring differences in the described effects across experiments. The results are reported in Table 2. The first model is the basic random-effects model displayed in the forest plot in Figure 2. The second and third model add the content of the correspondence sent and the legislative level of the targeted politicians. The estimates suggest that there is no statistically significant difference whether the correspondence is about a service request, policy position or career advice (baseline category). There is also no difference regarding the political unit of the legislator be it on local, state, or federal level.

When controlling for the publication status and the experiment quality in models (4) and (5), the estimated difference in response rate is even larger than in the basic random-effects model. The quality of an experiment appears to be associated with the outcome. As described in section 3.2, a high value refers to a high Risk-of-Bias, meaning a study with severe quality concerns. The negative and statistically significant estimate for quality suggests that studies with lower quality tend to find smaller differences in response rates. The estimates can be interpreted that actual response rate differences in advantaging female senders could be underestimated.

The publication status of an experiment is not associated with the response rate difference. Because all but one study employ a between-subject design, we do not test for this design difference. Finally, even the country fixed-effects cannot explain much of the variance between the

	Δ Response Rate				
	(1)	(2)	(3)	(4)	(5)
Constant	1.54***	1.21	3.96	5.37*	5.80
	(0.4)	(6.32)	(12.45)	(2.82)	(6.11)
Content: Policy		−1.44	−2.54		
		(2.94)	(6.89)		
Content: Service		0.12	−2.87		
		(4.12)	(7.58)		
Level: Local		0.09	0.70		
		(2.45)	(5.47)		
Level: State		−0.42	−4.93		
		(2.03)	(3.66)		
Level: Federal		0.99	−1.16		
		(2.33)	(3.64)		
Published				−1.20	−0.23
				(2.31)	(3.33)
Quality (RoB)				−1.43*	−2.82**
				(0.85)	(1.43)
Observations	27	27	27	27	27
Country FE	No	No	Yes	No	Yes
R^2	-	0.00	0.00	0.00	0.00
τ^2	0	0.19	2.14	0.49	0.59
	(0.73)	(1.50)	(8.29)	(1.3)	(3.73)

Note: *p<0.1; **p<0.05; ***p<0.01; All estimates are percentages. Standard errors in parentheses. The Quality measure of the study is a three-leveled scale based on the Cochrane Risk-of-Bias tool used in the critical appraisal of each experiment. Level contains three dummy variables denoting whether politicians from the respective level were part of the sample. Many experiments include politicians from several levels.

Table 2: Meta-Regression with Moderators to explain Gender Differences

studies. This estimate confirms the visual analysis of the forest plot: the results for countries with several experiments do not cluster around a country-specific effect (with the caveat that only few countries were investigated more than once).

4.2 Ethnicity

In total, 37 experiments randomizing ethnicity were identified in the systematic search. Figure 3 depicts differences in response rate.[4] The table lists the major ethnicity and the treatment group for each study. The ascriptive type of minority varies among countries, but discrimination applies to nearly all of them. The vast majority of experiments finds that minorities receive a lower response rate. No study yields a significantly higher response rate for minority senders. The random-effects model quantifies this tendency, suggesting that on average, minorities have a 4.36 percentage points lower response rate than senders from the ethnic majority.

Table 3 displays the meta-regression for ethnicity starting with the basic random-effects model summarized in Figure 3.

The meta-regressions focus on five sets of moderators. First, the content of the correspondence, such as service request, policy position or career advice, is unrelated to response rate. The same is true for the level of political unit, from local to national. These two estimates for ethnicity mirror our findings for gender discrimination. In model (4), the estimates suggest that studies using a within-design, i.e. sending several emails to each legislator, produced response rates with statistically significant differences. This estimate should be interpreted carefully though, as only four experiments employ this design. The estimate for published vs unpublished studies also produce statistically significant differences. In contrast to the gender experiments, study quality is not associated with the effect size.

[4]Results from Mendez and Grose 2018, Schakel et al. 2021, and Newland and Liu 2021 could not be included as the difference in response rate was neither provided nor could be computed.

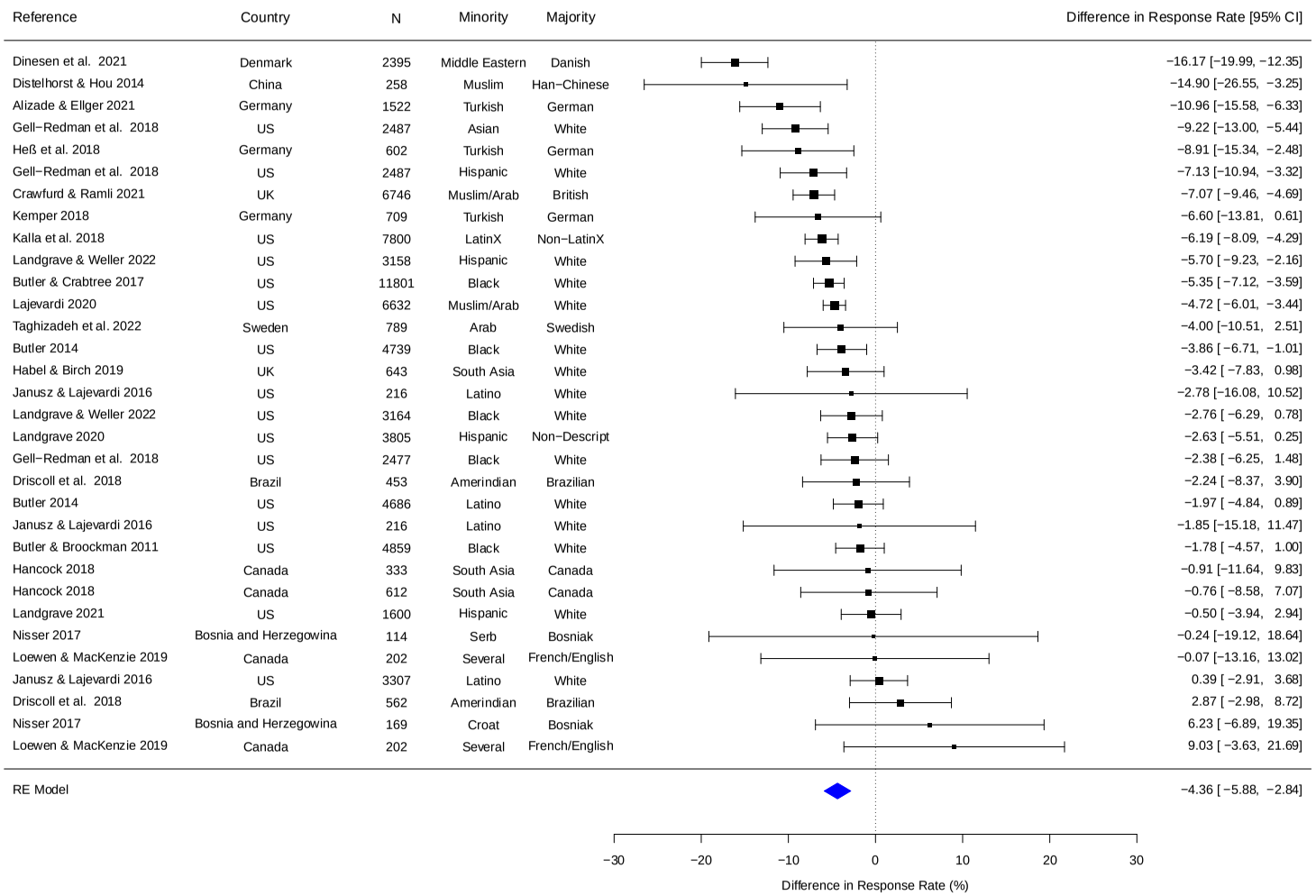

Figure 3: Meta-Analysis Results of Experiments randomizing Ethnicity

Differences in response rate are given in percentage points. Positive numbers mean that emails sent with aliases belonging to an ethnic minority got a higher response rate than emails sent from an alias from an ethnic majority.

	Δ Response Rate				
	(1)	(2)	(3)	(4)	(5)
Constant	-4.36^{***}	-9.09^{**}	2.03	-1.37	4.17
	(0.78)	(4.38)	(7.36)	(3.40)	(5.73)
Content: Policy		4.37	-4.23		
		(3.25)	(6.43)		
Content: Service		1.94	1.45		
		(2.89)	(2.00)		
Level: Local		-2.14	-2.67		
		(2.46)	(1.86)		
Level: State		3.48	-0.68		
		(2.32)	(2.20)		
Level: Federal		2.50	1.27		
		(1.97)	(1.88)		
Design: Within				5.43^{*}	3.24
				(3.12)	(2.20
Published				-4.15^{*}	-3.62^{*}
				(2.28)	(2.03)
Quality (RoB)				0.13	-0.09
				(1.32)	(1.10)
Observations	32	32	32	32	32
Country FE	No	No	Yes	No	Yes
R^2	-	0.24	0.73	0.06	0.78
τ^2	11.63	8.81	3.13	10.92	2.53
	(4.50).	(4.14)	(2.50)	(4.55)	(1.92)

Note: *p<0.1; **p<0.05; ***p<0.01; All estimates are percentages. Standard errors in parentheses. R^2 in decimal. The Quality measure of the study is a three-leveled scale based on the Cochrane Risk-of-Bias tool used in the critical appraisal of each experiment. Level contains three dummy variables denoting whether politicians from the respective level were part of the sample. Many experiments include politicians from several levels. The baseline category for Design is 'between'.

Table 3: Meta-Regression with Moderators to explain Ethnicity Differences

4.3 Co-Gender and Co-Ethnicity

Only few experiments analyze whether legislators are more or less responsive towards citizens belonging to the same ethnic or gender subgroup. The results for co-gender effects are mixed. On the one hand, most analyses find no evidence for female (male) legislators being more or less responsive to female (male) citizens (Magni and Leon 2021 for Latin America, Dinesen, M. Dahl, and Schiøler 2021 in Denmark, Dhima 2020 in Canada). On the other hand, experiments in France, Germany, Ireland, Italy and the Netherlands reported by Magni and Leon 2021 find women to be significantly more responsive to women (with an average margin of 8.4 percentage points). Two experiments (Gaikwad and Nellis 2021; Kalla, Rosenbluth, and Teele 2018), in India and the United States, find women to be less responsive to women, albeit one yields negative results only with a particular outcome measurement.

Ten experiments investigated co-ethnicity effects. The results on co-ethnicity are striking. All but one experiment investigating interactions between the ethnicity of the legislator and the constituent find statistically significant differences on the response rate. The magnitudes differ vastly, going up to 20 percentage points (Hancock 2018; Dinesen, M. Dahl, and Schiøler 2021). These differences suggest that, especially for ethnicity, homophily is a central mechanism for biases and recording both sender and recipient traits is crucial.

5 Discussion

This systematic review identified 33 studies reporting 57 original audit experiments on elected officials, randomizing either the gender or ethnicity of the putative constituents. Substantively, three findings stand out. First, legislators are more responsive to female constituents in most experiments (1.5 percentage points across studies) and less responsive to ethnic minorities (4.4 percentage points). Second, two central contextual features – the content of the message and the level of the political unit – appear to be unrelated to response rates. Third, features related to design might be related to response rates. For gender, experiments with higher quality, measured

by the Cochrane Risk-of-Bias tool, tend to find larger differences. For ethnicity, published work and within-legislator designs do so too. Fourth, legislators do not seem to be structurally more responsive to constituents of their own gender. However, studies investigating co-ethnicity yield substantially large effects. Because of the limited number of studies on co-ethnicity and shared gender, this last finding is tentative at best. In this spirit of caution, we discuss contextual differences, the limitations of audit studies and systematic reviews next.

Context. With its turn to credibility and causal inference, research in comparative politics renewed its sensitivity about research contexts and external validity. One response to questions about external validity is to combine results across multiple studies from particular settings. Our meta-analysis on responsiveness followed this logic. Slough and Tyson (2023) show that conceptual issues arise when aggregating these estimates. They stress that the same underlying construct and measurement need to be represented across studies. We accomplished this task because audit studies of discrimination in representative democracies present itself as a canonical design in the field. Specifying inclusion criteria and implementing them enabled us to provide a "harmonized" and externally valid framework. At the same time, audit experiments typically fall short in identifying the underlying psychological processes of discrimination of responsiveness. We cannot resolve the issue here, but Nathan and Sands (2023) offer a promising avenue for exploring how context and context affect political behavior, not just for citizens, but also for political elites.

Limitations of audits with legislators. Our review encountered some more or less well-known limitations of audit studies. A common and broadly discussed concern (Gerber and Green 2012) is the excludability assumption: does the name of the putative constituent *exclusively* signal ethnicity or gender? Names might also signal party preferences (Butler and Broockman 2011), socioeconomic class or migrant status (Landgrave and Weller 2022). Such threats to the excludability assumption limit the internal validity of the inference that might be abated using blocking on other features.

Second, studies differ in how they measure responsiveness, even when among those focusing solely on response rate. Some include auto-responses, some count them as non-response, and some studies drop observations with non-personalized responses entirely. Laudably, Magni and Leon (2021) conduct robustness checks with different operationalizations of 'response', with no substantial differences. If a response policy, such as providing proof, exists and is distributed unevenly between MPs, e.g., higher prevalence among conservatives, a measurement bias is prevalent.

A third question centers on subjects and detection. Studies should explicitly state if staff or legislators answer requests and if potential heterogeneity arises here (Landgrave and Weller 2020). In addition to design considerations, questions of who exactly is responding why and when become central for evaluating solutions to overcome possible discrimination. Relatedly, several cases of uncovering were reported to researchers, for instance in Campbell and Bolet (2021). Particularly on lower levels of the legislature, the risk of uncovering is high, as MPs often share their staff (Alizade and Ellger 2021). Besides ethical concerns and explicit mentions of suspicion by subjects, unknown uncovering, i.e. MPs suspecting to be part of an audit study but not making their suspicion transparent, constitutes a greater threat to the internal validity. It is hard to assess, how often unknown uncovering occurs.

Fourth, regular attrition is only a minor problem in audit experiments, as non-response is a valid outcome. Two types of attrition can be problematic if they are correlated with the treatment. Firstly, several studies report non-functioning email addresses, particularly on lower legislative levels. In most cases, less than 10% of the emails bounced, but some experiments suffer from up to 33% attrition (Nisser 2017). With proper randomization, it is unlikely that the functioning of an email is correlated with the treatment allocation. Non-functioning emails are only a major threat if they accumulate in specific subgroups. Otherwise, they only cause deviations between the intention-to-treat and the per-protocol effect. Second, MPs can request their data to be removed after being debriefed about the experiment. Only very few studies debriefed the experiment subjects, but in one case, 161 British MPs had to be removed (Schakel et al. 2021). It is highly likely

that requests for removal are not distributed evenly and thereby limiting the internal validity of an audit study.

Fifth, a common problem of audits with politicians is the lack of statistical power. No experiment on gender with less than 1,000 observations found statistically significant effects, despite point estimates similar to studies with significant effects. For experiments on ethnicity, the threshold is less clear, but studies with less than 600 observations risk being under-powered.

Limitations of the systemic review. Systematic reviews and meta-analyses aggregate information. Even with a structured procedure, the results should be considered with the necessary caution, as information aggregation always comes with information reduction. Three aspects are particularly important in our context.

First, response rate, our simple outcome measure, covers only one part of responsiveness. Other outcomes measuring the response quality, speed, friendliness, etc. capture different layers of discrimination. Structured comparisons of outcome measures have shown that discrimination might happen on more complex, deeper levels, such as the response helpfulness (Hemker and Rink 2017). Thus, it is possible that this meta-analysis only captures a fraction of responsiveness and possible discrimination therein.

Second, a meta-analysis necessitates counting multi-treatment designs, e.g. studies comparing two ethnicities to one majority, separately. Because several experiments then share the same comparison group, observations in the meta-analysis are not fully independent. The lack of independence can lead to an underestimation of the standard error and confidence intervals of the combined effect. Three experiments were split up into two separate observations (Butler 2014; Landgrave and Weller 2022; Nisser 2017), one experiment into three observations (Gell-Redman et al. 2018). Given the limited number of these splits, it is unlikely that substantive distortions occur here.

Third, the results from the identified studies show no signs of publication bias. A more extensive analysis for exploring potential differences are reported in the appendix. If anything, the

funnel plots and test statistics suggest that studies on ethnicity finding strong effects are underrepresented.

6 Outlook

This systematic review has several implications for practice and future research. The overall results for service responsiveness are relatively clear and robust: when communicating with their representative, women enjoy a slight advantage, ethnic minorities suffer from a stronger disadvantage. Given these results and the recent debates on ethics, future research will not benefit much more from simply redeploying existing experimental designs.

The results provide important information for practitioners, including MPs, staff, and parliaments, about the existence of inequalities in service responsiveness. Some research has already investigated ways to reduce discrimination, such as higher professionalization of staff (Landgrave and Weller 2022). Future research should explore the underlying mechanism of how legislators and staff interact with constituents and where biases could stem from.

As a simple next step, in-group effects can be researched more thoroughly. As co-ethnicity and co-gender analyses would not even require new audit experiments, but can be done by reanalysing existing data with merged information on the MPs' gender and ethnicity. These analyses could be done at relatively low cost and with substantively huge consequences.

References

Alizade, Jeyhun and Fabio Ellger (2021). "Do Politicians Discriminate Against Constituents with an Immigration Background?" en. In: *The Journal of Politics*.

Allegretti, Aubrey (2021). "MPs criticise academics for sending them fictitious emails for research." en-GB. In: *The Guardian*. Retrieved from www.theguardian.com.

Bischof, Daniel et al. (2022). "Advantages, Challenges and Limitations of Audit Experiments with Constituents." In: *Political Studies Review* 20.2. Publisher: SAGE Publications, pp. 192–200.

Borenstein, Michael et al. (2010). "A basic introduction to fixed-effect and random-effects models for meta-analysis." en. In: *Research Synthesis Methods* 1.2, pp. 97–111.

Butler, Daniel M. (2014). *Representing the advantaged: How politicians reinforce inequality*. Cambridge: Cambridge University Press.

Butler, Daniel M. and David E. Broockman (2011). "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators." en. In: *American Journal of Political Science* 55.3, pp. 463–477.

Butler, Daniel M., Christopher F. Karpowitz, and Jeremy C. Pope (2012). "A Field Experiment on Legislators' Home Styles: Service versus Policy." In: *The Journal of Politics* 74.2. Publisher: The University of Chicago Press, pp. 474–486.

Campbell, Rosie and Diane Bolet (2021). "Measuring MPs' Responsiveness: How to Do it and Stay Out of Trouble." en. In: *Political Studies Review*, pp. 1–9.

Casellas, Jason P. and Sophia J. Wallace (2015). "The Role of Race, Ethnicity, and Party on Attitudes Toward Descriptive Representation." en. In: *American Politics Research* 43.1, pp. 144–169.

Costa, Mia (2017). "How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials." en. In: *Journal of Experimental Political Science* 4.3, pp. 241–254.

Crawfurd, Lee and Ukasha Ramli (2021). "Discrimination by politicians against religious minorities: Experimental evidence from the UK." en. In: *Party Politics*, pp. 1–8.

Crowder-Meyer, Melody (2013). "Gendered recruitment without trying: how local party recruiters affect women's representation." In: *Politics & Gender* 9.4, pp. 390–413.

Dahl, Robert A. (1998). *On democracy*. New Haven: Yale University Press.

Desposato, Scott (2015). *Ethics and experiments*. London: Taylor & Francis.

Dhima, Kostanca (2020). "Do Elites Discriminate against Female Political Aspirants? Evidence from a Field Experiment." en. In: *Politics & Gender*, pp. 1–32.

Dinesen, Peter Thisted, Malte Dahl, and Mikkel Schiøler (2021). "When Are Legislators Responsive to Ethnic Minorities? Testing the Role of Electoral Incentives and Candidate Selection for Mitigating Ethnocentric Responsiveness." en. In: *American Political Science Review* 115.2, pp. 450–466.

Driscoll, Amanda et al. (2018). "Prejudice, Strategic Discrimination, and the Electoral Connection: Evidence from a Pair of Field Experiments in Brazil." en. In: *American Journal of Political Science* 62.4, pp. 781–795.

Druckman, James N. and Donald P. Green (2021). *Advances in Experimental Political Science*. Cambridge: Cambridge University Press.

Dynes, AM and L Martin (2021). "Revenue Source and Electoral Accountability: Experimental Evidence from Local US Policymakers." In: *Political Behavior* 43.3, pp. 1113–1136.

Einstein, Katherine Levine and David M. Glick (2017). "Does Race Affect Access to Government Services? An Experiment Exploring Street-Level Bureaucrats and Access to Public Housing." en. In: *American Journal of Political Science* 61.1, pp. 100–116.

Eulau, Heinz and Paul D. Karps (1977). "The Puzzle of Representation: Specifying Components of Responsiveness." en. In: *Legislative Studies Quarterly* 2.3, p. 233.

Gaikwad, Nikhar and Gareth Nellis (2021). "Do Politicians Discriminate Against Internal Migrants? Evidence from Nationwide Field Experiments in India." en. In: *American Journal of Political Science* 65.4, pp. 790–806.

Gell-Redman, Micah et al. (2018). "It's All about Race: How State Legislators Respond to Immigrant Constituents." en. In: *Political Research Quarterly* 71.3, pp. 517–531.

Gerber, Alan S. and Donald P. Green (2012). *Field experiments: Design, analysis, and interpretation*. New York: WW Norton.

Hancock, Lynn Marissa (2018). "{Reply: All} An Experimental Study of Legislative Responsiveness." English. ProQuest Dissertations and Theses. https://www.proquest.com/wpsa/docview/2186840299/abs Ph.D. United States: Yale University.

Hedges, Larry V. and Jack L. Vevea (1998). "Fixed- and random-effects models in meta-analysis." In: *Psychological Methods* 3.4, pp. 486–504.

Hemker, Johannes and Anselm Rink (2017). "Multiple Dimensions of Bureaucratic Discrimination: Evidence from German Welfare Offices." en. In: *American Journal of Political Science* 61.4, pp. 786–803.

Higgins, Julian P. T. et al., eds. (2022). *Cochrane Handbook for Systematic Reviews of Interventions*. Version 6.3. Cochrane, London.

Homola, Jonathan (2019). "Are parties equally responsive to women and men?" In: *British Journal of Political Science* 49.3, pp. 957–975.

Janusz, Andrew and Nazita Lajevardi (2016). *The Political Marginalization of Latinos: Evidence from Three Field Experiments*. en. SSRN Scholarly Paper ID 2799043. Rochester, NY: Social Science Research Network.

John, Peter and Florian Foos (2022). "Introduction to Symposium: Experiments with Politicians: Ethics, Power, and the Boundaries of Political Science." In: *Political Studies Review* 20.2. publisher: SAGE Publications, pp. 169–174.

Kaas, Leo and Christian Manger (2012). "Ethnic Discrimination in Germany's Labour Market: A Field Experiment." en. In: *German Economic Review* 13.1, pp. 1–20.

Kalla, Joshua, Frances Rosenbluth, and Dawn Langan Teele (2018). "Are You My Mentor? A Field Experiment on Gender, Ethnicity, and Political Self-Starters." en. In: *The Journal of Politics* 80.1, pp. 337–341.

Kemper, Jakob (2018). "Dem „Deutschen" Volke? Ein Feldexperiment mit Mitgliedern des Deutschen Bundestages zur Untersuchung der Zugänglichkeit politischer Eliten für Bürger mit Migrationshintergrund." de. DuEPublico: Duisburg-Essen Publications Online. https://duepublico.uni-

duisburg-essen.de/servlets/DocumentServlet?id=47520. Bachelor-Thesis. Germany: University of Duisburg-Essen.

Landgrave, Michelangelo and Nicholas Weller (2020). "Do More Professionalized Legislatures Discriminate Less? The Role of Staffers in Constituency Service." en. In: *American Politics Research* 48.5. Publisher: SAGE Publications Inc, pp. 571–578.

— (2022). "Do Name-Based Treatments Violate Information Equivalence? Evidence from a Correspondence Audit Experiment." English. In: *Political Analysis* 30.1, pp. 142–148.

Lloren, Anouk (2017). "Does direct democracy increase communicative responsiveness? A field experiment with Swiss politicians." en. In: *Research & Politics* 4.1, pp. 1–8.

Loewen, Peter John and Michael Kenneth MacKenzie (2019). "Service Representation in a Federal System: A Field Experiment." English. In: *Journal of Experimental Political Science* 6.2, pp. 93–107.

Loewen, Peter John and Daniel Rubenson (2022). "Value-added and Transparent Experiments." In: *Political Studies Review* 20.2. Publisher: SAGE Publications, pp. 243–249.

Magni, Gabriele and Zoila Ponce de Leon (2021). "Women Want an Answer! Field Experiments on Elected Officials and Gender Bias." en. In: *Journal of Experimental Political Science* 8.3, pp. 273–284.

Mansbridge, Jane (2015). "Should Workers Represent Workers?" en. In: *Swiss Political Science Review* 21.2, pp. 261–270.

McKenzie, JE et al. (2021). "Defining the criteria for including studies and how they will be grouped for the synthesis." In: *Cochrane Handbook for Systematic Reviews of Interventions*. Ed. by Julian P. T. Higgins et al. Version 6.3. Cochrane, London. Cochrane.

Mendez, Matthew S. and Christian R. Grose (2018). "Doubling Down: Inequality in Responsiveness and the Policy Preferences of Elected Officials." en. In: *Legislative Studies Quarterly* 43.3, pp. 457–491.

Moher, David et al. (2009). "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement." en. In: *PLOS Medicine* 6.7, pp. 1–6.

Nathan, Noah L. and Melissa L. Sands (2023). "Context and contact: Unifying the study of environmental effects on politics." In: *Annual Review of Political Science* 26. Publisher: Annual Reviews.

Naurin, Elin and Patrik Öhberg (2021). "Ethics in Elite Experiments: A Perspective of Officials and Voters." en. In: *British Journal of Political Science* 51.2, pp. 890–898.

Newland, Sara A. and John Chung-En Liu (2021). "Ethnic identity and local government responsiveness in Taiwan." en. In: *Governance* 34.3, pp. 875–892.

Nisser, Annerose (2017). "Cross-Ethnic Interactions and the Influence of Politics : Evidence from Online Spaces and a Field Experiment in Bosnia and Herzegovina." en. Konstanzer Online-Publikations-System (KOPS). http://nbn-resolving.de/urn:nbn:de:bsz:352-2–gwduw0vgdsk27. Ph.D. Germany: University of Konstanz.

Peters, Yvette (2018). "Democratic representation and political inequality: how social differences translate into differential representation." en. In: *French Politics* 16.3, pp. 341–357.

Pitkin, Hanna Fenichel (1967). *The concept of representation*. Oakland, California: University of California Press.

Putnam, Robert D., Robert Leonardi, and Raffaella Y. Nanetti (1993). *Making democracy work: Civic traditions in modern Italy*. Princeton, New Jersey: Princeton University Press Princeton.

R Core Team (2021). *R: A language and environment for statistical computing*. Retrieved from www.R-project.org/. Vienna, Austria.

Rhinehart, Sarina (2020). "Mentoring the Next Generation of Women Candidates: A Field Experiment of State Legislators." en. In: *American Politics Research* 48.4, pp. 492–505.

RStudio Team (2022). *RStudio: Integrated Development Environment for R*. Retrieved from http://www.rstudio.com/ Boston, MA.

Schakel, Wouter et al. (2021). "Unequal Responsiveness in MP-Citizen Communication: A Comparative Field Experiment." en.

Shea, Beverley J et al. (2017). "AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both." en. In: *BMJ* 358, j4008.

Slough, Tara and Scott A. Tyson (2023). "External Validity and Meta-Analysis." In: *American Journal of Political Science* 67.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12742, pp. 440–455.

Sterne, Jonathan A. C. et al. (2019). "RoB 2: a revised tool for assessing risk of bias in randomised trials." en. In: *BMJ* 366, p. l4898.

Sterne, Jonathan AC and Matthias Egger (2001). "Funnel plots for detecting bias in meta-analysis: guidelines on choice of axis." In: *Journal of clinical epidemiology* 54.10, pp. 1046–1055.

Stockemer, Daniel (2015). "Women's descriptive representation in developed and developing countries." en. In: *International Political Science Review* 36.4, pp. 393–408.

Teele, Dawn Langan, Joshua Kalla, and Frances Rosenbluth (2018). "The Ties That Double Bind: Social Roles and Women's Underrepresentation in Politics." en. In: *American Political Science Review* 112.3, pp. 525–541.

Thomsen, Danielle M and Bailey K Sanders (2020). "Gender Differences in Legislator Responsiveness." English. In: *Perspectives on Politics* 18.4, pp. 1017–1030.

Verba, Sidney (2003). "Would the dream of political equality turn out to be a nightmare?" In: *Perspectives on politics* 1.4, pp. 663–679.

Viechtbauer, Wolfgang (2010). "Conducting Meta-Analyses in R with the metafor Package." en. In: *Journal of Statistical Software* 36, pp. 1–48.

Appendix

Appendix A: Search Terms and Databases

Table A1 displays the search terms of the final search using the Web Of Science database. The 'TS'-operator searches the 'topic', covering title, abstract, author keywords and 'keywords plus' (index terms occurring in the references of the article).

Table A2 displays the search terms of the final search using 15 databases on ProQuest. The following databases included: Worldwide Political Science Abstracts, Social Services Abstracts, Sociological Abstracts, Sociology Database, African Writers Series, Applied Social Sciences Index & Abstracts, British Periodicals, Ebook Central, International Bibliography of the Social Sciences (IBSS), LGBT Magazine Archive, PAIS Index, Periodicals Archive Online, Periodicals Index Online, Political Science Database, Publicly Available Content Database. The 'su'-Operator searches the subject of the publication. These were used since they are more broad than specific keywords and work with non-exact matching. Thus, the subject 'audit' also includes 'audit study' for example.

Comb.	Definition	Results
A	TS=(Politician* OR elected official* OR elite*)	84,035
B	TS=(experiment* OR audit OR vignette*)	6,314,869
C	TS=(represent* OR responsive*)	3,010,052
D	A AND B AND C	677

Table A1: Search Terms used for Web Of Science

Comb.	Definition	Results
A	su(“Legislators”) OR su(“Politicians”) OR su(”Elites”) OR su(“Local government”) OR su(“State government”)	201,141
B	su(“Audit”) OR su(“vignettes”) OR su(“Responsiveness”)	13,770
C	su(“Discrimination”)	108,049
D	A AND B	613
E	A AND C	1,224
F	B AND C	123
G	D OR E OR F	1,942

Table A2: Search Terms used for ProQuest Databases

Appendix B: Experiments included in the Review and Covariates

Table B1 shows a list with all experiments included in the review and meta-analysis. The experiments reported by Schakel et al. 2021; Mendez and Grose 2018; Dynes and Martin 2021; Newland and Liu 2021 and Rhinehart 2020 are not included in the meta-analysis, as the relevant data was not available. Column 4 denotes the time of the experiment, not the publication date. Columns 5 and 6 are dummy variables whether the experiment randomized the gender or ethnicity of the senders. The last column shows the quality assessment based on the Cochrane Risk-of-Bias tool. Note that 'low' stands for low Risk-of-Bias, meaning a good study quality.

Table B1: Experiments included in the Review

ID	Reference	Country	Year	Gender	Ethnicity	Publication	Risk-of-Bias
1_0	Alizade & Ellger 2021	Germany	2018		1	Journal of Politics	low
2_0	Butler & Broockman 2011	US	2008		1	American Journal of Political Science	low
3_0	Butler & Crabtree 2017	US	2014		1	Journal of Experimental Political Science	some concerns
4_2	Butler 2014	US	2010		1	Cambridge University Press	some concerns
4_3	Butler 2014	US	2010		1	Cambridge University Press	some concerns
4_1	Butler 2014	US	2010	1		Cambridge University Press	some concerns
5_0	Crawfurd & Ramli 2021	UK	2014	1	1	Party Politics	some concerns
6_0	Dhima 2018	Canada	2015	1		Politics & Gender	low
7_0	Dinesen et al. 2021	Denmark	2017	1	1	American Political Science Review	some concerns
8_0	Distelhorst & Hou 2014	China	2014		1	Quarterly Journal of Policital Science	some concerns
9_1	Driscoll et al. 2018	Brazil	2010	1	1	American Journal of Political Science	high
9_2	Driscoll et al. 2018	Brazil	2011	1	1	American Journal of Political Science	high
10_0	Dynes et al. 2021	US	2016	1		American Politics Research	some concerns
11_0	Gaikwad & Nellis 2021	India	2018	1		American Journal of Political Science	some concerns
12_1	Gell-Redman et al. 2018	US	2013		1	Political Research Quarterly	low
12_2	Gell-Redman et al. 2018	US	2013		1	Political Research Quarterly	low
12_3	Gell-Redman et al. 2018	US	2013		1	Political Research Quarterly	low
13_0	Golder et al. 2019	New Zealand	2018	1		Political Science	high

Table B1: Experiments included in the Review

ID	Reference	Country	Year	Gender	Ethnicity	Publication	Risk-of-Bias
14_0	Habel & Birch 2019	UK	2014		1	Legislative Studies Quarterly	low
15_1	Hancock 2018	Canada	2017	1	1	Thesis	some concerns
15_2	Hancock 2018	Canada	2017	1	1	Thesis	some concerns
15_3	Hancock 2018	Canada	2017	1		Thesis	some concerns
16_0	Heß et al. 2018	Germany	2018		1	Soziale Welt	some concerns
17_1	Janusz & Lajevardi 2016	US	2013		1	Working Paper	some concerns
17_2	Janusz & Lajevardi 2016	US	2013		1	Working Paper	some concerns
17_3	Janusz & Lajevardi 2016	US	2015		1	Working Paper	some concerns
18_0	Kalla et al. 2018	US	2014	1	1	The Journal of Politics	low
19_0	Kemper 2018	Germany	2018		1	Thesis	high
20_0	Lajevardi 2020	US	2015		1	Politics, Groups, and Identities	some concerns
21_1	Landgrave & Weller 2022	US	2018		1	Political Analysis	some concerns
21_2	Landgrave & Weller 2022	US	2018		1	Political Analysis	some concerns
22_0	Landgrave 2020	US	2018		1	State Politics & Policy Quarterly	some concerns
23_0	Landgrave 2021	US	2018		1	Legislative Studies Quarterly	low
24_0	Lloren 2017	Switzerland	2015	1		Research and Politics	some concerns
25_1	Loewen & MacKenzie 2019	Canada	2010	1	1	Journal of Experimental Political Science	some concerns
25_2	Loewen & MacKenzie 2019	Canada	2010	1	1	Journal of Experimental Political Science	some concerns

Table B1: Experiments included in the Review

ID	Reference	Country	Year	Gender	Ethnicity	Publication	Risk-of-Bias
26_6	Magni & Leon 2021	Argentina	2016	1		Journal of Experimental Political Science	some concerns
26_7	Magni & Leon 2021	Brazil	2017	1		Journal of Experimental Political Science	some concerns
26_8	Magni & Leon 2021	Chile	2017	1		Journal of Experimental Political Science	some concerns
26_9	Magni & Leon 2021	Colombia	2017	1		Journal of Experimental Political Science	some concerns
26_1	Magni & Leon 2021	France	2017	1		Journal of Experimental Political Science	some concerns
26_2	Magni & Leon 2021	Germany	2016	1		Journal of Experimental Political Science	some concerns
26_3	Magni & Leon 2021	Ireland	2017	1		Journal of Experimental Political Science	some concerns
26_4	Magni & Leon 2021	Italy	2017	1		Journal of Experimental Political Science	some concerns
26_10	Magni & Leon 2021	Mexico	2017	1		Journal of Experimental Political Science	some concerns
26_5	Magni & Leon 2021	Netherlands	2017	1		Journal of Experimental Political Science	some concerns
26_11	Magni & Leon 2021	Uruguay	2017	1		Journal of Experimental Political Science	some concerns
27_0	Mendez & Grose 2018	US	2012		1	Legislative Studies Quarterly	low
28_0	Newland & Liu 2021	Taiwan	2017		1	Governance	low
29_1	Nisser 2017	Bosnia and Herzegowina	2017		1	Thesis	low
29_2	Nisser 2017	Bosnia and Herzegowina	2017		1	Thesis	low
30_0	Rhinehart 2020	US	2018	1		American Politics Research	some concerns

Table B1: Experiments included in the Review

ID	Reference	Country	Year	Gender	Ethnicity	Publication	Risk-of-Bias
33_1	Schakel et al. 2021	Germany			1	Working Paper	high
33_3	Schakel et al. 2021	Netherlands	2017		1	Working Paper	high
33_2	Schakel et al. 2021	UK			1	Working Paper	high
31_0	Taghizadeh et al. 2022	Sweden	2018		1	Parliamentary Affairs	some concerns
32_0	Thomsen & Sanders 2020	US	2016	1		Perspectives on Politics	some concerns

Appendix C: Analysis of Publication Bias

The distribution of effect sizes of different experiments can yield information on the prevalence of publication bias. Figure C1 shows the funnel plot for the meta-analyses on gender (left pane) and ethnicity (right pane). Following J. A. Sterne and Egger (2001), the y-axis depicts the standard error as measure for the study size and the difference in response rate as x-axis. In the absence of bias, the distribution on these two axes should correspond to a symmetrical funnel. The funnel plot of the gender meta-analysis looks fairly symmetric. Egger's regression test yields a non-significant intercept (p = 0.89), supporting the graphic interpretation of symmetry.

In contrast, many of the studies on ethnicity with small sample sizes, i.e., great standard error, fall to the right of the summary effect size, making the plot look slightly asymmetric. This means that potentially, small studies yielding a negative outcome (suggesting discrimination), are published less likely. However, Egger's regression test yields no significance of this graphic asymmetry (p = 0.18).

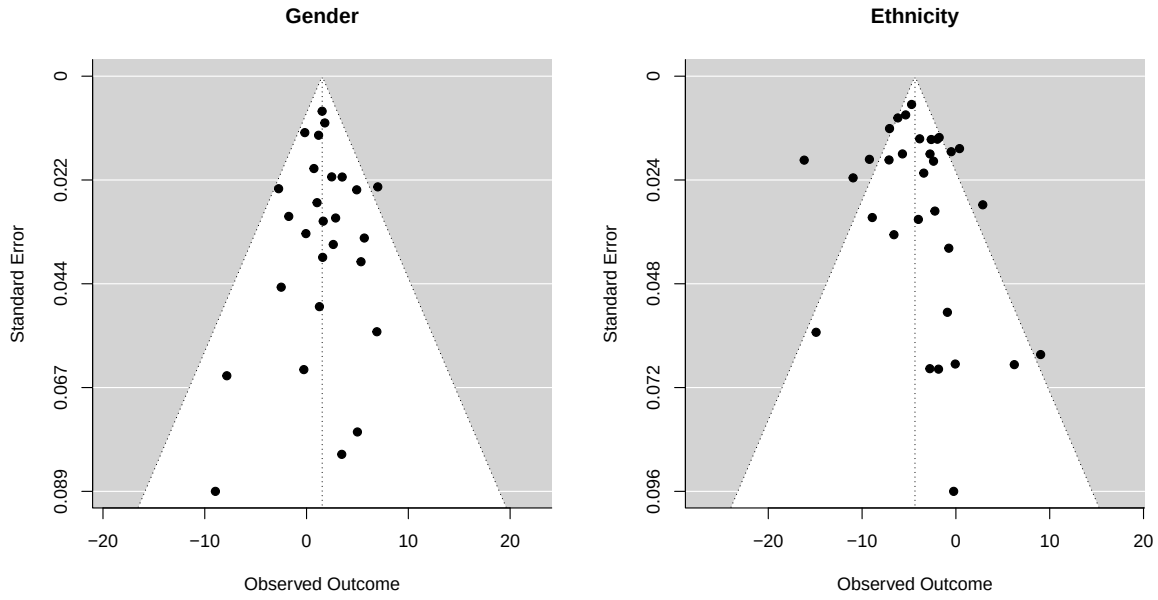

Figure C1: Funnel Plots for both Meta-Analyses