"Do Gender Quotas Really Reduce Bias? Evidence from a Policy Experiment in Southern Africa" - A Causal Re-analysis

Lauren Pendo, Robert Schmidt

Fall 2020

1 Introduction

In the 2018 paper *Do Gender Quotas Really Reduce Bias? Evidence from a Policy Experiment in Southern Africa*, Clayton analyzes the extent to which exposure to quota-elected female officials impacted explicit and implicit gender biases among the residents of electoral divisions (EDs) in Lesotho. The paper is a natural follow-up the work done by Beaman et al. [2], Bhavnani et al. [3], and others who showed that "women's presence in local leadership [in India] increases women citizens' political engagement and the likelihood that women will run in future elections." [1]

It has been hypothesized that exposure to female elected officials may change gender biases of the citizenry, with quotas as one way to ensure this exposure. Researchers in the field have posited that such quotas could potentially impact both explicit gender biases (direct decisions that reflect bias) and implicit gender biases (sentiments and preferences manifested through socially-held stereotypes). Given the apparent success of these quotas in India, Clayton's work aims to address whether quotas successfully generalize to other locales.

2 Study Overview

2.1 Design

The female-only official quota was imposed between 2005-2011 as treatment by the government of Lesotho randomly to approximatly 30% of the EDs. The author also independently verified the validity of the government's random assignment in the supplement to the article.

In the language of causal inference, the quota experiment was $\text{CRD}(D_1 = 0.3 \cdot D, D)$ on the D EDs in Lesotho. However, the paper is interested in the biases of the ED citizens; hence, our population is the individual members of the EDs. The paper estimates that there are about 1000 citizens per ED, with about 9-15 EDs for each of the 129 community councils that govern the country (so, we consider the experiment on the citizen level as $\text{CRD}(N_1 = 0.3 \cdot N, N)$ for the N citizens of Lesotho). Thus, the treatment $Z_i \in \{0, 1\}$ is whether or not a citizen is a member of a district without the quota ($Z_i = 0$) or a district with the quota ($Z_i = 1$).

2.2 Outcomes

Clayton's interest in the paper is the impact of the quota on both explicit and implicit gender bias. To test the impact on explicit bias, she examined responses to the 2012 Afro-Barometer survey for individuals who correspond to EDs with known quota status. Specifically, she considers three outcome variables:

- **Political bias**: $Y^{(\text{pol})}$, with potential outcomes $Y_i^{(\text{pol})}(1)$, $Y_i^{(\text{pol})}(0)$
- **Traditional bias**: $Y^{(\text{rights})}$, with potential outcomes $Y_i^{(\text{rights})}(1)$, $Y_i^{(\text{rights})}(0)$
- **Education bias**: $Y^{(\text{edu})}$, with potential outcomes $Y_i^{(\text{edu})}(1)$, $Y_i^{(\text{edu})}(0)$

Each of the outcomes is scored on a scale of $0-4$, with higher scores representing more gender-egalitarian answers. Since Clayton merged the survey data on the quota experiment information, we also have access to to the age, education level, gender, and religion group of the 996 respondents. Given the dataset size and number of informational features, we will be focusing on the explicit bias experiment in our causal re-analysis.

2.3 Causal Estimand and Assumptions

Since the randomness of the quota assignment was verified by the author, we can safely assume SUTVA: for each of the outcome variables $\omega \in \{\text{pol, rights, edu}\}$, $Y_i^{(\omega)}(\vec{Z}) = Y_i^{(\omega)}(Z_i) \ \ \forall\, i \in \{1, \ldots, N\}$. SUTVA also implies that there is no hidden version of the treatment. This seems reasonable given that the treatment is mandating that only women be elected as officials in a district.

For the design $\eta \sim \text{CRD}(N_1, N)$, we note the following four properties:

1. Probabilistic: there is a non-zero probability of assignment;
   $0 < P(Z_i \mid \vec{X}, \vec{Y}^{(\omega)}(1), \vec{Y}^{(\omega)}(0)) < 1$, for $\vec{X}$ as the observed demographic covariates.
2. Known assignment mechanism: 30% of the EDs received the quota assignment.
3. Individualistic: unit assignment is not dependent on other units;
   $P(Z_i \mid \vec{X}, \vec{Y}^{(\omega)}(1), \vec{Y}^{(\omega)}(0)) = P(Z_i \mid \vec{X}_i, Y_i^{(\omega)}(1), Y_i^{(\omega)}(0))$
4. Unconfoundedness: assignment is independent of the outcome;
   $P(\vec{Z} \mid \vec{X}, \vec{Y}^{(\omega)}(1), \vec{Y}^{(\omega)}(0)) = P(\vec{Z} \mid \vec{X})$

Taking cues from Clayton's paper, we use the average treatment effect as our causal estimand (for each of the outcome variables $\omega \in \{\text{pol, rights, edu}\}$):

$$\tau^{(\omega)-ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{(\omega)}(1) - Y_i^{(\omega)}(0) = \overline{Y}^{(\omega)}(1) - \overline{Y}^{(\omega)}(0)$$

2.4 Observational Study or Randomized Experiment?

While the randomization of quota assignment was performed on the ED level, Clayton's interest lies in the effect of the quota on an individual within an ED; hence, as discussed in section 2.2, she uses the Afro-Barometer survey given in 2012, given after the quota ended in 2011, as a proxy for the "effect" of the randomized experiment on the citizen level.

While this approach is pragmatic given the constraints of the experiment, one may question whether this aggregated dataset is truly a completely randomized experiment at the individual level, where the "individual" in question is a citizen rather than the ED which received the treatment. For instance, the more impossible task of assigning a quota to every citizen (whether that citizen can only vote for men or women) would be a more unambiguous "completely randomized design" at the citizen level. Hence, in our analysis, we will generally treat the study as observational and see the extent to which Clayton's results depend on the guarantees of a completely randomized experiment.

3 Exploratory Data Analysis

As a preliminary to our causal re-analysis, we investigated the data to examine the degree to which the covariates are balanced across treatment and control. Figure 1 details the distribution of each outcome variable across quota treatment; there does not appear to be a noticeable difference in the response across the quota assignment imposed by the government.

In addition, we examined the distribution of age versus the quota assignment in Figure 2. The distributions of age across treatment are distributed similarly, with treatment having an average age of 40 and control

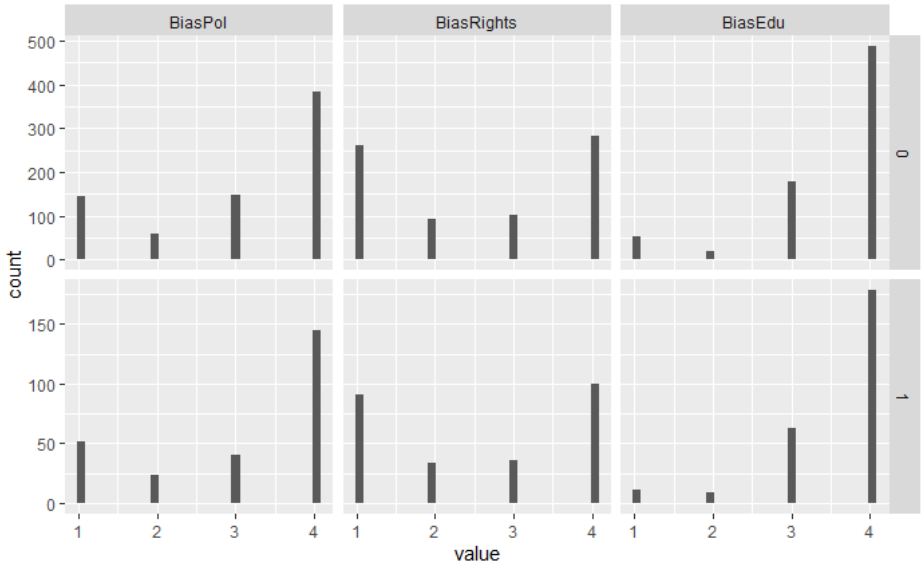

Figure 1: Distribution of outcome variables across the quota treatment.

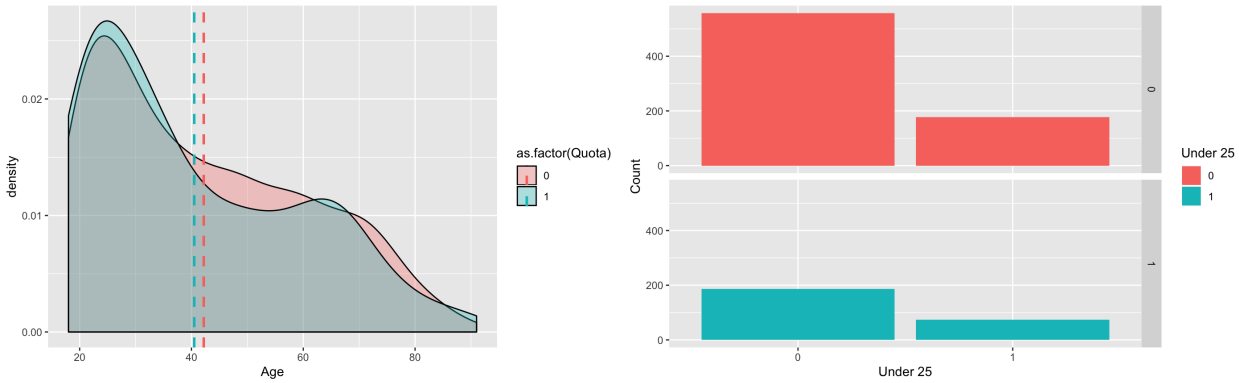

Figure 2: Distribution of age across Quota treatment.

an average age of 42. Overall, the treatment skews slightly younger, while the control group contains more participants across ages 40-80. Another notable observation is that 28.35% of participants who lived in districts with the quota were under 25, compared to 24.08% of participants without the quota.

Noting the work done by Beaman et al. in India [2], Clayton took particular interest in role model effects among adolescent females between the age of eleven and fifteen. While citizens in this age bracket cannot yet vote, the Afro-Barometer survey does include participants under the age of fifteen. Hence, Clayton binarized the continuous age variable into an over/under 25 subgroup. We perform the same binning in order to better align our causal re-analysis with her original work; the plot of the binned age variable ("Under 25") is shown in the rightmost sub-figure of 2.

Clayton also took particular interest in the relationship between bias and a citizen's gender. Figure 3 shows the distribution of males and females across treatment and control; we can see that gender is well balanced across the districts with and without a quota. The gender distribution is quite balanced across treatment and control, with 50.95% female in treatment and 49.66% female in control.

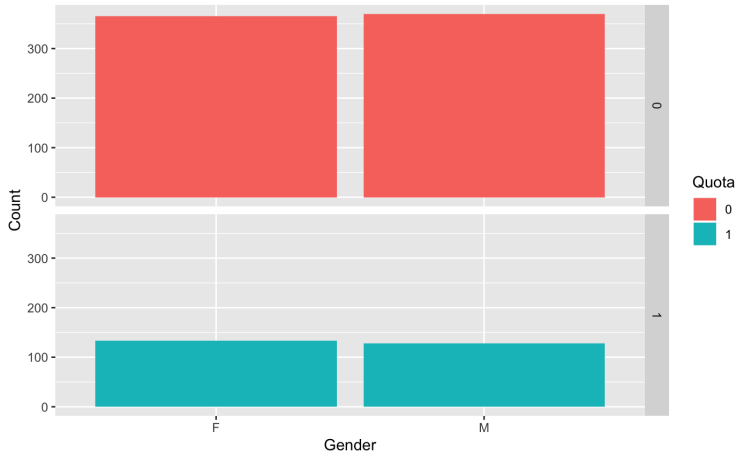

Figure 3: Distribution of gender across Quota treatment.

4 Neymanian Replication Study

Outcome	Mean treated	Mean untreated	ATE (95% CI)	p-value (two-sided)	Clustered p-value (two-sided)
Political bias	3.07	3.05	0.02 $(-0.18, 0.15)$	0.86	0.87
Traditional bias	2.56	2.55	0.01 $(-0.19, 0.18)$	0.98	0.99
Education bias	3.56	3.49	0.07 $(\mathbf{-0.18}, \mathbf{0.04})$	**0.24**	0.52

Table 1: Our replication of the paper's *Table 1* - discrepancies are marked in bold.

As a first step in our causal re-analysis, we attempted to replicate the results from Clayton's explicit bias study. Using standard t-testing, we see that our results essentially match the paper results, minus some differences in the confidence intervals and p-values given that Clayton clustered her standard errors by ED (education bias appears to be the most significantly affected by this discrepancy). Given the results from this Neymanian replication study, we decided to re-analyze the experiment from a Fisherian point of view, which takes far fewer assumptions into account.

5 Fisher Randomization

To understand the extent to which Clayton's results depend on Neymaninan assumptions of normality, we now analyze the results from a Fisherian perspective. Recall that the Fisher Randomization test makes no assumptions aside from SUTVA, which we verified in a previous section. The Fisher null hypothesis is as follows:

$$H_0^{(\omega)-\text{Fisher}} : Y_i^{(\omega)}(1) = Y_i^{(\omega)}(0) \;\; \forall\, i \in \{1, \ldots, N\} \;\; \text{and } \omega \in \{\text{pol, rights, edu}\}$$

Under Fisher's null, there are no treatment effects across every unit. Also note that Fisher's null implies the Neymanian null, and is thus a stronger hypothesis.

We performed the Fisher Randomization test on each outcome separately using 5,000 random permutations; the results are displayed in Table 2. In Figure 4, we can see a histogram of the difference in means estimator for different assignment permutations plotted against the observed ATE in the Clayton's experiment.

Outcome	ATE	p-value (two-tailed)
Political bias	0.02	0.936
Traditional bias	0.01	0.978
Education bias	0.97	0.254

Table 2: Results from the FRT on Experimental Data.

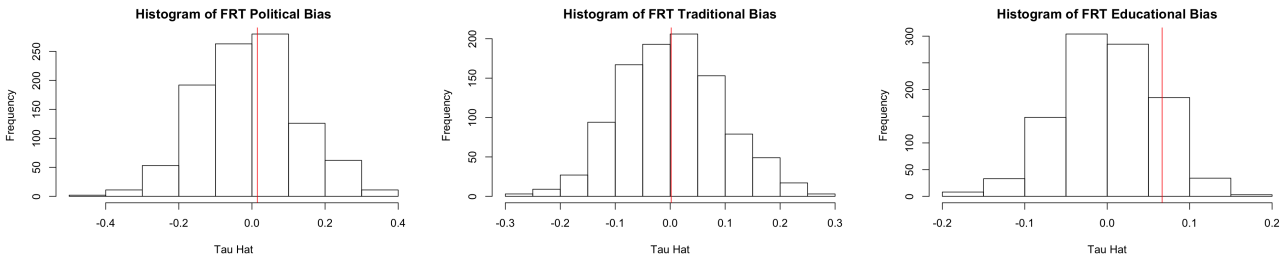

Figure 4: Histogram of FRT for Each Outcome.

We can see from a Fisherian perspective, none of the outcomes comes even close to the significance cutoff of 0.05. Both Traditional Bias and Education Bias have similar p-values from the Neymanian and Fisherian approach. The p-value of Political bias is higher in the Fisher approach compared to the Neymanian, indicating that the assumptions of the Neymanian approach may break down for this particular outcome variable.

6 Matching

6.1 Motivation

As supposed by Clayton, demographic factors may impact the likelihood of an individual to change perceptions of gender norms in response to the election quotas. In the article, Clayton mainly focuses on the impact of age and gender; she more specifically considers age as a binary variable, splitting the citizens as cohorts above and below 25 years of age. In order to correct for the impact of these demographic variables she uses "treatment by covariate interactions to identify model-based estimates for the quota's conditional average treatment effects" [1] within each of the four subgroups of age and gender. While she found no statistically significant results given that she performed a Bonferonni correction ($\alpha = 0.05/5 = 0.0125$) to account for the multiple testing, she did get one low p-value for young women's view on political rights ($p = 0.056$).

While in fact the quota venture was a randomized experiment, from the point of view of our re-analysis we can consider the impact of matching on covariates as if it were an observational study. Given that regressing on gender and age yielded more promising results than the first round of direct testing, it may prove informational to analyze the impact of matching on these two factors as well as on the other covariates (education level, poverty level, and religion).

6.2 Exploratory Matching

Recall that we have five demographic variables: age (over/under 25), gender, poverty level, religion, and education level. Hence, there are a number of possible covariates we can match on in addition to the two under consideration by Clayton in the paper. Since the dataset is imbalanced 30% treated / 70% control, we consider one-to-one matchings, and employ the Mahalanobis distance as our metric given its success on similar low-dimensional datasets. Here, we recall that the Mahalanobis distance between variables X_i and X_j is given by:

$$d_M(X_i, X_j) = \sqrt{(X_i - X_j)^T \hat{\Sigma}^{-1} (X_i - X_j)}$$

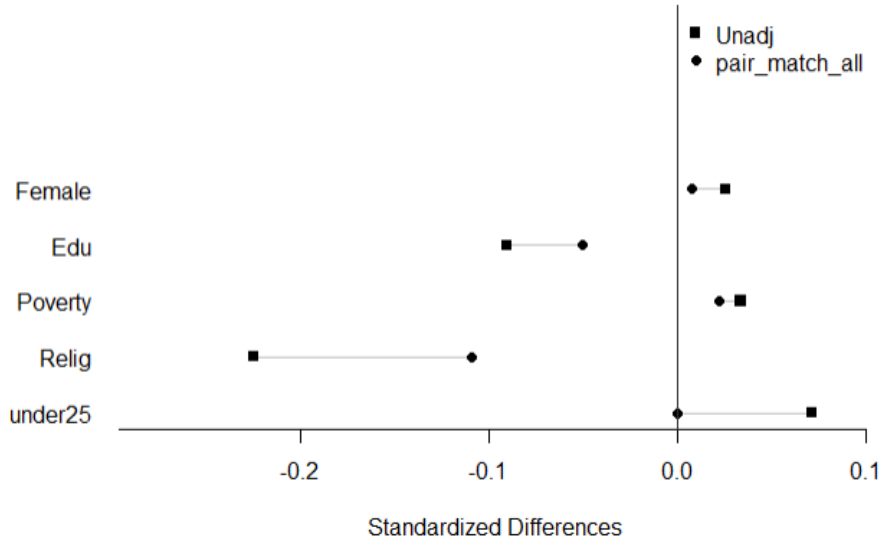

Figure 5: Results from matching on all observed covariates.

In the above expression, $\hat{\Sigma}$ is the sample covariance for the covariates in the dataset.

First we consider the results from matching on all variables as shown in Figure 5. We see that matching on all variables has a positive effect on the standardized differences of each covariate, especially gender and age. However, the remaining variables, and in particular religion, still have nonzero standardized differences.
A somewhat exhaustive search was performed on matching combinations for the remaining covariates (in terms of both pairs and triplets) in order to see if any other matching would be more conducive to a balanced analysis. The results from matching only on age and gender are shown in Figure 6. In the end, it would

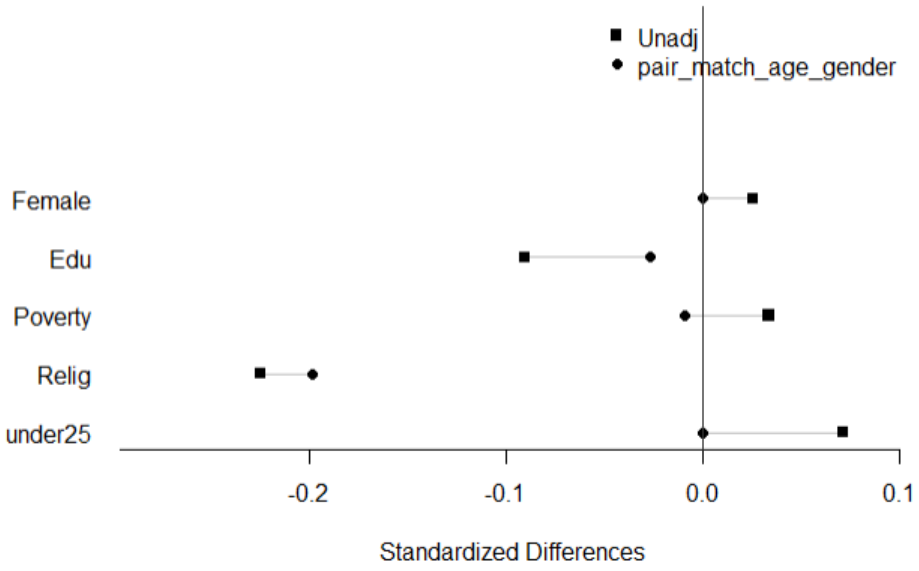

Figure 6: Matching on age and gender only.

appear that in matching only on age and gender rather than all of the covariates, we improve on age, gender,

poverty, and education, but lose some of the matching progress on religion. In order to better align our causal re-analysis efforts with the original paper's analysis on the effects of age and gender, we will focus more in-depth on the age/gender-only matching. Within this matching scheme, we performed a paired Fisher randomization test in order to test the strict null hypothesis of zero treatment effect across all units. The results from this FRT are shown in Table 3.

Outcome	ATE	p-value (two-tailed)
Political bias	0.03	0.37
Traditional bias	0.12	0.66
Education bias	0.14	0.13

Table 3: Results from the FRT on the age/gender matching.

Here, we note that the results have generally become more significant across the board as compared to Clayton's Neymanian tests with the exception of her more-significant result for education bias when adjusting for covariates via regression. We do note however that none of these results are statistically significant at a reasonable threshold. Since the p-values did tend to decrease, especially for political bias, we now perform a sensitivity analysis to see how the results hold up to deviations from ignorability as we view the paper from an observational lens.

7 Sensitivity Analysis

In addition to addressing deviations from ignorability, sensitivity analyses also allow us to understand if we have captured all the influential covariates in our matching. More specifically, for any matching pair k, let $\pi_{k1} = \pi(X_{k1})$ for probability π, and similarly for π_{k2}. Using these terms, we can construct the odds ratio:

$$\begin{aligned} \text{odds}_{k1} &= \frac{\pi_{k1}}{1-\pi_{k1}} \\ \text{odds}_{k2} &= \frac{\pi_{k2}}{1-\pi_{k2}} \\ \nu_k &= \frac{\text{odds}_{k1}}{\text{odds}_{k2}} = \text{odds ratio} \end{aligned}$$

In general, this then implies that:

$$p_k = P(Z_{k1} = 1 \mid Z_{k1} + Z_{k2} = 1) = \frac{\nu_k}{\nu_k + 1}$$

Under ignorability, we suppose that $\pi_{k1} = \pi_{k2}$, which accordingly implies that $\nu_k = 1$ and $p_k = 1/2$. Sensitivity analysis answers the question of what happens when $\pi_{k1} \neq \pi_{k2}$ slightly. Here, deviations are paramaterized on the ν_k scale:

$$\frac{1}{\Gamma} \leq \nu_k \leq \Gamma \;\; \forall\, k, \text{ and } \Gamma \geq 1$$

Note here that $\Gamma = 1 \implies \nu_k = 1 \implies$ ignorability. For a set amount of violation Γ, we calculate the following statistic in order to give a sense of how much the results depend on ignorability.

$$M(\Gamma) = \max_{\vec{\nu} \in [1/\Gamma, \Gamma]^k} \{p\text{-value}\}$$

We computed the upper bound using sensitivity parameter $\Gamma = 1.2$ and employing the senm function from the package authored by Rosenbaum; the results are summarized in table 4. To align our analysis to the work performed by Clayton, here we report the upper-bound on the two-sided p-value. Rosenbaum posits in his notes to the senm package that an upper bound on the two-sided p-value is the minimum of 1 and twice the smaller of each of the one-sided p-values provided by the senm function. As expected, many of the p-values have increased relative to the simple FRT matching results. In particular, it appears that the political bias and traditional bias results are quite sensitive to deviations from ignorability; the education bias result appears to have changed the least from the FRT matching outcome.

Outcome	|M Statistic|	Upper Bound on p-value
Political bias	2.66	0.97
Traditional bias	0.67	1.00
Education bias	6.17	0.32

Table 4: P-Value Sensitivity Analysis.

8 Subclassification on Propensity Score

As an alternative to matching, we performed sub-classification on propensity score using bins. Subclassification involves splitting the data into K strata, and calculating the propensity score within each group. For our analysis, we stratified into $K = 5$ groups (20th, 40th, 60th, 80th, 100th percentiles) of the propensity score as seen in Table 5. We can see that both the number of treated units and control units are relatively balanced across quartiles. In the case where there are K discrete propensity scores, the assignment for each strata is a CRD and the whole experiment is SCRD. Since our propensity scores our continuous, the last stratum contains all units with propensity scores for some π_{K-1} to π_K. Even under ignorability, we no longer have a CRD within each stratum, and thus the whole experiment is no longer SCRD. Thus, our τ^{ATE} estimate is biased, even for large samples. .

Quartile	Propensity	N Treatment	N Control
20	0.23	45	154
40	0.25	46	152
60	0.26	48	152
80	0.29	53	142
100	0.56	69	135

Table 5: Quartiles by Propensity Score.

We computed the stratified difference in means estimate of the ATE based upon the above strata as well as an estimate of the variance. We then computed a 95% confidence interval based on a Normal approximation. We can see in the results in Table 6 that 0 is contained in each of the confidence intervals. Thus, we cannot reject the Neymanian null hypothesis for any of the three outcomes. These results are in line with the results seen through the other approaches, where we have been unable to detect a statistically significant difference between treatment and control. The confidence intervals for both Political Bias and Traditional Bias from the sub-classification approach are very similar to the original results from the Neymanian experimental perspective as seen in Table 1.

Outcome	Stratified ATE	Variance	95% Confidence Interval
Political bias	0.026	0.008	(-0.14, 0.20)
Traditional bias	0.022	0.009	(-0.16, 0.21)
Education bias	0.08	0.003	(-0.03, 0.18)

Table 6: Subclassification on Propensity score

9 Discussion

At the crux of our re-analysis was the question of how much we could rely on the random quota assignment at the ED level holding at the individual level; to this end, we leveraged a number of analysis techniques from the observational study literature in order to better understand how much the study depended on the consequences of a completely randomized experimental design.

Through our exploratory data analysis, we verified that the covariates observed in the Afro-Barometer survey do tend to be balanced across treatments as supposed by Clayton in her supplemental analysis. Then, we analyzed her initial, direct testing through Clayton's own Neymanian lens, as well as from the Fisherian perspective. Here, we found that her results for political and traditional bias held even in the absence of normality assumptions, while her result for education bias did change in the Fisherian scheme.

Viewing the quota dataset an an observational study, we used matching on age and gender in order to possess potential counfounding at the individual level that may not have been evident in the original analysis. Here, we saw that the results for each outcome generally became more significant, although none of the results crossed the threshold into actual statistical significance. As an extension, we also performed a sensitivity analysis to see how much this increase in significance held under increased devations from ignorability; here we found that the results are in fact quite sensitive, with the p-values increasing quite dramatically even for $\Gamma = 1.2$.

As a final check on the assumptions from an observational point of view, we turned to subclassification on propensity score as an alternative to matching and to return to a Neymanian paradigm. Again, we found that the results were not significant.

In conclusion, it was quite informative to view the randomized quota experiment from an observational lens given how we were able to assess deviations from the properties of a completely randomized experiment - in the observational study paradigm, these are instead seen as assumptions. While we performed a fairly detailed analysis on the explicit bias experiment, there are still numerous observational techniques that can be applied to the remaining covariates, and the potential to apply the same analysis to the experiment on implicit bias. Overall, we view the amount of information gleaned from this causal re-analysis as a vindication of the techniques that have come from the field in recent years, and see numerous opportunities for rich re-analysis on this dataset.

References

[1] Amanda Clayton. Do Gender Quotas Really Reduce Bias? Evidence from a Policy Experiment in Southern Africa. *Journal of Experimental Political Science*, 5(3):182–194, 2018.

[2] Beaman Lori, Esther Duflo, Rohini Pande, and Petia Topalova. Political Reservation and Substantive Representation: Evidence from Indian Village Councils. *India Policy Forum*, 1:159–91, 2011.

[3] Bhavnani Rikhil. Do Electoral Quotas Work after They are Withdrawn? Evidence from a Natural Experiment in India. *American Political Science Review*, 103(1):23–35, 2009.