**Guide to Replication File**

The Long-Term Effects of Neighborhood Disadvantage on Political Behavior: The “Moving to Opportunity” Experiment

This analysis relies on two key datasets, neither of which can be made publicly available.

1. A snapshot of the national voter file from 2017 (i.e. including elections through the 2016 general) compiled by the voter data corporation L2. National voter file data is available for licensing and purchase through L2.
2. Data collected through the Moving to Opportunity Experiment by the Department of Housing and Urban Development (HUD). Access to this data is now administered by the Census Bureau; researchers interested in obtaining this data should contact the Census Bureau. We received access directly through HUD and indirectly through NBER under previous licenses.

In lieu of making these restricted datasets available, this replication archive consists of two components:

1. The code files we used to process and analyze the original data
2. A modified version of our final dataset that preserves individual privacy by dividing participants into <10—person clusters, with code to approximately reproduce our analyses (where possible) using the clustered data

We first discuss the methods used to produce results using the original data, then turn to a description of the clustered data.

The full, final dataset used for analysis is stored at the Census Bureau and can be obtained by researchers who meet privacy and data security requirements. The authors can provide further information about this process on request.

**Merging the voter file and MTO data**

The first step in our analysis involved merging these two datasets together. We linked MTO participants to their records in the voter file using information common to both sources: gender, first and last names, and birth dates.

Because we were not licensed to view personally identifiable information (PII) about MTO participants, including names and birthdays, the merge was performed by personnel at NBER. NBER housed and was licensed to use PII.

We provided NBER with the L2 national data file and code to execute the merge.

The merge proceeds in three steps.

1. *Generate clusters:* to make merging with the national voter file computationally tractable, we divide the MTO and voter file data into clusters based on gender and name. For example, for a male MTO participant named John Smith, we do not look for potential matches (i.e. generate the posterior probability of a match) in every record in the national voter file; instead, we might only look at a subset of male voters with names starting with Jo-.
2. *Run merge:* within each cluster, we identify the voter file records with the highest probability of being a match to each MTO participant.
3. *Export matches:* to preserve the privacy of MTO participants, NBER limited the number of potential matches and the precision of the voter file variables before providing us with the results of the merge.

The code files used in this process are in the folder Merge Files/.

The following code files are used to execute the merge (located in folder Merge Files/Merge Steps/):

1. 1\_cluster\_generation.R
   1. Takes in the MTO data and files listing all the first names of males and females in the L2 data.
   2. Identifies name clusters of a computationally tractable size and saves this list.
   3. Takes in the full voter file data, breaks it into clusters, and saves each cluster’s voter file data in a separate file.
   4. Exports summary information about each cluster.
2. 2\_cluster\_generation\_b.R
   1. Repeats the process in (1), for a larger cluster size.
3. 3\_main\_code.R
   1. Prepares for and executes the match\_function.R file, iterating over each set of clusters.
4. 4\_match\_function.R
   1. Takes in an L2 cluster file and the MTO data file
   2. Runs the fastLink algorithm to calculate potential matches between the voter file and MTO data
   3. Saves out, for each cluster, datasets of matched participants, unmatched participants, and summary statistics on the match
   4. Restarts RStudio and prepares for the next cluster to be run OR if all matches are complete, runs the reattach\_data file.
5. 5\_reattach\_data.R
   1. Reads in each cluster match file and aggregates them together.
   2. Calculates turnout rates based on the number of elections in which each participant did and could have voted (based on their birthdate and the elections present in their state’s voter file).
   3. Merges in Census data on the demographic composition of the Census tracts in which each family lived between treatment and 2010.
   4. Saves out datasets, removing name, birth date, and tract information.

The following code files were used by NBER to further anonymize the data and merge it with internal MTO survey files (located in folder Merge Files/Output Steps/): This code was provided to us by the NBER personnel who conducted the merge, and since we cannot re-run or test this code on the data it was executed on, we have made no simplifying edits.

1. 01\_create\_vote\_files\_a\_to\_c.do
   1. Takes in final merge data files and internal MTO files, linking them together using an NBER-created pseudo-id (to eliminate the possibility of us linking our data to the original MTO data).
   2. Rounds turnout rates to the nearest .1 and removes matches outside the top 5 posterior probabilities.
2. 02\_create\_survey\_files\_d\_n.do
   1. Takes in further MTO survey files and merges them with our data using the pseudo-id.
3. 03\_create\_revised\_vote\_file.do
   1. Revises the methods used to round the turnout rates.

After completing their cleaning and processing steps, NBER provided us with the following data files, each of which contains a common individual pseudo-id and some set of variables from the voter file or MTO’s internal files:

1. a\_mto\_fweights\_vote\_pseudo.dta
   1. Includes survey weights for each survey wave.
2. b\_mto\_vote\_match5\_pseudo\_v2.dta
   1. Includes the voter file match.
3. d\_mto\_base\_hhld\_pseudo.dta
   1. Includes baseline survey/administrative data for each household.
4. e\_mto\_base\_pers\_pseudo.dta
   1. Includes baseline survey/administrative data for each person.
5. f\_mto\_int\_rawsvy\_ad\_pseudo.dta
   1. Includes responses from the interim survey for adults.
6. h\_mto\_fin\_rawsvy\_ad\_pseudo.dta
   1. Includes responses from the final survey for adults.
7. i\_mto\_fin\_rawsvy\_yt\_pseudo.dta
   1. Includes responses from the final survey for youth.
8. j\_mto\_fin\_roster\_pseudo.dta
   1. Includes data on each person (roster) from the final survey.
9. k\_mto\_int\_adult\_analysis\_pseudo.dta
   1. Further data on adults from the interim survey.
10. l\_mto\_int\_adult\_analysis\_pseudo.dta
    1. Includes data from the interim survey for youth.
11. m\_mto\_int\_roster\_pseudo.dta
    1. Includes data on each person (roster) from the interim survey.
12. n\_mto\_fin\_analysis\_pseudo.dta
    1. Further data on participants from the final survey.
13. new\_pseudo\_id\_ra\_year.dta
    1. Includes data on respondents’ birth years.

After this point, all analyses were conducted by the authors in a secure data environment.

**Analysis files**

The following code files were used for data processing and analysis (located in the Analysis files/ subfolder):

1. 1\_combine\_data.R
   1. Takes in all the raw data files from NBER and merges them together
   2. Creates and exports a dataset including only one match for each matched MTO participant, combined\_dataset.csv
   3. Creates and exports a dataset with all possible matches for each observation, combined\_dataset\_withdups.csv
2. 2\_match\_comparison.R
   1. Takes in the combined dataset
   2. Calculates the differences in treatment effects on various outcomes separately for the matched and full samples, creating intermediate datasets of the differences
   3. Uses these intermediate datasets to create and export figure 3 from the main text
3. 3\_results.R
   1. Takes the combined dataset
   2. Conducts all analyses in the main text and appendix not created in other files
4. 4\_multiple\_matches.R
   1. Takes the combined dataset including duplicate matches
   2. Runs the key regressions on the with-duplicates data to produce Appendix section 4 tables
5. 5\_hte\_models.R
   1. Takes the combined dataset
   2. Re-runs the subgroup models presented in Figure 5 of the main text, exporting them in tabular format to produce the results found in the extended tables supplementary information
6. 6\_hte\_bart.R
   1. Takes the combined dataset
   2. Uses the R package bartMachine to search for possible sources of treatment effect heterogeneity for the results reported in appendix section A6
7. 7\_ces\_baseline.R
   1. Takes in the 2020 Cooperative Election Study from the Harvard Dataverse
   2. Calculates baseline registration rate estimates for Appendix section A1.1

**Clustered Data**

Because HUD procedures prohibit sharing any figure based on data from 10 or fewer individuals, we were able to construct a dataset that grouped participants into clusters with at least 11 people. In this dataset, each row represents 11 or more participants and shows the mean value for each variable for the participants in the cluster.

Many of our analyses can be conducted using this clustered dataset. However, it is important to note that analyses using the clustered dataset *will not exactly recover* estimates obtained from the full data. Averaging variables over the cluster members induces measurement error, as we discuss further below.

Our primary goal in creating the clusters was to allow for as-close-as-possible replication of our key models regressing participation measures on treatment assignment and study site, including randomization weights. Our secondary goal was to allow for replication of results that require subsetting the data by gender, as gender was an important factor in the success of the match process.

To create clusters, we first divided all participants by age group, treatment assignment, and study site. Each of these age x treatment x site groups has at least 11 members, so all clusters are homogeneous on these dimensions. We further divided these groups by randomization weight, which reflects each subject’s probability of assignment to each treatment group; this weight varies by site and over time. All but a handful of groups by randomization weight have at least 11 people. For the remaining groups with fewer than 11 people, we combined observations with another group identical on age x treatment x site and with the next-closest weight.

These age x treatment x site x groups ranged in size from 11 people (the minimum) to over 300. For groups large enough to accommodate multiple clusters (i.e. at least 22 people), we further divided participants into clusters with as close to 11 people as possible. Because of the importance of gender to our analyses, we attempted to keep as many clusters homogeneous by gender as possible. To do this, we assigned clusters to group members within gender groups. For example, in a group with 22 people, including 10 women and 12 men, we would assign all the women and 1 man to a single cluster, and all the men to a single cluster. This rule resulted in 94% of clusters containing at least 90% men or at least 90% women. Cluster assignment within groups within genders was random.

Within each cluster, we then took the mean value of all variables needed to conduct our main analyses. The result is a 1389-row dataset where each row represents 1 cluster of 11-21 participants. The variables in this dataset are identical to those in the full final analysis dataset (including recodes executed in the file “3\_results.R”), with the addition of an indicator for each cluster (“cluster”) and a measure of the number of people in the cluster (“n\_cluster”). We also exclude variables which are not used for the analyses possible using the cluster data.

The clustered dataset can be used to replicate many of the key analyses in the paper and appendix. However, there are three key limitations to the clusters.

First, the clustered dataset cannot be used for analyses that require subsetting the data on variables for which the clusters are not homogeneous. For example, models that condition on a participant’s voter registration status (like models predicting post-registration turnout) cannot be run, because most clusters contain a mix of people who were and were not registered to vote; we cannot limit an analysis to only those who were registered. Other analyses that subset on race or marriage status also cannot be run. Because we ensured nearly all clusters were homogeneous on gender, analyses subsetting on gender can be run, but the results do not match those from the full dataset as well because the gender heterogeneous clusters must be excluded.

Second, we cannot recover the standard errors of estimates from the full data. The standard errors of estimates from the clustered data will differ for three reasons: a) there are about 11 times fewer observations in the clustered dataset, which decreases analyses’ degrees of freedom and increases the size of the standard errors, b) treating each cluster’s mean as a single observation masks variation within the cluster’s component individuals, so the variance in each estimate with clustered data will be understated relative to the full dataset, and c) analyses with the full data include family-level clusters, which cannot be preserved in the clustered dataset; without accounting for family clustering, the clustered dataset analyses will understate the standard errors.

Finally, measurement error will cause point estimates from analyses using clustered data to vary slightly from those using the full data reported in the paper. Because cluster assignments are block randomized, the mean of each variable in the clustered data will be the same in expectation as those in the full data; however, the means will vary slightly due to randomness in the assignment process. This variation will lead to point estimates that differ from those using the full data. For our key analyses regressing participation on treatment and study site, the only variable averaged over clusters is the participation outcome; measurement error from clustering is confined to this variable, so point estimates from the clustered data closely match those from the full data. However, models that include more variables averaged within clusters—like the models adding many controls—will include error for each of these additional variables, so point estimates will vary more widely from those using the full data.

With these limitations in mind, we produced a code file to replicate point estimates for as many of the paper’s analyses as possible using the clustered data. The code file is structured identically to the key results file for the full data (“3\_results.R”), and it includes notes for each section when an analysis cannot be replicated or somehow departs from the method used for the full data.

The clusters were created and analyzed using the following code files:

1. create\_clusters.R
   1. Takes the final analysis dataset and produces clusters.csv, a file with the mean value for each variable for >10-person clusters.
2. cluster\_results.R
   1. Replicates all possible analyses in the main text and appendix using the clusters.csv dataset. Keep in mind that point estimates will not match exactly, due to measurement error, and that there will be more error in models that draw on more variables (especially those with many controls).