Replication package for: A Welfare Analysis of Tax Audits Across the Income Distribution

William C. Boning, Nathaniel Hendren, Ben Sprung-Keyser, Ellen Stuart

October 22, 2024

1 Overview

This replication package describes, but does not contain, the restricted-access code for data cleaning and analysis. The tabulations used to create the figures and graphs in the paper as well as the code that produces those tables and figures are not restricted and are provided as part of this replication package. The replication package is available at the XX repository.

In this document, we describe the underlying sources of data, the code used to clean and construct the analysis dataset, and the code used to create the tables and figures in the paper and appendix.

2 Data Availability

The majority of the data used in this paper is derived from tax data, accessed by William C. Boning, an employee of the United States Department of the Treasury, and Ellen Stuart, who gained access through the IRS Statistics of Income Joint Statistical Research Program. Under section 6103 of the Internal Revenue Code, this data cannot be shared. For this reason, we are unable to include this data in the replication package. Researchers can apply for access to the IRS tax data via the Joint Statistical Research Program: https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program. Additionally, the IRS treats the associated metadata including variable names and table names as restricted data, and the code that uses those variable names cannot be shared.

Individuals with access to the tax data should contact William C. Boning (William.Boning@treasury.gov) to access the data and code used for this project.

The main source of data used in this analysis is a database housed on IRS servers that contains administrative data on the universe of tax returns, information

returns, transactions on taxpayers' accounts, and enforcement actions (Internal Revenue Service 2023). We also use information from a series of internal IRS studies of program costs (Internal Revenue Service 2010-2020), internal documentation of the sample design of the NRP Form 1040 studies from 2006-2014 (Internal Revenue Service 2007-2015), and data from the 2023 IRS Taxpayer Compliance Burden Survey (Internal Revenue Service 2024). We use publicly available General Schedule (GS) & Locality Pay Tables (Office of Personnel Management 2003-2021) and inflation data (Bureau of Labor Statistics 2022).

2.1 Administrative Data

From the administrative tax database (Internal Revenue Service 2023), we make use of information from the following set of tax forms and files:

- Form 1040, including information about income, deductions and credits, and tax liability.
- A dataset of information from National Research Program (NRP) audits of Forms 1040, with information on examiners' proposed adjustments to line items.
- A dataset of IRS enforcement actions, which includes information about the stage of the enforcement process, hours spent by IRS staff, additional tax and penalties assessed, and revenue remitted.

We extract this data using Stata 17. As the code that does this extraction would need to be fully redacted, it is not included in this replication package. Researchers with access to the tax data should contact William C. Boning for access to this code.

2.2 Internal Accounting Data

We use a series of reports (Internal Revenue Service 2010-2020) that capture internal program-level accounting data compiled by the IRS' CFO. These reports span fiscal years 2010 to 2020 and contain information about programs like field audit and field collection, such as total and line-item costs. We received them from the IRS CFO office on November 18, 2021. While these studies are not publicly available, rounded values of the relevant numbers are included in Appendix B of the paper.

2.3 Internal Form 1040 National Research Program Sample Design Documentation

We use a series of documents (Internal Revenue Service 2007-2015) that describe the sampling design including strata and strata weights for the 2006-2014 NRP Form 1040 studies. We use the included strata descriptions and strata weights to construct matched returns for the NRP returns. These files are restricted

to internal IRS use only. William C. Boning accessed them from internal IRS documentation libraries on April 5th, 2022.

2.4 Internal 2023 IRS Taxpayer Compliance Burden Survey Data

We use a tabulation of the results from an internal IRS 2023 survey of taxpayers (Internal Revenue Service 2024) asking about the burden they experienced to comply with post-filing enforcement. We received a file of tabulated results summarizing reported burden (time, money, and total monetized burden) from the survey by enforcement area and TPI bin on March 22nd, 2024. The survey and tabulated results are restricted to internal IRS use only.

2.5 Other miscellaneous data

We use two publicly available datasets in this project.

First, we use the GS payscale data from 2003-2021 (Office of Personnel Management 2003-2021), available at https://www.opm.gov/policy-data-oversight/pay-leave/salaries-wages/, to estimate the direct wage costs associated with each audit.

Second, we use the CPI Inflation Calculator from the U.S. Bureau of Labor Statistics to adjust for inflation. The CPI Inflation Calculator is available at https://data.bls.gov/cgi-bin/cpicalc.pl.

Summary files containing this data are included in the replication package William C. Boning can provide to replicators with access to the IRS data.

3 Computational requirements

The analysis was performed on IRS servers. Stata 17 was used for pulling the data from the database, processing the raw data inside the secure environment, performing the analysis, and creating the tables, figures, and in-text numbers.

3.1 Directory structure

The main project directory on the IRS servers contains the following subdirectories:

- `/ado`
- `/csv`
- `/do`
  - `/multipliers`
  - `/nrp_deterrence`

    - /ops_deterrence
    - /percentiles
- /make_figures_and_tables
    - /code
        * /analysis
        * /prep
    - /data
        * /cleaned
        * /figures
        * /outside
    - /Figures
        * /paper
        * /slides
    - /Tables
- /dta
- /gph
- /nrp_docs
- /log
- /pdf_arial
- /sql
- /ster
- /txt
- /xlsx

The following user-created packages are required and can be installed from `SSC` (as happens automatically at the beginning of `makefile_rev_costs.do`)

- ftools (Correia 2016)
- gtools (Bravo 2018)
- estout (Jann 2004)
- coefplot (Jann 2013)

The entirety of the Stata code takes approximately 80 hours to run on the IRS server environment, depending on server congestion.

4 Instructions

To replicate the dataset creation and analysis, replicators should first gain access to tax data housed on IRS research servers. Members of the public will not be able to do this. Replicators with access to the restricted tax data should contact William C. Boning (William.Boning@Treasury.gov) for access to the restricted data and code files used for this project.

After obtaining the restricted data and code files, replicators can replicate the costs and revenues results by updating the project folder reference at the top of, and then running, `do/makefile_rev_costs.do`. `do/makefile_rev_costs.do` runs a series of do files that perform the following steps:

1. Pull and prepare data
    (a) Import GS payscale data
    (b) Import inflation data
    (c) Import burden survey data
    (d) Pull and prepare basic NRP data to merge to revenue data
    (e) Pull and prepare data on TPI percentile thresholds
    (f) Pull exam data from revenue database and perform initial data cleaning
    (g) Pull income information from Forms 1040 data
    (h) Pull risk scores
    (i) Tabulate risk scores
    (j) Pull collections data from revenue database and perform basic data cleaning
    (k) Pull collections data from revenue database only for cases ever in field collections and perform basic data cleaning
    (l) Pull ZIP codes used to determine the location adjustment for GS scale wages
    (m) Clean up exam data before making unique-by-audit-id dataset
    (n) Make unique-by-audit-id dataset
2. Make tabulations and statistics used in the paper
    (a) Tabulate unique-by-audit-id dataset by income
    (b) Tabulate imputed taxpayer burden from unique-by-audit-id dataset by income

(c) Tabulate unique-by-audit-id dataset by income and whether any additional tax was assessed.

(d) Tabulate imputed taxpayer burden from unique-by-audit-id dataset by income and whether any additional tax was assessed.

(e) Tabulate costs and revenues for NRP audits, with NRP weights.

(f) Convert tabulations into present-value revenue and cost numbers used in the paper.

(g) Convert tabulations by whether any additional tax was assessed into revenue and cost numbers used in the paper.

Replicators who have obtained access to the restricted data and code can replicate making the database-driven inputs to the cost multipliers used in the paper by running `do/multipliers/makefile_multipliers.do`. `do/multipliers/makefile_multipliers.do` runs a series of do files that perform the following steps:

1. Pull and clean cost data from revenues/costs database.
   (a) Pull cost data for individual examinations cases, in two steps
   (b) Pull cost data for individual collections cases, in two steps
   (c) Collapse individual cases cost data
   (d) Pull cost data for business examinations and collections cases, in two steps
   (e) Collapse business cases cost data
2. Make totals used to calculate costs multipliers.
   (a) Make .csv file of total costs for use in multipliers calculations
   (b) Make .csv file of total costs for field collection for sues in multipliers calculations.

Replicators who have obtained access to the restricted data and code can replicate the NRP deterrence by running `do/nrp_deterrence/makefile_match_deterrence.do`. `do/nrp_deterrence/makefile_match_deterrence.do` runs a series of do files that perform the following steps:

1. Pull and construct the dataset:
   (a) Pull NRP Form 1040 studies with study year and strata
   (b) Create file with average corrections overall and by TPI
   (c) Pull potential control observations with variables needed to make strata
   (d) Make strata for potential controls for each study year

(e) Compile and prepare unmatched data from NRP

(f) Match NRP to controls

(g) Pull and append panel of information from Forms 1040 for NRP and matched controls

(h) Prepare taxpayer-level fields for the matched sample.

(i) Prepare the taxpayer-tax-year fields in the matched panel.

(j) Make audit indicators

(k) Merge on audit indicators

(l) Create file with average corrections overall and by TPI

(m) Pull potential control observations with variables needed to make strata.

(n) Make strata for potential controls for each study year

(o) Compile and prepare unmatched data from NRP.

(p) Match NRP to controls.

(q) Pull and append panel of information from Forms 1040 for NRP and matched controls.

(r) Prepare taxpayer-level fields for the matched sample.

(s) Prepare the taxpayer-tax-year fields in the matched panel.

(t) Make indicators for audit for each tax year

(u) Merge with indicators for audit for each tax year

2. Produce analysis tabulations and regression results:

(a) Tabulate means by NRP or by NRP and study year

(b) Run single difference regressions

(c) Run difference-in-difference regressions by TPI quantile

(d) Run single-difference regressions by grouped NRP study years

(e) Run single-difference regressions by presence of business income

(f) Run single-difference regressions by whether there was any audit correction

(g) Run difference-in-difference regressions by whether there was any audit correction

Replicators who have obtained access to the restricted data and code can replicate the deterrence results for matched operational audits in Appendix Figure A.XIV

by running the file `do/ops_deterrence/makefile_ops_deterrence.do`, which calls a series of do files that perform the following steps:

1. Pull deciles of field/office-audited risk score distribution within tax year and exam class
2. Prepare risk score deciles for use with the database
3. Pull deciles of field/office-audited TPI distribution within tax year and exam class
4. Pull panel data for treated and matched controls.
5. Prepare panel data and plot means over time for treated and matched control taxpayers.

Researchers who have obtained access to the restricted data and code interested in replicating the figures and numbers reported in the paper must then run `metafile.do`, which is included in the `make_tables_and_figures` subfolder. A public version of `metafile.do` is also available for researchers only interested in replicating the figures in the paper. The pre-processing code is removed from the publicly available version of the code. The restricted-access `metafile.do` runs a series of do files that perform the following steps:

1. Clean the collapsed data
   (a) Add labels to all the variables
   (b) Calculate cost components for NRP, correspondence, and operational audits by year
   (c) Calculate revenue components for NRP, correspondence, and operations audits by year
   (d) Apply 3% discount to future cost and revenue
   (e) Merge in multipliers
   (f) Compute revenue over cost ratio by audit type and year
   (g) Split data by TPI bin
2. Create figures in the paper
   (a) See the table below for the do files that correspond to each figure

The concordance between do files and figures is included in Table 1 below. Some figure formatting (e.g. text callouts, legend positions) are done in powerpoint. The tables are made entirely in Excel using data output from the steps outlined above. For table one and all appendix tables, see Tables/table_1.xlsx and Tables/appendix_tables.xlsx in the replication package, respectively. Additionally, there are several numbers mentioned in the text of the paper that are not derived from any tables or figures. These numbers are calculated by .do files called by the restricted-access version of metafile.do. While we cannot provide the calculation

.do files in the public replication archive, a concordance between these in-text numbers and the restricted-access .do files that produce the numbers is included in Table 2.

Table 1: Concordance between do files and tables/figures

	Analysis .do file	**Output file(s)**
Figure 1	analysis/waterfall_cost_rev_future_disc.do 1	waterfall_costs_rev_ops
Figure 2	analysis/scatter_cost_rev_rc_inc.do 1	scatter_costs_rev_tpi_ops
		scatter_rc_tpi_ops
Figure 3	analysis/line_rc_audit_prob.do ops	connected_hours_wages_year_ops
		line_rc_prob_audit_ops
Figure 4	analysis/rc_45_degree.do	scatter_rc_tpi_45degree
	analysis/bar_roc.do	bar_roc_exp_inc
Figure 5	analysis/scatter_cost_rev_rc_inc.do 1	scatter_rc_alt_Oh_tpi_ops
Figure 6	analysis/event_study_taxes_paid.do	coeffplot_detmult_tpi
	analysis/coeffplot_detterrence.do	event_study_taxes_paid
Figure 7	analysis/coeffplot_deterrence_plus_rev.do	coeffplot_marginal_rc_deterrence_tpi_appendix
		coeffplot_marginal_rc_deterrence_tpi
Figure 8	analysis/waterfall_mvpf.do	waterfall_mvpf_20_30
		waterfall_mvpf_90_100
Figure 9	analysis/scatter_mvpf.do	scatter_mvpf_tpi_det
Appendix Figure 2	analysis/waterfall_cost_rev_future_disc.do 2	waterfall_costs_rev_corr
Appendix	analysis/scatter_rev_costs_corr.do	scatter_costs_rev_tpi_corr
Figure 3	analysis/connected_corr_inc.do	scatter_rc_tpi_corr
		connected_hours_wages_inc_corr
		connected_audit_rates_corr
Appendix Figure 4	analysis/scatter_alt_oh_allocations.do ops	scatter_costs_rev_tpi_alt_oh_allocation
		scatter_rc_tpi_alt_oh_allocation
Appendix Figure 5	analysis/cum_rev_shares_inc_w99.do	rev_share_2003_ops_wins
		costs_share_2003_ops_wins
Appendix Figure 6	analysis/scatter_cost_rev_rc_year.do	connected_rc_year_ops
		connected_costs_rev_year_ops
Appendix Figure 7	analysis/scatter_rc_inc_corr.do	scatter_rc_tpi_10p_cor_pctl5_ops
		scatter_rc_tpi_20p_cor_pctl5_ops
Appendix Figure 8	analysis/scatter_labor_hours_inc	scatter_hours_wages_tpi_ops
Appendix Figure 9	analysis/rc_45_degree.do	scatter_rev_tpi_45degree
Appendix Figure 10	analysis/line_rc_audit_prob.do nrp	connected_rc_labor_year_nrp
		connected_hours_wages_year_nrp
		line_rc_prob_audit_nrp
Appendix Figure 11	analysis/scatter_cost_rev_rc_inc_year_nrp.do	connected_rev_tot_inc_yr_nrp
		connected_costs_tot_inc_yr_nrp
Appendix Figure 12	analysis/event_study_taxes_paid_bus_income.do	event_study_taxes_paid_bus_income
Appendix Figure 13	analysis/event_study_taxes_paid_4yr_groups.do	event_study_taxes_paid_alt_years
Appendix Figure 14	analysis/event_study_taxes_paid_2yr_groups.do"	event_study_taxes_paid_alt_years_v2
Appendix Figure 15	analysis/event_study_taxes_paid_ops.do	event_study_taxes_paid_ops
Appendix Figure 16	analysis/coeffplot_deterrence.do	coeffplot_detmult_tpi_unwinsorized
Appendix	analysis/scatter_burden.do	scatter_hours_spent
		scatter_Total Burden
		scatter_Spending
Appendix Figure 17	analysis/no_change_costs.do	no_change_tpi
Appendix Figure 18	analysis/scatter_burden.do	scatter_Hours Spent_changesplit
		scatter_Spending_changesplit
		scatter_Total Burden_changesplit
Appendix Figure 19	analysis/coeffplot_eitc.do	event_study_eitc_noref

Table 2: Concordance between .do files and in-text numbers

Text	Page	Do file
The evidence suggests that, in the case of a major audit expansion, approximately 27% of total costs are likely to be fixed	3	waterfall_cost_rev_future_disc.do
In present discounted value, these additional tax payments raise 3.2 times as much revenue as the revenue raised from the initial audits.	4	analysis/coeffplot_deterrence.do
We estimate that in our 7–11 year sample frame we capture 96–99% of total labor hours spent on audits.	12	cum_rev_shares_inc_w99
We estimate that 86–96% of all revenue is collected within 7–11 years after filing, with the exact percentages varying by TPI bin	13	cum_rev_shares_inc_w99
The IRS audited 0.92% of all returns from tax year 2010 but just 0.56% of returns from 2014	15	line_rc_audit_prob
Footnote 22: In 2010, 0.21% of all returns were selected for in-person audit compared to 0.11% in 2014.	15	line_rc_audit_prob
Footnote 24: We find similar results for the revenues per audit for correspondence audits: there was a 38% decline in the audit rate for correspondence audits from 2010 to 2014. The average return to a correspondence audit was $1,194 in 2010 and $1,294 in 2014.	15	line_rc_audit_prob
By comparison, in our baseline estimates total marginal non-labor costs are 34.3% as large as total labor and fringe costs	18	waterfall_cost_rev_future_disc.do
The yearly impact on taxes paid is around $300 per audited taxpayer, which is approximately 30% of the $1,026 in revenue collected by the initial audit.	20	event_study_taxes_paid_bus_income.do
When we sum the deterrence effect over 14 years and apply a 3% discount rate, we estimate that an average NRP audit produces $3,258 in additional taxes paid via the individual deterrence channel	20	event_study_taxes_paid.do
For taxpayers without business income the future revenue collected is roughly 3.81 times initial NRP revenue. Consistent with previous literature, we find slightly less persistent deterrence effects for taxpayers with business income, with a ratio of 2.81	20	event_study_taxes_paid_bus_income.do
Using a matched difference-in difference-design, we obtain a multiplier of approximately 2.5, consistent with the results for NRP audits	20	event_study_taxes_paid.do
In fact, this calculation suggests that such a policy is welfare enhancing as long as the average welfare weight on audited high income individuals is less than 1.13 times the welfare weight placed on audited low-income individuals.	26	waterfall_mvpf.do
We find that raising $1 from audits imposes private welfare costs of $1.08 on noncompliant taxpayers and private welfare costs of $0.07 on compliant taxpayers.	28	waterfall_mvpf.do

References

Bravo, Mauricio Caceres, "GTOOLS: Stata module to provide a fast implementation of common group commands," 2018. Statistical Software Components S458514, Boston College Department of Economics, https://ideas.repec.org/c/boc/bocode/s458514.html. 3.1

Bureau of Labor Statistics, "Consumer Price Index Inflation Calculator," 2022. Bureau of Labor Statistics, Washington DC, https://data.bls.gov/cgi-bin/cpicalc.pl, accessed 2022-07-14. 2

Correia, Sergio, "FTOOLS: Stata module to provide alternatives to common Stata commands optimized for large datasets," 2016. Statistical Software Components S458213, Boston College Department of Economics, https://ideas.repec.org/c/boc/bocode/s457874.html. 3.1

Internal Revenue Service, "Form 1040 NRP Sample Designs 2006-2014," 2007-2015. Washington DC, accessed 2022-04-05. 2, 2.3

_ , "Cost Based Performance Measures," 2010-2020. Washington DC, accessed 2023-04-15. 2, 2.2

_ , "Administrative tax data files 2003-2021," 2023. Washington DC, accessed 2023-04-15. 2, 2.1

_ , "2023 IRS Taxpayer Compliance Burden Survey," 2024. Department of Treasury, Washington DC, accessed 2024-03-22. 2, 2.4

Jann, Ben, "ESTOUT: Stata module to make regression tables," 2004. Statistical Software Components S439301, Boston College Department of Economics, https://ideas.repec.org/c/boc/bocode/s439301.html. 3.1

_ , "COEFPLOT: Stata module to plot regression coefficients and other results," 2013. Statistical Software Components S457686, Boston College Department of Economics, https://ideas.repec.org/c/boc/bocode/s457686.html. 3.1

Office of Personnel Management, "General Schedule (GS) & Locality Pay Tables," 2003-2021. Office of Personnel Management, Washington DC, https://www.opm.gov/policy-data-oversight/pay-leave/salaries-wages/, accessed 2021-12-04. 2, 2.5