# The IRS Research Bulletin

## Proceedings of the 2024 IRS / TPC Research Conference

**Research, Applied Analytics & Statistics**

Publication 1500 (Rev. 6-2025) Catalog Number 11546J Department of the Treasury **Internal Revenue Service** www.irs.gov

*Papers given at the*

# *14th Annual Joint Research Conference on Tax Administration*

## *Cosponsored by the IRS and the Urban-Brookings Tax Policy Center*

**June 14, 2024**

Compiled and edited by Corbin Miller*
Research, Applied Analytics, and Statistics, Internal Revenue Service

*Prepared under the direction of Reza Rashidi, IRS Acting Chief Research and Analytics Officer

## Foreword

This edition of the IRS Research Bulletin (Publication 1500) features selected papers from the 14th Annual IRS-Tax Policy Center Research Conference held on June 13, 2024, at the Urban Institute in Washington, DC. Conference presenters and attendees included researchers from many areas of the IRS, officials from other U.S. government or international agencies, and academic and private sector experts on tax policy, tax administration, and tax compliance. This marks our second in-person conference after several years of being virtual only. Videos of the presentations are archived and available on the Tax Policy Center website.

The conference began with welcoming remarks by Robert McClelland, Senior Fellow of the Urban-Brookings Tax Policy Center (TPC), and Barry Johnson, then Chief Data and Analytics Officer in the IRS Office of Research, Applied Analytics and Statistics (RAAS). The remainder of the conference included ses-sions on innovations in the use of tax data, tax avoidance, tax audits, and reducing the filing burden. The key-note speaker was Danny Werfel, Commissioner of Internal Revenue, who offered insights on his vision for the future of the IRS and using data to improve public service.

We trust that IRS executives, managers, employees, stakeholders, and tax administrators will find this volume to be valuable and increase their access to the latest tax administration research. Furthermore, it is our hope that this research contributes to improvements in tax administration, sparks additional innovative research, and increases cooperation among tax administration researchers everywhere.

## Acknowledgments

This year's IRS-TPC Research Conference was the result of preparation over many months by dozens of people. The conference program was assembled by a committee representing research organizations throughout the IRS. Members of the program committee included: Robert McClelland (TPC, co-chair), Brett Collins (RAAS, co-chair), Devi McKalko (RAAS), Bizuayehu Bedane (RAAS), Anne Dayton (Small Business/Self Employment, SBSE), Trevor Speed (SBSE), Chris Wilson (RAAS), April Harding (Online Services), and Brittany Jefferson (Taxpayer Services).

We also wish to acknowledge Leonard Burman (TPC), William Boning (Treasury, Office of Tax Analysis), Arnstein Øvrum (Norwegian Tax Administration), and Robert Weinberger (TPC) for serving as discussants and Brittany Jefferson (IRS, Wage and Investment), and Devi McKalko, Melissa Vigil, and John Guyton (IRS, RAAS) for serving as moderators.

This volume was prepared by Lisa Smith and Daniel Martinez (layout and graphics), Spencer Adams (editor), and Beth Kilss (contractor), all members of the IRS Statistics of Income Division, in co-ordination with Corbin Miller (economist, SOI Division) who was the liaison with the conference presenters in preparing the papers for publication. The authors of the papers are responsible for their content, and views expressed in these papers do not necessarily represent the views of the Department of the Treasury or the Internal Revenue Service.

We appreciate the contributions of everyone who helped make this conference a success.

Reza Rashidi
Acting Chief Data and Analytics Officer, Internal Revenue Service

# 14th Annual IRS-TPC Joint Research Conference on Tax Administration

## Contents

## 4. Simplifying the Filing Burden



## 5. Appendix

# 1



## Harnessing Data for Better Research

**Bryant ◆ Hertz ◆ Pierce ◆**
**Beckhusen ◆ Landivar ◆ Laughlin ◆ Sanders**
**Grusky ◆ Hout ◆ Martin-Caughey ◆ Miranda**

**Bedane**

# A Large-Scale, High-Quality U.S. Occupational Database: Results from Merged IRS and ACS Write-Ins*

*Victoria L. Bryant, Thomas N. Hertz, Kevin Pierce (IRS, SOI), Julia Beckhusen, Liana Christin Landivar, Lynda Laughlin, Carl Sanders (U.S. Census Bureau), David B. Grusky (Stanford University), Michael Hout (New York University), Ananda Martin-Caughey (Brown University), Javier Miranda (Halle Institute for Economic Research, University of Jena)*

## 1. Introduction

Administrative data, especially income tax returns, have greatly advanced social science knowledge about the persistence of income across generations (Mazumder (2005) and Mitnik, Bryant, and Grusky (2024)). We have learned much about the extent of geographic differences in mobility (Chetty, Hendren, et al. (2014)), about the causal effects of place on mobility (Chetty and Hendren (2018)), and about cross-national differences in mobility (Corak (2013) and Mitnik, Bryant, Grusky, and Weber (2015)). Further, administrative income tax returns data are key to modernization in statistical agencies. For example, the U.S. Census Bureau uses individual tax return data provided by the IRS to enhance the accuracy and comprehensiveness of its data in the Nonemployer Statistics program[1] or the National Experimental Wellbeing Statistics experimental income and poverty statistics.[2]

But there is much more that we could do with these types of data to improve our knowledge and statistical programs. There are three key challenges that can only be addressed in a cost-effective manner by developing the capacity to analyze occupations with tax return data. These new occupation data, if developed and made available, would make it possible (a) to secure better estimates of long-term trends in mobility, (b) to analyze economic and occupation mobility simultaneously and thus reconcile apparent inconsistencies in the trend data and explore possible tradeoffs between them, and (c) to peer into the activities workers perform inside the firms that employ them, thus allowing us to better understand production and firm outcomes. From a statistical program perspective, having this type of data opens a range of opportunities to further enhance statistical programs, and their underlying data. Broadly speaking, these types of data would open a wealth of new research opportunities.

From a research perspective, the public cares about social mobility, but relies too much on anecdote. Scholars cannot fill the evidence gap because existing data fall short. We have very little post-1973 evidence on occupation mobility; the estimates of trends in occupation and economic analyses of mobility conflict with one another; and the possibility of systematic underestimates of mobility arise when occupation and economic mobility are examined in isolation from one another.

These problems can be addressed by exploiting the occupation fields available in tax return data. However, the information on occupation in tax records is not in a form that analysts can currently use. While taxpayers are asked to write their occupations on their tax forms, the Internal Revenue Service makes no effort to validate those entries, and there has been relatively little systematic research on occupation that draws on tax return data.

In this research program, we propose to machine-code the occupational information on tax records and classify and score that information. To do so, we will work with a tax return file linked to American Community Survey (ACS) records that already have occupational data suitably coded. That will allow us to assess the accuracy of our algorithm and characterize the measurement error in the coder.

An accurate automatic coder that can be deployed accurately at scale will allow constructing an occupational database with a sample size within an order of magnitude of the full U.S. worker universe. This reflects a multiple order of

* The views expressed are those of the authors and not those of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product. This research was performed at a Federal Statistical Research Data Center under FSRDC Project Number 2596. (CBDRB-FY23-P2596-R10780).

[1] https://www.census.gov/programs-surveys/nonemployer-statistics/technical-documentation/methodology.html

[2] https://www.census.gov/data/experimental-data-products/national-experimental-wellbeing-statistics.html

magnitude increase in the yearly records that can be used in economic and sociological analysis, from around 5 million workers in the 2019 ACS to nearly 130 million workers in Tax Year 2019 IRS records. When linked with other historical data sets, the resulting dataset may then be used to analyze occupation and income mobility simultaneously by estimating, for example, the intergenerational elasticity of incomes within and between occupations.

Some countries have comprehensive statistical systems collecting worker-level occupation that facilitate this kind of analysis. Our proposal to classify and score the occupational field in U.S. tax returns may be understood as a step toward developing a comparable U.S. statistical system. As Grusky and Cumberworth (2010) argued, an American comprehensive monitoring system would allow us to measure trends in income, occupational, and educational mobility at once, allowing us to identify which types are changing and which are not (without being confounded by countervailing changes in other forms). If we succeed in extracting reliable occupational data from tax records, then we can approximate that kind of analytical leverage in the United States, at least with respect to income and occupation.

These data are especially attractive because post-1973 data on occupational mobility are only available in small surveys. The tax return data could be combined with Census linkages from earlier time periods to estimate long-run trends in opportunity in the United States. Because the U.S. Census Bureau has gathered occupational data since 1870, Long and Ferrie (2013) used fathers and sons linked in 19th century data and estimated change in mobility over time by comparing the 19th century census-based mobility tables with the CPS-based mobility tables in Featherman and Hauser (1978). Our project will help fill in the crucial gap between 1973 and the present.

The tax records we propose to use include a short description of each taxpayer's occupation on the Form 1040 and information about the taxpayer's employer on the Form W-2. Because the occupation data are unstandardized text, they have only rarely been used. If they could be converted—with help from selected material on other forms—to a standard format, then they can be exploited for a host of basic science and policy analyses. We have stressed above their usefulness for analyses of mobility but in fact a wide range of basic science and policy applications would open up (some of which will be noted below).

The primary constraint to this point in using the IRS occupation write-in data is the "free response" nature of the taxpayer information. Computing and methodological constraints had prevented quantitative analysis of free text responses until recent developments in computational language modeling; see Minaee et al. (2024) for a review. Our use of recent advances in Large Language Models allows us to overcome the computational and modeling issues associated with free text response and can further serve as an example for statistical agencies looking to utilize their free-text-response data.

## 2. Data

Our data come from two primary sources: IRS tax data and the American Community Survey.

**TABLE 1. Sample Counts**

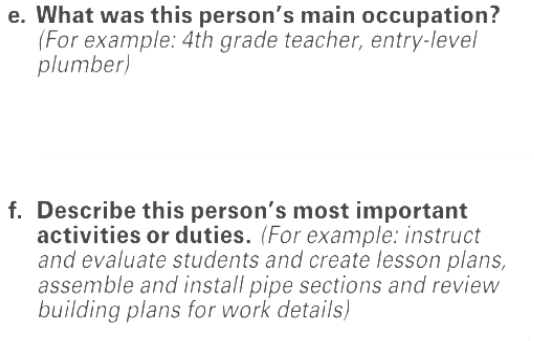
| Data Source | Count |
|---|---|
| 2019 1-Year American Community Survey, Person Level | 4,718,000 |
| Tax Year 2018 E-Filed IRS Form 1040 Returns | 128,300,000 |

### *2.1 The American Community Survey Write-ins*

The American Community Survey (ACS) is a nationally representative survey run by the U.S. Census Bureau. The detailed socioeconomic data, including occupation data, that was previously collected in the decennial census long-form questionnaire is now collected by the ACS, and the ACS is now the largest household survey source of occupation data. We begin our analysis with the 2019 1-Year ACS. The 2019 ACS consists of a sample of 4,718,000 individuals (see Table 1; all counts here and below are rounded for disclosure protection).

Occupation data in the ACS is captured using several write-in fields that consist of verbatim answers to questions about the occupation and industry of each resident over the age of 15 in the household. For this study, we use the written response to the question "What was this [reference] person's main occupation?"[3] (Figure 1) and a write-in capturing information on industry of employment which is used in the ACS occupation coding operation to determine the best occupation code.

**FIGURE 1. 2019 ACS Occupation Write-in Prompts**




The Industry and Occupation Autocoder ("autocoder"), a set of logistic regression models, dictionaries, and edit procedures, assigns an occupation code to around 40% of the ACS responses (Beckhusen (2020)). Records that are not assigned an occupation code or the assigned code's score falls below a certain quality threshold are sent to clerical coders for manual occupational coding. Following the occupational coding procedure, records go through an editing process to address missing data and check for logical consistency between industry, occupation, education, and other variables.[4] We keep edited data if the edits are all based on actual responses but drop the case of an imputed occupation or imputed data used in an edit.

Some example write-in responses are shown in Table 2. Occupations for all household members are potentially reported by the single individual completing the ACS, typically the householder or head of household. Most ACS respondents fill out the questionnaire online or using a paper form; in the event of nonresponse, some respondents are interviewed via telephone or personal visit.

Valid ACS occupation write-ins are those that received a valid occupation code after the auto-coding, human coding, and edit process but pre-imputation. See Beckhusen (2020) for details on the occupation coding process.

## *2.2 Occupation Data from Form 1040*

We use occupation data from the Tax Year 2018 (Processing Year 2019) Form 1040 E-Files. This sample consists of 128,300,000 tax returns (Table 1). The 1040 "Occupation" field is an optional field that allows space for a write-in response for both a primary filer and spouse (if applicable). Using these responses, we create a person-level dataset of individual reported 1040 occupation. This occupation is reported by the filer (or tax preparer).

[3] Starting in 2019, the questionnaire includes the following examples: "4th grade teacher, entry-level plumber", which may prompt respondents to provide more detail than in previous years (Martin-Caughey (2023)).

[4] Education and earnings are explicitly considered in the ACS edit stage but are not used in the LLM-based coder in Section 4 below.

**TABLE 2. Example Occupation Write-ins, ACS**

| Occupation |
|---|
| VICE PRESIDENT PAYROLL |
| SENIOR PRODUCT ENGINEER |
| TEACHER OF HANDICAPPED |
| CARE GIVING |
| FUNERAL DIRECTOR |
| ASSIST REAL ESTATE AGENT |
| SCAFFOLDING SUPERVISOR |
| MACHIN OPERATOR [sic] |
| TRASH COLLECTOR |

There are no consequences to the entry in the "Occupation" field of Form 1040, and the IRS only uses these responses for research purposes. Tax preparation software potentially plays an important role in the occupation reports of e-Filers, as it may both provide guidance and potentially backfill previous year responses. The instructions from one tax preparation software (TaxAct) reads: "Enter what best reflects your current occupation. Common entries include: Student, Laborer, Factory Work, Owner-Operator, Self Employed, Homemaker, Unemployed, Retired, etc."[5]

Industry data from tax records is pulled by linking a worker's W-2 to an Employer Identification Number, and then that EIN is linked to the Longitudinal Business Database to get the NAICS Code (2017 coding scheme). In case of multiple W-2s, the highest earning W-2 is used.

### *2.3 Combining the ACS and IRS Form 1040*

Both the ACS write-ins and IRS Form 1040 are linkable at the individual level by the Protected Identification Key (PIK). The Census Bureau uses the Person Identification Validation System (PVS; see Wagner and Layne (2014)) to assign each person record a unique PIK to facilitate record linkage. Form 1040 collects Social Security Numbers (SSN), and this information is used in the PVS to assign a PIK. The ACS does not collect SSNs, and therefore relies on probabilistic matching to the PVS reference file using name, date of birth, address, household composition, and other data fields. We expect some mismatch due to the higher quality IRS PIKs matching using SSNs to the probabilistic ACS PIKs. Additionally, not all survey records can be assigned a PIK. This may happen when the record has insufficient identifying information, or the person is missing from the PVS reference file. In general, about 90 to 93% of survey records are assigned a PIK and about 98% of federal administrative records are assigned a PIK (Mulrow et al. (2020)).

Our final sample is created by considering individuals who

- have a valid PIK, and
- have a valid ACS write-in, and
- have a valid IRS write-in

using our validity rules discussed above. After merging data sets, we have a paired data set of 1,588,000 individuals with variables for PIK, ACS write-in, IRS write-in, and additional demographic and income information pulled from the ACS responses.

### *2.4 Weighting*

The population of 1040 e-Filers who have an occupation write-in could potentially and substantially differ from the overall U.S. population; there is nonrandom selection into a PIK match (Bond et al. (2014)), selection into filing a tax return at all given filing cutoff rules for household structure and income, selection between filers and e-filers (Kopczuk and

[5] https://www.taxact.com/support/1665/2023/occupation

Pop-Eleches (2007)), and there may be differential nonresponse to the occupation field even conditional on e-filing. For example, if "Lawyers" were 10 times more likely to file a tax return with an occupation field than "Cashiers", the model would fit lawyers more aggressively relative to cashiers when considering overall prediction rates for the underlying population.

For the subset of matched IRS/ACS individuals, we estimate

$$\hat{p}(X) = Pr(\text{any observed IRS write-in} | X, \text{any observed ACS write-in}),$$

that is, the probability of observing a worker in the ACS with an occupation write-in who has a nonempty IRS occupation write-in field, as a function of observed characteristics X.

We can then form inverse probability weights for the matched sample using the ACS base weights multiplied by $1/\hat{p}$. For example, those in occupations less likely to file a tax return would be estimated as less likely to have an IRS write-in given an ACS write-in, and thus upweighted relative to their base ACS weight in later analyses. We estimate $\hat{p}$ as a nonparametric function of age, sex, years of schooling, state, work status the previous week, years since last employment (if not working), and number of weeks worked last year using the LightGBM algorithm of Ke et al. (2017).

Summary statistics from reweighting are shown in Table 3. Our prior is that non-prime age workers and those with less schooling are less likely to file and less likely to e-file (given e-filing costs), and so would become upweighted relative to their ACS base weights. We find that the estimated weights vary from the base ACS weights: the IRS reweights have a standard deviation of 0.78 (after mean renormalized to 1). Rows 2 and 3 of Table 3 show that there is a negative correlation between both age and the new weight and years of schooling and the new weight: both older and more educated workers are estimated to be more likely to have an IRS response given a valid ACS response than younger and less educated workers, and so get weighted down.

**TABLE 3. IRS Reweights Summary Statistics**

| Statistic | Value |
|---|---|
| Standard Deviation | 0.78 |
| Corr(Age, IRS Weight) | -0.08 |
| Corr(Years of School, IRS Weight) | -0.36 |

The weighting assumptions assume ignorable nonresponse given observables, which is not true in the context of self-reported survey income (Bollinger et al. (2019)) and certainly not strictly true here, but any induced biases are left to future research.

### *2.5 Potential Reasons for ACS/IRS Mismatch*

There are a variety of possible reasons for different write-ins or different assigned occupation codes between the IRS and ACS matched data. Here we briefly discuss seven reasons we might expect a mismatch.

**Text field differences:** The IRS provides less space to fill in details about an occupation. The ACS collects information used to code a person's occupation in two write-in fields: the person's main occupation and their most important activities or duties in that occupation. Each write-in field and the combination of these write-in fields provide the respondents with more space in which to respond, generating the possibility that the ACS collects more detailed information than the single IRS occupation response space.

Table 4 shows the average number of characters per valid entry in the ACS "Occupation" field, the ACS "Duties" field, and the IRS "Occupation" field. The ACS has two informational advantages over the IRS: first, the occupation field itself provides slightly more characters on average than the IRS field, 14.2 in the ACS vs. 11.6 in the IRS. Second, the "Duties" field allows for the possibility of significant clarification/specification of more detailed characteristics of the occupation, with an average of 28.1 more characters written. From this, we would expect it to be easier to code the ACS responses to one of the 565 Census 2018 occupational codes than the IRS responses.

**TABLE 4. Characters in IRS and ACS Write-in Responses**

| Question | Mean Character Count |
|---|---|
| ACS Occupation | 14.2 |
| ACS Duties | 28.1 |
| IRS Occupation | 11.6 |

**Tax software backfills:** A worker may have changed occupations and reported it correctly in the ACS but failed to update on Form 1040. This type of error is more likely to occur for individuals who use tax preparation software that auto-fills the occupation from the previous year. We do not have information on the mode of tax form completion.

**Time frame mismatch:** An individual may have changed occupations between filling out the ACS and Form 1040. Here, both write-ins are "correct", but they refer to different time periods. We know the dates the ACS and IRS responses were processed but not necessarily the dates the forms were completed.

**Additional demographic information:** The ACS undergoes an editing process to impute missing data and to check for logical consistency between related variables. The editing procedures may assign a different occupation code to a record based on additional information provided in that record such as firm type, industry, and educational attainment. The IRS does not edit the occupation field nor make use of the information provided in additional data fields.

**Industry mismatch:** Differences between LBD NAICS codes and ACS industry codes are possible, particularly for those with multiple jobs or those in multi-establishment firms where we do not know the specific job/establishment to use to assign a NAICS code. The only industry code available at the universal level is the LBD NAICS codes. Industry information is used to assign respondents to occupations for both the ACS and in our proposed method for the IRS, but if our method cannot necessarily use the "correct" industry codes used in ACS coding it will make matching the ACS assigned occupation code more difficult.

**Multiple job holders:** While the ACS specifies that the respondent report the occupation for the job held in the "last week" or the job at which they worked the most hours last week in the case of multiple jobs, the IRS field simply reads "Occupation."

**Instruction violations:** Some respondents may not follow directions, intentionally or unintentionally, on the ACS, Form 1040, or both. Although respondents are required by law to respond to the ACS and report accurate information to the IRS, in practice, there is little risk to providing inaccurate information in these fields.

**Mis-PIKs:** For a small percentage of cases, the ACS and IRS data refer to different people due to probabilistic matching of ACS workers to PIKS and potentially SSNs entered on Form 1040.

## 3. Token Similarities in ACS vs IRS Write-ins

To compare ACS and IRS write-ins, which are both unstructured text strings, we first use an approximate string-matching algorithm called the Token Set Ratio (TSR). The algorithm calculates the similarity ratio by comparing the number of matching characters and the total number of characters in each string.[6] The measure ranges from 0 and 100, with 100 indicating a perfect match. For example, "machine operator" and "michine operator" would have a TSR of 97. TSR is effective at capturing exact matches and matches that contain minor spelling errors, but it cannot be used to match synonyms and strings with varying levels of detail.[7] Therefore, it is simply a first step in understanding how often respondents are providing nearly identical responses. Our next steps will involve methods to semantically score matches.

Table 5 shows the distribution of Token Set Ratio matches across the paired dataset. If all entries were identical, this distribution would have 100% of the mass at 100. The first two columns show that approximately 33% of the sample gets

[6] TSR is a variation of Levenshtein distance, which calculates the minimum number of character insertions, deletions, or substitutions needed to transform one string into another. TSR is insensitive to word order, and it removes duplicate words within strings before calculating the ratio. For example, "history professor" and "professor history" would be considered identical.

[7] Semantically equivalent terms like "lawyer" and "attorney" may fail to register as matches.

an exact match by TSR, which includes both exact matches and matches where one of the answers is a subset of the other. The median TSR agreement is 57.

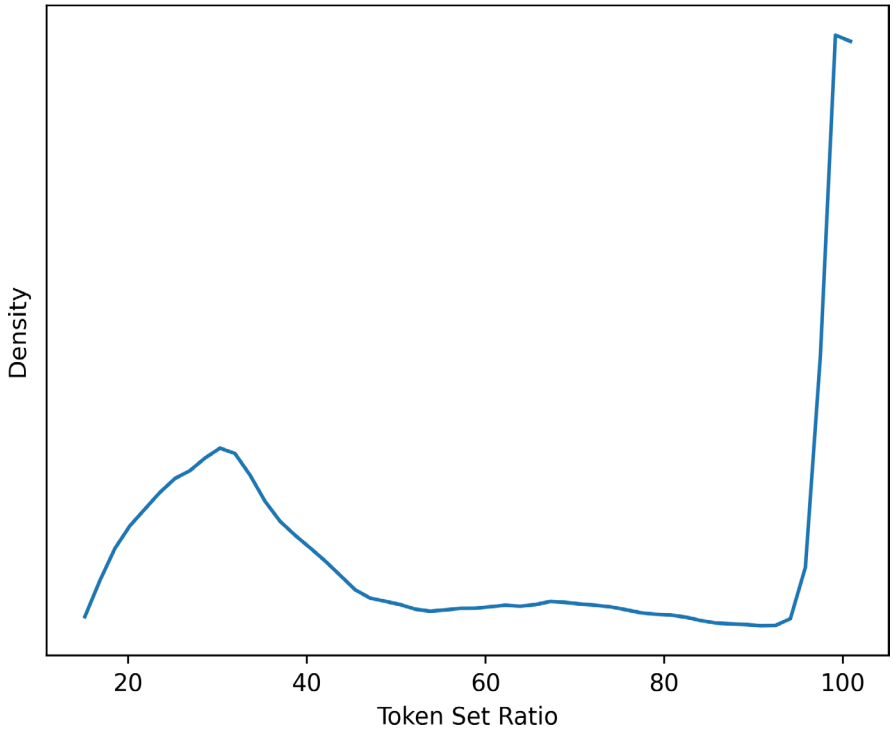
Without a frame of reference, it is hard to interpret TSR comparison numbers, so the third column shows the results from a simulated "bad" match. To generate these numbers, we uniformly shuffled the IRS write-ins across people and re-computed these "wrong person" TSR comparisons. Almost the entire sample has a score of 48 or below in this case, so we interpret 48 as a safe bound to consider two TSRs a real match.

Using the cutoff of 48 for a real match, we find that 75% of the sample would be considered matched. A very conservative cutoff for a true match of a TSR of 57 would put half the sample as correctly matched. Approximately 25% of the paired responses have TSR scores 30 or below, which would put them near the median or below in the randomized match and can confidently be considered nonmatches.

**TABLE 5. TSR percentiiles**

| Pseudo percentiiles | TSR | Random TSR |
|---|---|---|
| 5 | 16 | 12 |
| 25 | 30 | 23 |
| 50 | 57 | 29 |
| 63 | 95 | 32 |
| 67 | 100 | 33 |
| 75 | 100 | 36 |
| 90 | 100 | 43 |
| 99 | 100 | 48 |

**FIGURE 2. Actual Token Set Ratio Distribution**



The kernel density estimate, a visual way to consider the distribution of TSR scores across the sample, is shown in Figure 2. There is a spike around 30, which we interpret as "misses", then uniform levels between approximately 50 and 90, and a large spike near 100.

Occupations themselves may be harder or easier to describe, or have more common ways to describe them; consider "Registered Nurse" as a job versus a contractor who workers on roofs, fences, house interiors, plumbing, etc. Table 6 shows TSR statistics for the 10 largest occupation codes represented in the paired data. "Elementary and middle school teachers" have a TSR agreement of 84.45 on average, which is the highest average score in this group. On the other hand, the average score for "Janitors and building cleaners" is 53.11, which would be considered a marginal match by our criteria. Low scores may reflect high levels of within-occupation job title heterogeneity, which have been found in the ACS and other surveys (Martin-Caughey (2021) and Martin-Caughey (2023)).

There are significant concerns with the IRS data that individuals using tax preparation software may have their form pre-filled with the previous year's occupation, and never take the time to update the field as it is not required or checked. Tax preparers or filers may provide less detailed information in the occupation field than respondents may provide in the ACS write-in fields used to classify a person's occupation. Additionally, it may be the case that younger workers are in jobs that have common alternate ways of describing them. Older individuals may both be more diligent in updating their tax forms and be in jobs that have more succinct and accurate descriptions. On the other hand, retirees who still work may have "retired" on their tax form but report their actual job on the ACS. Table 7 shows there is an inverted-U shape relationship between age and TSR. Workers below 18 are effectively not a match on average; this category presumably over-represents individuals who are mismatched by PIK, since there will not be many individuals under age 18 who show up as filers or spouses on a tax return.

**TABLE 6. Token Set Ratio by Occupation**

| Occ. Name | Occ. Code | Mean TSR | SD TSR | Count |
|---|---|---|---|---|
| Managers, all other | 0440 | 55.52 | 32.14 | 49,500 |
| Registered nurses | 3255 | 68.91 | 37.10 | 40,500 |
| Elementary and middle school teachers | 2310 | 84.45 | 27.77 | 38,000 |
| Driver/sales workers and truck drivers | 9130 | 68.25 | 33.91 | 35,000 |
| First-line supervisors of retail sales workers | 4700 | 59.10 | 32.96 | 30,500 |
| Secretaries and administrative assistants | 5740 | 59.50 | 32.64 | 28,500 |
| Retail salespersons | 4760 | 58.56 | 32.62 | 26,500 |
| Customer service representatives | 5240 | 55.32 | 31,56 | 24,000 |
| Accountants and Auditors | 0800 | 69.60 | 33.64 | 20,000 |
| Janitors and building cleaners | 4220 | 53.11 | 32.90 | 19,500 |

**TABLE 7. Token Set Ratio Scores by Age**

| Age Category | Mean Age | Mean TSR | Count |
|---|---|---|---|
| <18 | 15.95 | 50.88 | 10,500 |
| 18-30 | 25.80 | 59.20 | 273,000 |
| 30-45 | 37.85 | 63.08 | 478,000 |
| 45-65 | 55.36 | 62.09 | 689,000 |
| >65 | 70.60 | 59.51 | 126,000 |

Finally, all the demographic and occupational factors above are correlated with each other, and the conditional correlations of observable individual characteristics and the TSR score of their write-in responses can give us a fuller picture

of who matches write-ins and who does not. We run a linear regression of TSR score onto a constant, a dummy for sex, a quadratic on age minus 18, the natural log of the ACS reported wage and salary earnings for the previous 12 months, years of schooling, and two variables that summarize the predicted quality of the Census's PIK match. The regression was run with fixed effects based on groups "occupation code 0-100", "occupation code 101-200", ..., "occupation code 9700-9760" (9760 being the final occupation code).

The results from this regression are shown in Table 8. The results show that, all else equal, men have a marginally higher match than women, with 0.357 TSR points being a near-negligible amount. On the other hand, higher earners have higher write-in match scores, with the predicted difference between someone earning $20,000 per year and $200,000 per year of 5.91 TSR points (2.565×(ln(200,000)–ln(20,000))). Those with more education have closer matches, with one additional year associated with 0.40 TSR points. Even conditional on other demographics, the relationship between age and TSR has an inverted-U shape, with younger people having worse matches, the best[8] matches coming from those around age 52. Finally, "PVS Score" and "Bad PVS Match Cat." are two variables that indicate the probability of a correct PIK match, with PVS Score being a continuous variable with higher values being a more likely correct match,[9] and Bad PVS Match Cat. being a dummy variable that takes 1 if the PIK system had to do multiple attempts to identify the individual, which is associated with the resulting match being less certain. A higher PVS score is associated with a better occupational write-in match, and those not in the best PVS match category have significantly higher level of disagreement between their write-ins, either due to mis-PIKs or selection of difficult-to-PIK people into lower-quality responses in the 1040 occupation field.

**TABLE 8. Regression Coefficients**

| Variable | Coefficient Estimate | Robust S.E. |
|---|---|---|
| Male | 0.357 | 0.060 |
| ln(Wage and Salary Earnings) | 2.565 | 0.020 |
| Years of Schooling | 0.407 | 0.010 |
| Age – 18 | 0.154 | 0.007 |
| (Age – 18) $^2$ | -0.002 | 0.0001 |
| PVS Score | 0.376 | 0.005 |
| Bad PVS Match Category | -4.412 | 0.194 |

Notes: Occupation category fixed effects included. All p-values < .001.

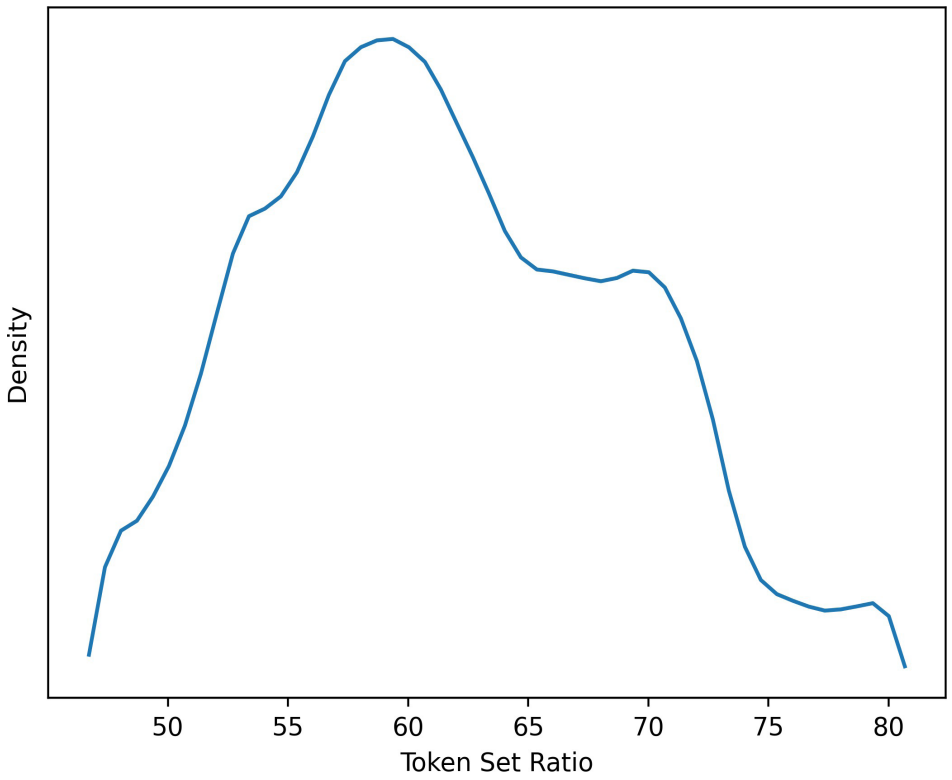
After running this regression, we generated the predicted TSR value for everyone given their demographics we include in the regression. A kernel density plot of the resulting distribution is shown in Figure 3. As expected, the predicted values are more tightly clustered than the actual values, and the multi-modality of the actual TSR distribution is not replicated. What this instead shows is that there is significant variation in who we predict will have a match between the IRS and ACS data. For some groups, e.g. women with low earnings, low schooling, who are far from 52 years old (either above or below), who are not cleanly matched by the Census PVS system, and who are in difficult-to-describe occupations,[10] our predicted TSR is near or below 50 and would be considered a marginal match at best. On the other hand, highly educated high earners who are well matched by the Census PVS system have average predicted scores of 75 or more, which would be categorized as strong matches.

[8] The maximum comes when t satisfies $0.1543 - 2 \times 0.002285 \times t = 0$, and since t is age minus 18 the resulting maximum is equivalent to 51.76 years.

[9] The Census PVS system contains diagnostic information about "how difficult" it was to find a match, e.g. what additional data sources considered to differentiate people with similar names and addresses.

[10] For disclosure protection purposes, we did not report the coefficients for occupation categories. The unconditional TSR levels across large occupations are calculated in Table 6.

**FIGURE 3. Predicted Token Set Ratio Distribution**




## 4. Semantic Similarity via Large Language Models

The case of the ACS write-in of "Lawyer" and an IRS write-in of "Attorney" makes interpreting the results from the previous section difficult to take as dispositive: near-zero token similarity does not imply near-zero chance of being the same occupational report. Additionally, high token similarity does not imply similar occupations: "Paramedic" and "Paralegal" have a token set ratio of 56, putting them near the median of the overall IRS/ACS token matches.

In this section, we use a Large Language Model (LLM) to estimate the semantic similarity between the ACS and IRS write-ins. We estimate the relationship between a given text occupation write-in plus industry and the final coded ACS occupation. We estimate this model twice. First, we use the ACS occupation write-ins and industry code, which gives an estimate of the coding procedure used by the ACS autocoder and clerical coders. Second, we use the IRS write-in and industry code from tax forms—information that would be available in the IRS universe—as the input data to predict the ACS occupation codes for everyone with IRS write-ins in the ACS. If these two models make the same predictions for a given person, we consider those responses semantically similar, but if the IRS write-ins are less informative, mismatched, or just incorrect (see Section 2.5) we can quantify the degree of disagreement in the two sources.

### *4.1 Model*

Mapping from a text string write-in to numerical model inputs is done using the BERT (Bidirectional Encoder Representation from Transformers) model from Devlin et al. (2018), a widely-used open source LLM.[11] BERT can take a free text string as an input, and creates an output of one numerical vector per write-in.[12] These vectors can be used either for generating text as a function of other text, as in LLM chat models, or as numerical inputs into other models, such as here. Due to computational constraints inside the privacy-protected computing environment, we used a BERT model with 2 layers, 2 attention heads, and a 128-dimensional token encoding, known as BERT-Tiny (Turc et al. (2019)).[13] Slightly larger versions of the BERT model were tested without noticeable improvements in prediction quality.

Previous-generation text embedding methods such as convolutional neural nets (with a variety of architectures, see for example Lai et al. (2015), Word2Vec (Mikolov et al. (2013)), GloVe (Pennington, Socher, and Manning (2014)), and FastText (Bojanowski et al. (2017)) were tested and were outperformed by our preferred BERT approach in their accuracy of predicted ACS occupation codes compared to the actual codes.

The model below is estimated twice, once on ACS write-ins and industry and once on IRS write-ins and industry, but both times with the target coded occupations drawn from the ACS sample. This approach allows the IRS write-ins to have their own relationship to the ACS codes that might not hold in the ACS itself; for example, if there are common default write-ins in the IRS data because of instructions that do not appear in the ACS data, our IRS-estimated model would learn that response is uninformative in the IRS even if it might have been informative in the ACS.

Consider a write-in $W$ consisting of a text string along with a coded occupation $j$ and an industry code $k$. The multinomial logit model we estimate can be written

$$\Lambda(j|k,W) = BERT(W;\theta_W)'\gamma_{1,j} + \theta_k'\gamma_{2,j}$$

where $\Lambda$ is the logistic function and $\theta_W$, $\gamma_{1,j}$, $\theta_k$, and are vectors of parameters to be estimated. The first term on the right-hand side contains a function $BERT(W;\theta_W)$ that maps a given text string to a D-dimensional vector, with $D << dim(W)$, using a set of parameters $\theta_W$ that map words and positions in sentences into numbers. The parameter vectors $\theta_k$ are industry-specific $D_K$-vectors that map each industry index to a vector of reals.

Both the *BERT* and the industry "embedding" $\theta_k$ are dimension reduction approaches. In the case of industry dummies, the natural approach would be allowing for a full set of industry dummies in each occupation category, which would be $J{\times}K$ parameters; with this approach, there are $K{\times}D_K{+}J{\times}D_K$ parameters. In the empirical implementation, we have $J=570$, $K=250$, and $D_K=30$, dropping the number of industry-specific parameters from 142,000 to 24,600. The dimension reduction of the BERT function is of course more dramatic, as a 15-character write-in of 27 (the alphabet plus spaces) possible characters in each entry has a total of $27^{15}$ (over 2.9 sextillion) possible configurations. The implemented BERT function we estimate contains a relatively modest 4 million parameters relating the underlying text to the final 128-dimensional vector of real numbers.

The model is estimated with standard machine learning techniques for multinomial logits, and the result of the estimation is a set of probability distributions over occupations generated by a given write-in and industry:

$$\hat{p}(j|k,W) \propto \exp\left(BERT\left(W;\hat{\theta}_w\right)'\hat{\gamma}_{1,j} + \hat{\theta}_k'\hat{\gamma}_{2,j}\right), \forall j \in \{0,1,\ldots,J\}$$

First, to verify the validity of our estimation approach, Table 9 shows the results of estimated models using only ACS data. The first column describes the specification estimated: first, a constant-only multinomial logit, which approximately[14] matches aggregate occupational shares. The second row includes industry dummies by occupation, which approximately

[11] The "Transformers" architecture of the model, which allows models to consider the context of words across long gaps in text, is the base of most modern LLMs, such as GPT/ChatGPT.

[12] We take the standard approach of treating the [CLS] token as representative of the write-in. See Toshniwal et al. (2020) for other options.

[13] https://github.com/google-research/bert

[14] Estimation includes regularization via a dropout layer, tending to slightly push all estimates towards equal probabilities across categories.

matches the occupation shares within industry. The third row matches using the write-in information only, while the fourth row gives the full specification including both a BERT encoded write-in and the industry embedding terms.

The first results column of Table 9 shows the optimized value of the multinomial loss function, which is converted into a McFadden pseudo-$R^2$ measure in the second column. The third through last columns give different match rates that can be used to evaluate model fit: "Match Rate" gives the probability that the most likely predicted occupation from the model is the actual code, and the Top k columns give the probability the actual code is in the top k most likely occupations according to the model.

**TABLE 9. ACS Estimation Results**

| Model | Loss | Pseudo $R^2$ | Match Rate | Top 2 | Top 5 | Top 10 |
|---|---|---|---|---|---|---|
| No regressors | 5.38 | 0.0 | 0.03 | 0.05 | 0.11 | 0.2 |
| Industry dummies | 3.33 | 0.38 | 0.25 | 0.38 | 0.55 | 0.67 |
| LLM Text Only | 1.00 | 0.81 | 0.74 | 0.85 | 0.93 | 0.96 |
| LLM Text + Industry | 0.73 | 0.86 | 0.81 | 0.90 | 0.96 | 0.97 |

Notes: All statistics reported calculated on the validation (non-estimation) holdout sample of 1/8 of the observations. McFadden pseudo-$R^2$ used in the second column. Match Rate is the probability the model prediction is the same as the true Census final code. Top *k* is the probability the Census final code is in the top *k* of most likely model predictions.

We find that in the ACS, the LLM-based autocoder with the full text plus industry specification matches the true code in 81% of cases, and the true code is in the top 5 most likely model predictions in 96% of cases.

The LLM can be compared to both the Census autocoder and Census human coder stages of the ACS coding process. Recall that an occupation/industry joint write-in is only passed to a human coder if the autocoder is unable to make a confident code prediction, so the cases with an assigned autocode should be "easier" in an algorithmic sense than cases coded by humans. Table 10 shows that in cases where the Census autocoder was able to assign a code, our coder predicted the final Census code 97% of the time.[15] In contrast, for write-ins that had to be sent to a human coder, we predicted the exact code in 76% of cases.

**TABLE 10. ACS Autocoder and LLM Comparison**

| Write in Category | Pr(LLM pred. code  Census final code) |
|---|---|
| Census Autocoded | 0.97 |
| Census Handcoded | 0.76 |

Notes: All statistics reported calculated on the validation (non-estimation) holdout sample of 1/8 of the observations. McFadden pseudo-$R^2$ used in the second column.

Table 11 gives the same results where the input variables are drawn from the IRS data and the target is the ACS code. For simplicity, we only estimate the full model and give numbers for the regressor-less model for comparison. We find that the IRS data matches 42% of coded cases, with the actual code being in the top 10 of the model predictions 77% of the time. This is expected given the problems associated with using the IRS data that we discussed in Section 1.5. However, the result is consistent with economically significant information in the IRS write-ins about occupations, even if it cannot be used to generate a full 4-digit Census code with very high confidence. Our next steps include investigating whether different levels of occupational aggregation can lead to high confidence in IRS occupation predictions.

[15] The final Census code can differ from the Census autocoder code due to the later edit process.

**TABLE 11. IRS Estimation Results**

| Model | Loss | Pseudo $R^2$ | Match Rate | Top 2 | Top 5 | Top 10 |
|---|---|---|---|---|---|---|
| No regressors | 5.38 | 0.0 | 0.03 | 0.05 | 0.11 | 0.2 |
| LLM Text + Industry | 2.76 | 0.49 | 0.42 | 0.54 | 0.68 | 0.77 |

Notes: All statistics reported calculated on the validation (non-estimation) holdout sample of 1/8 of the observations. McFadden pseudo-$R^2$ calculated in the second column. The "No regressors" row is the same as the first line of Table 9 by construction.

### *4.2 Semantic Similarity Measures*

Our measure of semantic similarity uses the results from the two models estimated above. Taking the estimated parameters, we compare each worker's induced ACS occupation probability distribution from both the ACS inputs data and the IRS input data. Comparing these probability distributions gives an interpretable comparison of our parameter estimates between the two models without concerns about scale/normalizations.

For an example, say a worker in 2019 wrote "Lawyer" in their ACS write-in field and their firm was human-coded to industry 7270, "Legal Services". These inputs would (hypothetically) lead to an approximately 0.9 probability of being coded to a Lawyer and approximate 0.1 probability of being coded to a Paralegal or Legal Assistant, and a near-zero chance of being any other occupation. If the matched IRS record for this worker read "Attorney" as their write-in and the NAICS code of their firm was "Legal Services", then the probability distribution of the IRS-based model (using only IRS write-in and IRS industry as inputs) should look almost identical to that of the ACS distribution. We would evaluate this as the best possible semantic match between ACS and IRS responses.

The mathematical measure of semantic distance we use is given by

$$SemD(W_{acs}, W_{irs}|k) = \frac{1}{2}\sum_{j=1}^{J}|\hat{p}_{acs}(j|k, W_{acs}) - \hat{p}_{irs}(j|k, W_{irs})|,$$

the Total Variation Distance (TVD) between the estimated probability measures, with  and  being the model evaluated at the ACS and IRS parameters, respectively. The total variation distance measure ranges from 0 (exact same probability distribution), to 1 (the probability distributions both put 1 probability on different outcomes). TVD is a natural way to think about whether two distributions are distinguishable based on empirical frequencies (Ostrovski (2017)).[16]

There are other state-of-the-art methods to compare the semantic similarity across the write-ins: for example, the SentenceTransformers model of Reimers and Gurevych (2019)[17] is a similar approach, explicitly developed to compare the similarity of text strings. The implementation details of SentenceTransformers are quite similar to our approach, where both entries are embedded using BERT and then the embeddings are compared. In testing, SentenceTransformers did not perform well with our large number of occupational categories.

[16] Other distance comparisons between distributions such as Kullback-Leibler divergence are possible as well, although in the case of K-L divergence, infinite distances can be generated by differences between 0 and arbitrarily small predicted probabilities that would be hard to detect in data.

[17] https://sbert.net/

Summary statistics for the Total Variation Distance between ACS and IRS write-ins are shown in Table 12. As above for Token Set Distance, we compare the percentiles of the TVD distribution to percentiiles of a shuffled TVD distribution where we randomly permute the IRS responses while leaving the ACS responses fixed; this is an estimate of what the data would look like if there was no true connection between individual-level responses. We find that about 50% of the paired write-ins have distance below the 1st percentile of the shuffled data, and 90% have paired write-ins that are closer than the 25th percentile of the shuffled data. These direct comparisons show that there are similarities in responses across the ACS and IRS in the semantic space, which helps reinforce the evidence at the token level in the previous section.

**TABLE 12. Total Variation Distance between ACS and IRS Write-ins,percentiles**

| Pseudo percentiles | TVD | TVD Shuffle |
|---|---|---|
| 1 | 0.02 | 0.75 |
| 5 | 0.07 | 0.91 |
| 10 | 0.14 | 0.95 |
| 25 | 0.41 | 0.98 |
| 50 | 0.75 | 0.99 |
| 75 | 0.93 | 1.00 |
| 90 | 0.98 | 1.00 |
| 95 | 0.99 | 1.00 |
| 99 | 1.00 | 1.00 |

The results from semantic similarity are consistent with those from token similarity: there is significant information in the 1040 occupation write-ins that can be used at the tax filer universe-level.

## 5. Conclusion and Future Directions

In this paper, we described and analyzed a new dataset consisting of matched American Community Survey (ACS) and 1040 occupation reports. This dataset allows validation and quality analysis of the IRS's large Form 1040 occupational write-in database by comparing it with the high-quality ACS write-in and coding process. We analyzed the similarity between the two datasets both along the token and semantic dimensions. We found a bimodal distribution of response quality in the token dimension, with over 50% of the ACS sample a high-quality token match with its IRS counterpart, but also a significant set of seeming no-matches.

Alongside the dataset itself, to run the semantic analyses, we created a Large Language Model-based occupational coder that can map occupation write-in responses to ACS occupational codes. This autocoder allows for coding of the entire IRS occupational write-in database, which will allow aggregate comparisons of the responses to what we would expect from the representative ACS occupation numbers.

The natural next step in our research program is to use the LLM-based autocoder to code the entire set of occupations listed on tax returns. This creates a database of standard tax information with over 128 million coded occupational observations within a given year. This is orders of magnitude above the largest existing databases of coded individual worker occupations. Additional improvements could include determining the best level of occupational aggregation to be used in coding IRS occupations, extending the analysis of longitudinal problems due to tax preparation software backfills, and training the LLM model on our full timeframe of available data (2011 to 2020). There are several research projects currently active that will use the U.S. tax filer universe-level coded data sets, looking at previously unanswerable questions about occupations across time, locations, generations, and careers.

## References


Beckhusen, Julia B (2020). "Recent changes in the Census Industry and Occupation classification systems." US Census Bureau, American Community Survey Technical Paper 78.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017). "Enriching word vectors with subword information." Transactions of the Association for Computational Linguistics 5, pp. 135–146.

Bollinger, Christopher R, Barry T Hirsch, Charles M Hokayem, and James P Ziliak (2019). "Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch." Journal of Political Economy 127.5, pp. 2143–2185.

Bond, Brittany, J David Brown, Adela Luque, and Amy O'Hara (2014). "The nature of the bias when studying only linkable person records: Evidence from the American Community Survey." Center for Administrative Records Research and Applications Working Paper 8, pp. 2–30.

Chetty, Raj and Nathaniel Hendren (2018). "The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects." The Quarterly Journal of Economics 133.3, pp. 1107–1162.

Chetty, Raj, Nathaniel Hendren, Patrick Nathaniel, and Emmanuel Saez (2014). "Where is the land of opportunity? The geography of intergenerational mobility in the United States." The Quarterly Journal of Economics 129.4, pp. 1553–1623.

Corak, Miles (2013). "Income inequality, equality of opportunity, and intergenerational mobility." Journal of Economic Perspectives 27.3, pp. 79–102.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pretraining of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805.

Featherman, David L. and Robert M. Hauser (1978). "Chapter 5: Social stratification in a service economy." In: Opportunity and Change. New York: Academic Press, pp. 219–311.

Grusky, David B. and Erin Cumberworth (2010). "A national protocol for measuring intergenerational mobility?" Stanford Center for Study of Poverty and Inequality 2.8.

Ke, Guolin et al. (2017). "LightGBM: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems 30.

Kopczuk, Wojciech and Cristian Pop-Eleches (2007). "Electronic filing, tax preparers and participation in the Earned Income Tax Credit." Journal of Public Economics 91.7-8, pp. 1351–1367.

Lai, Siwei, Liheng Xu, Kang Liu, and Jun Zhao (2015). "Recurrent convolutional neural networks for text classification." In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 29. 1.

Long, Jason and Joseph Ferrie (2013). "Intergenerational occupational mobility in Great Britain and the United States since 1850." American Economic Review 103.4, pp. 1109–37.

Martin-Caughey, Ananda (2021). "What's in an occupation? Investigating within-occupation variation and gender segregation using job titles and task descriptions." American sociological review 86.5, pp. 960–999.

Martin-Caughey, Ananda (2023). "Category Cohesion: Using a similarity index to understand the measurement and meaning of occupations." U.S. Census Bureau Working Paper SEHSD-WP2023-28.

Mazumder, Bhashkar (2005). "Fortunate sons: New estimates of intergenerational mobility in the United States using social security earnings data." Review of Economics and Statistics 87.2, pp. 235–255.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient estimation of word representations in vector space." arXiv 1301.3781.

Minaee, Shervin et al. (2024). "Large language models: A survey." arXiv 2402.06196.

Mitnik, Pablo, Victoria Bryant, and David B. Grusky (2024). "A very uneven playing field: Economic mobility in the United States." American Journal of Sociology 129.4, pp. 1216–1276.

Mitnik, Pablo, Victoria Bryant, David B. Grusky, and Michael Weber (2015). New Estimates of intergenerational economic mobility using administrative data. Tech. rep. Internal Revenue Service, Statistics of Income Division.

Mulrow, E, A Mushtaq, S Pramanik, and A Fontes (2020). "Assessment of the U.S. Census Bureau's Person Identification Validation system." NORC at the University of Chicago.

Ostrovski, Vladimir (2017). Testing equivalent of multinomial distributions. Statistics & Probability Letters 124, pp. 77–28.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global vectors for word representation." In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543.

Reimers, Nils and Iryna Gurevych (2019). "Sentence-BERT: Sentence embeddings using Siamese BERT-Networks." In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.

Toshniwal, Shubham et al. (2020). "A cross-task analysis of text span representations." arXiv 2006.03866.

Turc, Iulia, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "Well-Read students learn better: The impact of student initialization on knowledge distillation." CoRR abs/1908.08962.

Wagner, Deborah and Mary Layne (2014). "Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications' record linkage software." US Census Bureau, Center for Administrative Records Research and Applications Working Paper.

# Disaggregating Tax Compliance Burden: A Comparative Study

*Bizuayehu Bedane (IRS, RAAS)*

## 1. Introduction

The tax compliance burden includes out-of-pocket costs, time, efficiency loss, and psychological costs. Tran-Nam et al. (2000) defines tax compliance cost as the sum of out-of-pocket expenses and the imputed value of time and resources, subtracting any perceived benefits of tax compliance. In contrast, Guyton et al. (2003) defined compliance burden as out-of-pocket, time, efficiency, and psychological costs. However, efficiency and psychological costs present challenges in measurement and are frequently disregarded in compliance cost calculations.

The time cost involves gathering information, record-keeping, understanding tax laws, and preparing and submitting tax returns. On the other hand, the financial cost includes expenses related to tax software, tax preparation services, and printing and mailing tax forms (Guyton et al. (2023)).

Methodologically, tax compliance research faces many challenges, from data availability to non- response bias and monetization of compliance time. Eichfelder and Vaillancourt (2014) identify these challenges, highlighting the intricate process of survey design, data collection, low response rate, and cost measurement. Several studies underscore the significance of mitigating non-response bias and have implemented various techniques to alleviate it (Lignier et al. (2014); Evans et al. (2013); Schoonjans et al. (2011); Brick et al. (2010); Contos et al. (2012); and Smulders et al. (2012)). Despite these challenges, researchers have employed innovative approaches and methodologies to shed light on the structure and composition of tax compliance costs, offering valuable insights into its complexities.

Another critical challenge in tax compliance research is monetizing compliance time. Various methodologies exist for assigning a monetary value to tax compliance time, including using a constant rate based on market wages (Schoonjans et al. (2011)), employing a variable monetization rate (Contos et al. (2012)), or relying on wage rates reported by survey respondents (Smulders et al. (2012); Evans et al. (2016)). However, the choice of monetization technique significantly impacts the calculation of total tax compliance costs, thereby complicating comparisons across different studies.

This study also reviewed two multi-country data sources and studies and compared them with the IRS methodology. The first is the Standard Cost Model (SCM), widely used in the European Union. The SCM is excellent for impact assessment, cross-border comparison, and relevance for all forms of a legislative framework. However, SCM lacks representativeness.

The second data source is the World Bank's Doing Business 'Paying Taxes,' which collects data from 189 countries for small and medium businesses. The World Bank data provides consistency and a substantial volume of expert estimates. However, it doesn't distinguish businesses by size and is unsuitable to examine how tax compliance costs vary across different businesses. It was also criticized for its data irregularities and inconsistencies in some countries.

The Internal Revenue Service (IRS) has collected compliance costs data from individual and business taxpayers since 1984. The Individual Taxpayer Burden (ITB) survey gathers data from Wage and Investment (W&I) and self-employed taxpayers (SE) about the time and out-of-pocket costs incurred in preparing and filing their tax returns. The ITB surveys were conducted in 1984, 1999 (for W&I taxpayers only), 2000 (specifically for SE taxpayers), 2007, and annually since 2010. Business Taxpayer surveys were conducted in 1984, 2004, 2009, and 2012, with plans for subsequent surveys to occur annually or every three years after that. The IRS conducts simulations using the taxpayers' burden model (TBM), ITBM for individual taxpayers, SBBM for small business taxpayers, and BTBM for large business taxpayers' (Guyton et al. (2023)).

Empirical evidence suggests compliance costs vary based on income levels, firm size, preparation methods, technology adoption, information access, and educational background. Blaufus, Eichfelder, and Hundsdoerfer (2014) establishes a positive correlation between taxable income, educational status, and compliance costs for German taxpayers. Furthermore, Berger et al. (2017) confirm that the compliance cost, as a percentage of pretax income, is highest for those in the lowest income quintile.

Self-employed taxpayers face a higher compliance burden, as indicated by Eichfelder and Vaillancourt (2014), whose study suggests that compliance costs are 1% of employees' income but significantly higher for the self-employed. Another study by Blaufus, Eichfelder, and Hundsdoerfer (2011) emphasizes the considerable compliance costs of self-employed taxpayers in Germany.

The expenditure and time invested in tax compliance activities vary depending on the specific task. Research by Guyton *e*t al. (2023) indicates that reporting and substantiating income constitute over half of individual income tax compliance costs. Similarly, studies like DeLuca *e*t al. (2007) emphasize that many of these costs arise from fees paid to professional tax preparers. On the other hand, Eichfelder and Vaillancourt (2014) suggest that time and personal expenses form the predominant share of the cost burden for business taxpayers. This assertion is supported by DeLuca *e*t al. (2003), whose study on small businesses in the U.S. underscores the significant impact of preparation methods on compliance costs, with a substantial amount of time allocated to record-keeping.

For business taxpayers, compliance burden negatively correlates with firm size, with small businesses experiencing disproportionately higher costs (Smulders *e*t al. (2012); Evans et al. (2014); Evans *e*t al. (2016); Slemrod and Venkatesh (2002)).

The complexity of the tax code adds to the rising compliance burden. Benzarti (2020) found that compliance costs influence taxpayers' choice between itemized and standard deductions. Berger et al. (2017) estimate that the complexity of the tax code cost individuals over $104 billion in Tax Year 2017.

The main objective of this paper is to conduct a comparative study of tax compliance costs. The primary approach is to disaggregate tax compliance costs based on different sub-groups, such as firm size and income. Administrative data and published articles are the primary sources for this study, employing a comparative analysis of existing studies and descriptive statistics based on data availability. The main contribution of this study is providing a comprehensive review of challenges in tax compliance cost, a summary of compliance cost studies, and comparisons of U.S. individual and business taxpayers' compliance costs with data from selected countries (U.K., Canada, Australia, and Germany).

This study examined the tax compliance burden, highlighted its conceptual underpinnings and methodological challenges, and compared and contrasted U.S. taxpayers' tax compliance burden with that of the U.K., Australia, Canada, and Germany. The case studies presented in this study offer insight into the structure and composition of tax compliance burdens. These studies from the United States to Germany, Australia to Canada reveal compliance costs' variability and regressive nature, influenced by income levels, business size, and tax code complexity. By comparing these case studies, we gain a nuanced understanding of the various factors influencing compliance costs across different nations and contexts.

The key findings of this research can be summarized as follows. Firstly, tax compliance studies face numerous challenges, including data scarcity, non-response bias, and variability in the valuation of tax compliance time. Consequently, comparisons between tax compliance studies should be approached with caution. Secondly, the study indicates that tax compliance costs exhibit a regressive pattern, with firm size and income negatively correlated with compliance burdens. Thirdly, it is observed that individual taxpayers in the United States shoulder higher tax compliance costs compared to the countries examined in this study (Germany and Canada).

Conversely, compliance costs for small businesses in the United States are lower than those in the United Kingdom.

This paper is organized as follows. Section 2 covers tax compliance concepts, methods, and challenges. Section 3 reviews studies about the structure and composition of tax compliance costs. Section 4 covers case studies that compare U.S. individual and business taxpayers' compliance costs with those of selected countries, and Section 5 concludes.

## 2.1 Tax Compliance: Concepts, Methods, and Challenges

### *2.1 Concepts*

This section examines the complexities and methodologies surrounding tax compliance research. Tax compliance costs are defined differently across various studies. Before delving into the findings of these studies, this section explores the conceptual underpinnings of tax compliance.

Tran-Nam et al. (2000) distinguishes between social and taxpayer compliance costs. Social costs encompass efficiency loss (deadweight loss), administrative expenses, and compliance costs. Tax compliance costs include out-of-pocket expenditures plus the imputed value of time and resources minus the benefits of tax compliance. Administrative costs denote the government's expenses in tax collection.

Contrarily, Guyton et al. (2003) divide the total taxpayer burden into tax liability and excess burden, further breaking down the excess burden into compliance, psychological, and efficiency costs. The classification of excess burden aligns with the concept of social costs of tax compliance in Tran-Nam *e*t al. (2000). Compliance burden comprises out-of-pocket payments and time costs. Psychological costs refer to the dissatisfaction, frustration, and anxiety stemming from interactions with the tax system, which are challenging to quantify. Efficiency loss results from tax-induced distortions, leading to a change in consumer and producer surplus, which are difficult to measure and often omitted from compliance cost assessments.

Generally, tax compliance costs (Guyton *e*t al. (2003); Tran-Nam *e*t al. (2000)) include expenses borne by taxpayers to fulfill their tax obligations, preparing and filing time, and out-of-pocket outlays. Tran-Nam *e*t al. (2000) also argue that taxpayers benefit from tax compliance, such as cash flow and managerial benefits. Managerial benefit denotes enhanced decision-making resulting from experience gained in record-keeping and related tax compliance activities, though challenging to measure and typically excluded from compliance cost calculations.

Tax compliance costs can be further categorized into private and public compliance costs, where the public aligns with administrative compliance expenses. The studies examined in this study primarily concentrate on private tax compliance, encompassing out-of-pocket payments and monetized time expended to fulfill tax obligations.

### *2.2 Challenges in Tax Compliance Research.*

Tax compliance studies face several challenges, including data availability, non-response bias, survey design, defining cost burdens, and valuing compliance time. Eichfelder and Vaillancourt (2014) identify four main issues in survey-based cost measurement methodologies: non-response bias, low response rates from small businesses, potential biases in survey questionnaire framing, the valuation of compliance time, and the allocation of time burdens among internal staff and advisory costs.

Data availability is often limited, with previous studies relying on surveys, qualitative interviews, case studies, and administrative data. Conducting these surveys involves selecting appropriate sampling groups, drawing representative samples, and choosing suitable times and geographic areas (Eichfelder and Vaillancourt (2014)). Moreover, a lack of panel data made comparison over time and across observations impossible in all the studies reviewed.

Low response rates, particularly from small businesses, pose another significant challenge. For instance, response rates in the studies reviewed range from less than 1% (Hansford and Hasseldine (2012)) to 42% (Marcuss et al. (2013)).

Non-response bias and how survey questions are framed can significantly impact tax compliance cost estimates. Eichfelder and Hechtner (2016) studied these effects using Belgian business data, finding that framing temporal aspects of cost measurement (annually versus monthly) could drastically change estimates. For small businesses, estimates could be reduced by as much as 53% or increased by up to 112%, with an average change of 39% downward or 65% upward.

Additionally, Lignier et al. (2014), Evans et al. (2013), Schoonjans et al. (2011), Brick et al. (2010), Contos et al. (2012), and Smulders et al. (2012) highlighted the critical role of addressing non-response bias, which stems from systematic differences between those who respond to surveys and those who do not.

Evans et al. (2013) and Tran-Nam et al. (2014) employed wave analysis to tackle non-response bias. They segmented the survey response data into three waves: early, middle, and late responses. Subsequently, they used a chi-square test for selected questions to compare the responses from the early and late groups across 100 questionnaires. The findings indicate no statistically significant difference between the early and late waves.

Similarly, Eichfelder and Hechtner (2016) observed no substantial evidence of non-response bias affecting compliance cost estimates. Slemrod and Venkatesh (2002) acknowledged the potential for non-response bias due to differences in sampling rates among taxpayer groups. To address this, they calculated a set of weights based on the ratio of the total taxpayer population to the number of responses. Likewise, Blaufus et al. (2014) and Blaufus et al. (2019) employed weighting factors to mitigate potential sample selection biases. Stark and Smulders (2019) noted the risk of sampling bias, given that their study's respondents tended to be high-income earners and well-educated.

To calculate the total compliance cost, we must combine taxpayers' time on various tax-related activities with their direct out-of-pocket expenses. There are several methods to quantify the time spent on these activities: using a constant cost based on the average market wage (Schoonjans et al. (2011)), applying variable monetization rates (Contos et al. (2012)), charging the hourly rates of external service providers as seen in the EU Standard Cost Model (Pedersen et al. (2013)), or using valuations reported by respondents themselves (Smulders et al. (2012), Evans et al. (2016)). The chosen method for monetizing time can significantly affect the total tax compliance costs.

For example, Contos et al. (2012) pointed out that variations in monetizing compliance time can substantially influence the overall compliance cost. Differing from previous approaches, they introduced a variable monetization method, arguing that it better captures the varying opportunity costs of time spent on tax compliance by taxpayers and their employees. In their research, the variable monetization rates ranged from $8 to $90 per hour, whereas the fixed monetization rate was $28.73. Their study found that the average compliance cost for U.S. businesses was $11,600 using variable rate monetization and $10,300 using constant rate monetization, as estimated through the Business Taxpayers Burden Model (BTBM).

Blaufus et al. (2019) also thoroughly examined the obstacles to monetizing compliance time. Given these challenges, the study employed various methods to quantify tax compliance time, including before-tax and after-tax wage rates.

### *2.3 Methods and Issues*

Globally, there are few multi-country data sources, such as the World Bank's Doing Business and the European Union data. These studies generally rely on small samples and may use different data collection methods, including mail, in-person, telephone, or email interviews.

The Standard Cost Model (SCM) is a methodology used across the European Union to assess the compliance costs associated with various taxes. As outlined by the E.U. in Pedersen *e*t al. (2013), the SCM specifically targets the administrative burdens tax compliance places on private businesses. It is intended for microeconomic analysis, aiding in both the ex-ante impact assessments of proposed regulations and the ex-post simplification of existing ones. The SCM defines compliance costs to include all expenses related to adhering to regulations, except for direct financial costs and long-term structural impacts. These costs encompass internal and external labor expenses and any necessary expenditures.

Applicable to businesses of all sizes, the SCM is versatile for impact assessments, including cross-border transactions. It is relevant to all forms of taxes and legislative frameworks, supports segmentation, and facilitates comparisons between countries (Pedersen et al. (2013)). However, as noted by Eichfelder and Vaillancourt (2014), the SCM has its drawbacks, including issues with representativeness, a failure to consider temporary compliance costs, and excluding non-mandatory expenses like those for tax planning.

The World Bank's tax methodology evaluates the ease of tax compliance across 189 economies by analyzing the total tax rate alongside the administrative burden involved. This burden is measured by the hours spent annually on tax preparation, filing, and payment and the number of tax payments required each year. Critical indicators of tax payments assessed include the total number of taxes paid, payment methods, payment and filing frequencies, and the number of agencies involved (World Bank (2018)).

The "time (hours per year)" indicator from the World Bank (2018) quantifies the hours spent annually to prepare, file, and pay three major tax types: corporate income tax, value-added or sales tax, and labor taxes, which include payroll taxes and social security contributions.

Preparation time involves gathering all necessary information to compute taxes due and calculate payments. Filing time consists of completing and submitting all required tax forms to the tax authorities. Payment time accounts for the hours needed to make payments, either online or in person, and includes any delays experienced during in-person payments.

While the World Bank's methodology provides consistency (Pedersen *et* al. (2013)) and a substantial volume of expert estimates (Eichfelder and Vaillancourt (2014)), it also comes with limitations. Firstly, the data does not distinguish between micro, small, medium, and large firms, preventing any inference about how compliance costs might vary across different-sized businesses (D'Andria and Heinemann (2023)). Secondly, in some developing countries, the methodology has faced criticism for producing unrealistically large figures (Eichfelder and Vaillancourt (2014)), and irregularities have been documented (D'Andria and Heinemann (2023)). For instance, an investigation into data irregularities in Doing Business reports for 2018 and 2020 highlighted inconsistencies in countries such as China, Saudi Arabia, and the UAE (Machen *et* al. (2021)). These irregularities were identified based on consultations with tax experts, as the reported compliance costs were not representative sample estimates.

Consequently, there is considerable variance in cost estimates, ranging from the smallest to the largest burden estimates. Since 1984, the IRS has conducted surveys to assess the tax compliance burden of Individual Taxpayers (ITB) and Business Taxpayers (BTB). For Individual Taxpayers, surveys were conducted in 1984, 1999 (for Wage and Investment taxpayers only), 2000 (specifically for self-employed taxpayers), 2007, and annually since 2010. Business Taxpayer surveys were conducted in 1984, 2004, 2009, and 2012, with plans for subsequent surveys to occur annually or every three years after that. The 2004 survey targeted small business taxpayers exclusively, with updates in 2009 introducing a separate survey instrument for large businesses (Guyton *et* al. (2023)).

The ITB surveys categorized tax returns by preparation method[1] and then further stratified within these categories based on five complexities[2] levels. The survey instrument comprises questions regarding the resources, time, and out-of-pocket costs taxpayers incur. The IRS data collected from individual and business taxpayers is representative and employs a robust methodology. One potential challenge with the IRS survey is respondents' inability to differentiate the time used to prepare their federal and state tax returns.

The IRS conducted simulations utilizing the ITBM, SBBM (Contos *et* al. (2009)), and BTBM. The IRS Taxpayer Burden Model (TBM), developed in 2002 and updated in 2010, employs a log-linear model specification. The dependent variable, the logarithm of compliance cost, is estimated as a function of various independent variables. The TBM model controls the type and volume of taxpayer activities necessary to fulfill their tax obligations (Guyton *et* al. (2023)).

Several studies used a log-linear regression model (Blaufus et al. (2011); Contos et al. (2009); Marcuss et al. (2013); Contos et al. (2012); Slemrod and Venkatesh (2002); Blaufus, Hechtner, and Jarzembski (2019)), where the dependent variable is the logarithm of compliance cost as a function of various factors, including firm size, income, and taxpayers' characteristics.

[1] Third-party, self-prepared using tax preparation software, self-prepared by hand, and VITA-prepared.

[2] low, medium-low, medium, medium-high, and high.

All the examined studies rely on cross-sectional data and employ a linear regression model for analysis. However, these approaches fail to capture the change in taxpayers' behavior over time. As Hsiao (2007) and Hsiao (2022) noted, panel data offer numerous advantages by blending inter-individual discrepancies and intra-individual dynamics. Panel data enhances the analytical process by providing increased degrees of freedom, facilitating more precise inferences of model parameters. Moreover, it excels in modeling and capturing complex human behavior compared to single cross-section or time-series data. Panel data simplifies computation and strengthens statistical inference by effectively controlling unobserved individual and time heterogeneity.

These enhancements to analysis and inference in tax compliance cost research suggest that tax authorities and other institutions would benefit from acquiring longitudinal surveys to grasp the dynamic and intricate nature of the structure of tax compliance costs.

In conclusion, understanding tax compliance costs requires careful consideration of these methodological challenges and limitations. These aspects crucially influence the accuracy and usefulness of the findings in policy-making and economic analysis.

## 3. Tax Compliance Cost and Structure: Empirical Evidence

### *3.1 Individual Tax Compliance Costs*

This section provides an overview of research concerning the compliance costs that individual and business taxpayers face in various countries. The total compliance costs include time spent and direct out-of-pocket expenses. As highlighted in section two, these studies should be cautiously approached due to challenges such as time valuation and non-response bias.

Guyton et al. (2003) assessed data from 15,447 U.S. taxpayers, distinguishing between 6,366 wage and investment (W&I) taxpayers and 9,081 self-employed taxpayers. Their analysis revealed that the average compliance cost per taxpayer was $149, with self-employed individuals incurring an average of $363, compared to $75 for W&I taxpayers. The average time spent on tax compliance was 25.5 hours—59.5 hours for the self-employed and 13.8 hours for W&I taxpayers. Marcuss et al. (2013), utilizing data from the Individual Taxpayers Burden (ITB) 2010 survey, found that over half of the compliance costs for U.S. individual income tax were linked to income reporting and substantiation. The study indicated that average compliance costs tended to stabilize with increasing Adjusted Gross Income (AGI), ranging from 0.5 to 2.2% of AGI.

Blaufus, Hechtner, and Jarzembski (2019) analyzed tax compliance costs in Germany using data from 18,196 taxpayers in North Rhine-Westphalia. They found that taxpayers spent between 9.13 and 10.23 hours on tax-related tasks, incurring an average of ($96) 106 euros, with total average compliance costs ranging from 228 ($205) to 321 euros ($289). The study noted that significant time was devoted to collecting and sorting receipts and completing tax forms, highlighting the extensive time commitment required for compliance.

Tran-Nam, Evans, and Lignier (2014) studied compliance costs for Australian personal taxpayers by surveying 517 individuals stratified by income and tax complexity. They reported an average compliance cost of AUD 796.85 ($773), with figures ranging from AUD 370 ($359) for lower-income taxpayers to AUD 3,998 ($3882) for high-income individuals, suggesting a regressive cost structure where the ratio of gross compliance costs over taxable income decreases as taxable income increases.

Further studies highlight the variability in tax compliance burdens based on employment type, income levels, and geographic location. A U.K. study reported an average compliance cost for individual taxpayers of £498 ($329), with an average time expenditure of 4.5 hours. In South Africa, Stark *et* al. (2019) estimated the average compliance cost at ZAR 6,905 ($483). The most recent Canadian study by Vaillancourt and Li (2024) places the average individual tax compliance cost at $130.

**TABLE 1. Individual Taxpayers Compliance Costs from Selected Studies (2003–2024)**

| Study | Country | Group | N | Response Rate | Hours | Average Cost | Year |
|---|---|---|---|---|---|---|---|
| GOST (2003) | USA | W&I | 6,366 | 61% | 14 | $75 | 1999 |
| | USA | SE | 9,081 | 56% | 60 | $363 | 2000 |
| MPA (2010) | UK | All | 320 | 32% | 8 | $329 | 2000 |
| M (2013) | USA | All | 7,685 | 42% | 13 | $373 | 2010 |
| BEH (2014) | Germany | All | 629 | | 10-14 | $218/$329 | 2007 |
| | Germany | EM | | | 7-9 | | 2007 |
| | Germany | SE | | | 21-36 | | 2007 |
| TEL (2014) | Australia | All | 517 | 13% | | $773 | 2011/12 |
| BHJ (2019) | Germany | All | 18,196 | 54% | 9-10 | $205/$289 | 2015 |
| SS (2019) | S. Africa | EM | 556 | | 29.5 | $232 | 2016/17 |
| | S. Africa | SE | 556 | | 29.5 | $1,707 | 2016/17 |
| VL (2024) | Canada | All | 1,523 | | 1.5 | $130 | 2023 |

Notes: G (2003) = Guyton et al. (2003). MPA (2010) = Mathieu, Price, and Antwi (2010). M (2013) = Marcuss et al. (2013). BEH (2014) = Blaufus, Eichfelder, and Hundsdoerfer (2014). TEL (2014) = Tran-Nam, Evans, and Lignier (2014). BHJ (2019) = Blaufus, Hechtner, and Jarzembski (2019). SS (2019) = Stark and Smulders (2019). VL (2024) = Vaillancourt and Li (2024). W&I=Wages and Income, SE=Self Employed, EM= employment income. Exchange rates for reference, 2000: 1 pound=0.661 USD (source: https://data.oecd.org), 2007: 1 USD = 1.37 Euro, 2011: 1 USD=1.03 AUD, 2015: 1 USD=1.1 Euro, 2016: 1 USD=14.3 ZAR.
Source: https://www.imf.org/external/np/fin/ert/GUI/Pages/CountryDataBase.aspx

### *3.2 Business Taxpayers Compliance Costs*

Contos *et* al. (2009) focused on small business flow-through entities in the U.S., finding that C corporations bore an average compliance cost of $8,958, S corporations $8,498, and partnerships $6,717. Notably, partnerships faced a higher relative burden, with compliance costs representing 1.5% of total receipts, compared to 0.77% for C corporations and 0.87% for S corporations. In the U.K., Hansford and Hasseldine (2012) found that small businesses incurred an average tax compliance cost of £21,362 ($13,330), and the median cost per full-time employee decreased from £4,410 ($2,752) to £448 ($280) to £361 ($225) as turnover increased, indicative of a regressive cost structure favoring larger firms.

A study from Malaysia by Sapiei, Abdullah, and Sulaiman (2014) reinforced this pattern, showing that compliance costs constituted 0.057% of sales for small businesses versus only 0.001% for the largest corporations. Schoonjans *et* al. (2011) reported similar findings for 151 Flemish SMEs, with an average compliance cost of 342 euros. Microenterprises (firms with less than 20 employees) had an average total tax compliance cost (TCC) relative to assets of 3.2%. In comparison, small firms with more than twenty employees have an average TCC of only 0.7%, illustrating regressivity.

In Australia, Ligneir, Evans, and Tran-Nam (2014) highlighted the burdensome nature of tax compliance for smaller entities. They reported average annual compliance costs of A$3,392 ($3,293) for micro businesses, A$12,169 ($11,815) for small businesses, and A$54,605 ($53,015) for medium-sized companies, with business size and tax complexity as significant predictors of these costs.

Further extending this analysis, Evans *et* al. (2014) examined small businesses in the U.K., Canada, and South Africa, confirming that tax compliance costs are significant, regressive, and consistent over time. Smulders *et* al. (2012) provided additional evidence from South Africa, where a survey of 5,865 small businesses showed an average annual tax compliance cost of R63,328 ($8,722) and a time burden of 255 hours. Similarly, an Ethiopian study involving 1,003 businesses reported an average compliance cost of $306 per business. The study also found that the tax compliance cost as a share of turnover tends to decrease as business turnover increases (from 4.7 to 5.39 to 5.51%), highlighting the regressivity in smaller businesses (Yesegat *et* al. (2017)).

Moreover, a 2022 E.U. report (Legge *et* al. (2022)) noted that compliance costs for enterprises within the European single market ranged from 1 to 2% of turnover, varying with business size and tax system complexity. The study found that the relative burden of tax compliance (total tax compliance cost to turnover ratio) was the highest for micro sized business (1.9%), followed by small business (0.8%), and medium sized business (0.35%), indicating a regressive trend. Stamatopoulos *et* al. (2017) indicated that large businesses faced compliance costs of $12,710 in Greece. A 2016 study by Evans *et* al. in Australia examined 79 large enterprises and international groups, finding that tax compliance costs, though substantial, were regressive and did not show a decline over time. The study noted that the average compliance cost for large corporations relative to their turnover was 0.04%.

Contos *et* al. (2012) found that compliance costs varied with the size of the business and its organizational structure. Using the BTM methodology and a variable monetization rate, they determined the average compliance cost for large businesses to be $11,600. The result from the robust OLS regression model indicated that the coefficient for high complexity is positive and statistically significant at the 1-percent level, suggesting that activities of higher complexity increased the overall compliance burden. Slemrod and Venkatesh (2002) also examined tax compliance costs across both large and mid-sized businesses, corroborating the regressive nature of these costs. They demonstrated that compliance costs, relative to the firm's size, were disproportionately higher for smaller firms than their larger counterparts.

> "…firms in the $5 million to $10 million asset category spent on average $35,443 on total compliance costs, while firms in the $100 million to $250 million category—firms 10 to 50 times the size of the $5 million to $10 million firms—spent on average $243,942 on total compliance costs—only seven times the average amount spent by the smaller firms." (pp.15)

These studies collectively demonstrate that tax compliance costs impose a significant and regressive burden on smaller businesses relative to their larger counterparts.

**TABLE 2. Compliance Costs of Businesses from Selected Studies (2002–2017)**

| Study | Country | N | Firm Sizes | Cost per Turnover | Cost per Firm | Resp. Rate |
|---|---|---|---|---|---|---|
| SV (2002) | USA | 443 | Large<br>Medium | | $134,954 | |
| CGLN (2009) | USA | 7,049 | Small | | $6,644 | |
| CGLLN (2012) | USA | 22,000 | All | | $11,600 | 31.5% |
| SSFF (2012) | S. Africa | 5,865 | Small | | $8,722 | 6.7% |
| HH (2012) | UK | 41 | Small<br>Medium | | $13,330 | <1% |
| SAS (2014) | Malaysia | 98 | Small<br>Medium<br>Large | Avg=0.01%<br>Small=0.057%<br>Large=0.001% | $14,412 | 20.7% |
| LET (2014) | Australia | 682 | Small<br>Micro<br>Medium | 14% | $10,684 | 7.5% |
| ELT (2016) | Australia | 79 | Large | 0.04% | $1,750,277 | 42% |
| YCC (2017) | Ethiopia | 1,003 | All | 4.7% | $406 | |
| SHE (2017) | Greece | 285 | Large | | $12,710 | 27.9% |

Notes: SV (2002) = Slemrod and Venkatesh (2002). CGLN (2009) = Contos et al. (2009). CGLLN (2012) = Contos et al. (2012). SSFF (2012) = Smulders et al. (2012). HH (2012) = Hansfor and Hasseldine (2012). SAS (2014) = Sapiei, Adbullah, and Sulaiman (2014). LET (2014) = Lingier, Evans, and Tran-Nam (2014). ELT (2016) = Evans, Lignier, and Tran-Nam (2016). YCC (2017) = Yesegat, Coolidge, and Corthay (2017). SHE (2017) = Stamatopoulos, Hadjidema, and Eleftheriou (2017). Annual average exchange rates: 1 USD=7.26 ZAR (2014), 1 pound=0.624 USD (2011), 1 USD=3.27 MYR (2014), 1 USD=1.03 AUD (2011), 1 USD=0.753 Euro (2013).
Sources: https://data.oecd.org and https://www.imf.org/external/np/fin/ert/GUI/Pages/CountryDataBase.aspx

### *3.3 Tax Compliance Cost Structure: Internal, External, and Non-labor Costs.*

This section delves into the structure of compliance costs, considering internal, external, and non-labor expenses as examined by various studies across different nations.

In Australia, large businesses incur an AUD 3 million compliance cost, with internal expenses constituting 45.7%, external costs at 34.2%, and non-labor costs at 20.05% (Evans et al. (2016)). Slemrod and Venkatesh (2002) discovered that for U.S. firms, 58.7% of compliance costs are attributed to internal personnel, 24.8% to external expenses, and 16.5% to non-labor outlays. Stamatopoulos et al. (2017) revealed that in their study, 52.6% of compliance costs for external service providers, 20% for educational expenses, 17% for acquisitions, and 10.2% for internal personnel. The E.U. 2022 report (Legge et al. (2022)) states that most enterprises outsourced their VAT and CIT tax activities. Specifically, 76% of small-sized enterprises outsourced VAT obligations, while 84% of micro-enterprises outsourced CIT obligations (see Table 3 for details).

**TABLE 3. Estimated Share of Tax Compliance Activities**

| | VAT | | | | CIT | | | |
|---|---|---|---|---|---|---|---|---|
| Firm Size | Micro | Small | Medium | LSE | Micro | Small | Medium | LSE |
| Internal | 26% | 24% | 33% | 28% | 16% | 20% | 29% | 24% |
| External | 74% | 76% | 67% | 72% | 84% | 80% | 71% | 76% |

Notes: N=2,479; Source: VVA/KPMG (2021) in D'Andria and Heinemann (2023).

**TABLE 4. Compliance Time Allocated by Activity for U.S. Businesses, Percent (2010–2023)**

| Year | Recordkeeping | Tax Planning | Form Completion and Submission | All Other Time |
|---|---|---|---|---|
| 2010 | 53.1 | 12.5 | 21.9 | 12.5 |
| 2011 | 50.0 | 12.5 | 21.9 | 12.5 |
| 2012 | 56.5 | 13.0 | 26.1 | 4.3 |
| 2013 | 54.2 | 16.7 | 20.8 | 8.3 |
| 2014 | 54.2 | 12.5 | 25.0 | 8.3 |
| 2015 | 54.5 | 18.2 | 22.7 | 9.1 |
| 2016 | 54.5 | 18.2 | 22.7 | 4.5 |
| 2017 | 52.4 | 14.3 | 23.8 | 4.8 |
| 2018 | 52.6 | 15.8 | 26.3 | 5.3 |
| 2019 | 50.0 | 15.0 | 25.0 | 5.0 |
| 2020 | 52.4 | 14.3 | 23.8 | 9.5 |
| 2021 | 54.5 | 18.2 | 22.7 | 9.1 |
| 2022 | 48.0 | 20.0 | 24.0 | 8.0 |
| 2023 | 50.0 | 16.7 | 25.0 | 8.3 |
| Average | 52.6 | 15.6 | 23.7 | 7.8 |

Source: Compiled from 1040 instructions https://www.irs.gov/pub/irs-pdf

Furthermore, research conducted in South Africa by Stark and Smulders (2019) found that individuals allocated 80% of their time to tax compliance activities, with 11% designated for tax adviser fees and 9% for incidental expenses.

Evans et al. (2014) findings suggest that SMEs from the U.K. and Australia spend two-thirds of their time on recording information, while Canadian and South African businesses spend roughly half of their time on this function (see Table 13 in the appendix). The average record-keeping time (2010–2023) allocated by U.S. business taxpayers took half of the total time (see Table 4).

The average form completion and submission time (2010–2023) allocated by U.S. individual taxpayers is 37% followed by record keeping (36%) (see Table 14 in the appendix).

Moreover, Evans *e*t al. (2016) study of large businesses in Australia shows that record keeping and preparation and lodgment relation to taxes was the largest item expenditure, accounting for 37% of the total, followed by tax planning (27%) and professional advice (23%).

### *3.4 Drivers of Compliance Costs*

Several factors contribute to determining tax compliance costs, including income, tax code complexity, and firm size. The tax code's complexity exacerbates the growing compliance burden (Evans et al. (2016); Blaufus et al. (2019); Lazos et al. (2022)). For example, self-employed taxpayers in the U.S. face higher average complexity compared to W&I taxpayers, leading to increased time and monetary costs associated with completing tax forms (Guyton et al. (2003)). Benzarti (2020) found that compliance costs influence taxpayers' decisions between itemized and standard deductions, with itemizing costs ranging from 0.6 to 0.8% of adjusted gross income (AGI). Berger et al. (2017) estimated that the tax code's complexity costs individuals over $104 billion in Tax Year 2017, averaging $596 per taxpayer. Marcuss et al. (2013) analyzed the impact of complexity on the tax compliance burden, using a proxy for activity type and volume, and found that heightened high-complexity activity increases total compliance burden. Specifically, the coefficients for Low, Medium, and High categories are 0.006, 0.01, and 0.039, respectively, indicating that an additional dollar of activity in the high category increases compliance cost by 3.9%, more than in the medium and low category. Blaufus et al. (2019) also affirmed that tax code complexity increases compliance costs.

Evans *e*t al. (2016) identified three drivers of tax compliance costs: the complexity and uncertainty of tax rules, administrative compliance requirements from tax authorities, and international exposure. In Australia, 95% of respondents agreed that the tax law is complex, with complexity scoring the highest, followed by compliance and regulatory demands from tax authorities.

Various studies concluded that business size significantly influences tax compliance costs, with regressive compliance costs; smaller businesses incur higher compliance costs. The regression results from Evans et al. (2016) highlighted that business size strongly predicts tax compliance costs. The coefficients for controlling the effect of size—measured by the logarithm of annual turnover and the number of entities in the group—were positive and less than one, indicating although tax compliance costs increase with business size, the increase is less than proportional. Contos et al. (2012) noted that business size is negatively related to compliance costs. The controls for firm size, measured by the logarithm of total assets and the logarithm of total receipts, were less than one, 0.188 and 0.139, respectively, indicating that compliance costs increase less than proportionally as size increases. Evans et al. (2014) conducted a study across small businesses in Australia, the U.K., Canada, and South Africa, finding that tax compliance costs regress as business size increases. Similarly, the European Union (Legge et al. (2022)) tax compliance SMEs report suggests that compliance costs are negatively related to firm size. The regression results for small, medium, and large size firms, relative to micro-sized firms, were negative and significant. Compared to their turnover, small-sized enterprises spend 1.17 percentage point less than micro-sized enterprises to comply with tax obligations. This figure is 1.82 percentage points for medium-sized enterprises and 2.2 percentage points for larger firms.

Research by Blaufus, Eichfelder, and Hundsdoerfer (2014) established a positive correlation between taxable income and compliance costs for German taxpayers. The regression coefficient for taxable income is smaller than one, indicating economies of scale in tax compliance activities and suggesting that the relative cost burden of tax compliance is higher for

taxpayers with a lower taxable income. Additionally, Berger *et* al. (2017) confirmed that compliance costs, as a percentage of pretax income, are highest for individuals in the lowest income quintile. Average compliance costs as a share of pretax income decrease from 0.8% for the bottom quintile to 0.7% for the second and third quintile, and 0.6% for the fourth quintile, before increasing back to 0.7% for the top quintile. Blaufus, Hechtner, and Jarzembski (2019) indicated that income ranks among the most significant determinants of tax compliance costs.

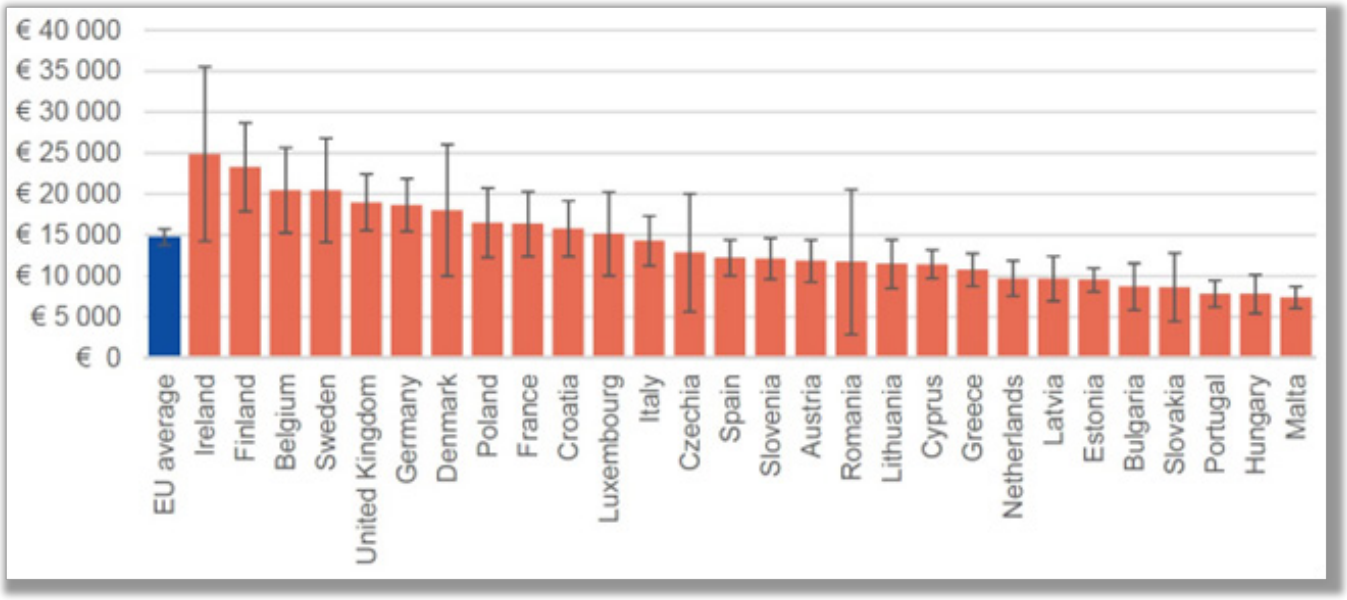
## 4. Comparison of Individual and Business Taxpayers Compliance Cost: Case Study

This section compares the tax compliance costs of U.S. individual and business taxpayers with those of selected countries. Initially, it examines the compliance costs of U.S. businesses using the World Bank's Doing Business data, comparing them to the OECD average and other specific nations. It then contrasts the costs for U.S. individual taxpayers with those from Australia, Germany, and Canada, and similarly for U.S. business taxpayers with two other selected countries, chosen based on data availability. All compliance costs are converted to U.S. dollars using the IMF's average annual exchange rate for the respective country and year.

Table 7 presents data for individual taxpayers across various countries. It shows that U.S. individual taxpayers face higher compliance costs than those in Germany and Canada but lower than those in Australia.

Table 11 compares the compliance costs for U.S. business taxpayers with those in the selected countries. Results indicate that U.S. SMEs incur higher costs than their Australian counterparts but lower than those in the U.K. The 2022 tax compliance study across 28 E.U. countries (including the U.K.) involved 2902 samples, revealing total compliance costs of E.U. 14,745 ($17,449.7) (Figure 1), with an average compliance cost per turnover of 1.9% (See Figure 3 in the Appendix). The compliance cost for E.U. SMEs exceeds that of U.S. businesses with asset sizes ranging from $1 million to $10 million and $10 million to $500 million (see Table 5 for details).

**FIGURE 1. Cross-Country Comparison of Tax Compliance Costs for Businesses in Absolute Values**




Source: VVA/KPMG (2022), based on 2,479 sampled firms. Note: Brackets indicate 95% confidence intervals.

**TABLE 5. Income Tax Compliance Costs from TBM by Asset Size, 2009**

| Total Assets ($ millions) | C Corps | S Corps | Partnerships | All |
|---|---|---|---|---|
| **Panel A: Average Compliance Costs ($)** | | | | |
| $0 to $0.10 | $4,800 | $4,400 | $4,600 | $4,600 |
| $0.10 to $1 | $14,000 | $12,000 | $11,300 | $12,200 |
| $1 to $10 | $34,400 | $27,800 | $23,700 | $26,500 |
| $10 to $500 | $112,400 | $76,300 | $68,600 | $78,300 |
| $500 or more | $630,000 | $331,800 | $316,800 | $468,000 |
| all asset Sizes | $14,900 | $8,900 | $13,400 | $11,600 |
| **Panel B: Total Compliance Costs ($ billions)** | | | | |
| $0 to $0.10 | $4.5 | $11.9 | $6.8 | $23.1 |
| $0.10 to $1 | $7.7 | $13.3 | $10.5 | $31.5 |
| $1 to $10 | $6.1 | $8.2 | $15.3 | $29.6 |
| $10 to $500 | $4.2 | $2.7 | $8.5 | $15.5 |
| $500 or more | $2.8 | $0.1 | $1.4 | $4.3 |
| All asset sizes | $25.3 | $36.3 | $42.5 | $104.1 |

Note: C corporations are entities filing Form 1120; S corporations are entities filing Form 1120S; and partnerships are entities filing Form 1065. Source: Contos et al. (2012).

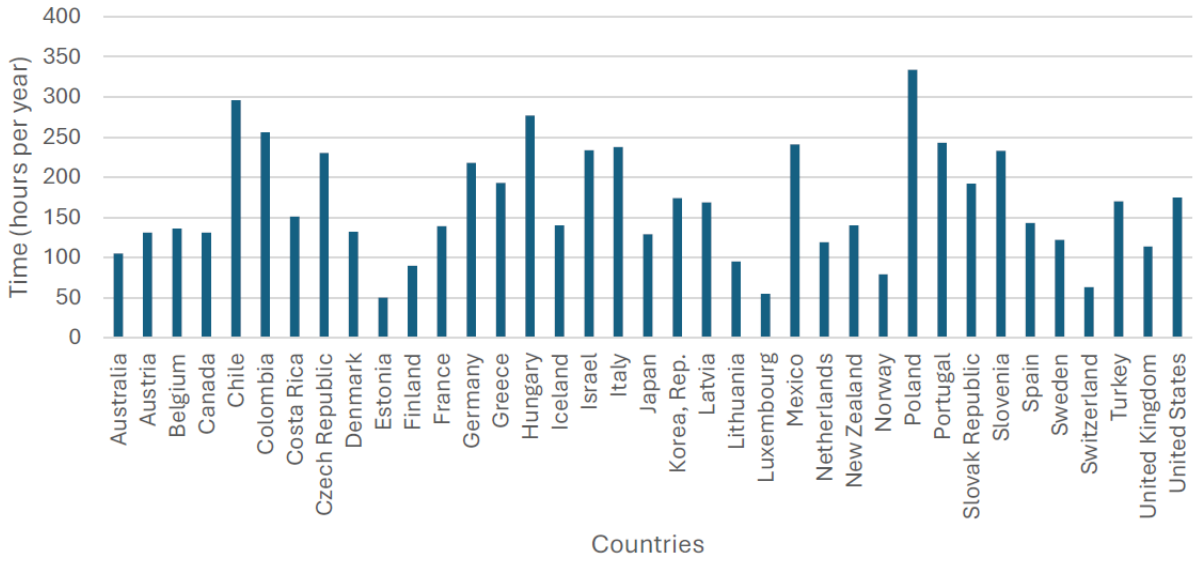
The World Bank's 2018 Doing Business data also serves as a resource to compare the burden of U.S. business compliance globally. This dataset includes a section on tax compliance costs (Paying Taxes), showing that the average medium-sized U.S. firm spends 175 hours on tax compliance, higher than the OECD high-income average of 158.8 hours and more than in the U.K., Australia, Canada, and Japan, yet less than in Germany and Italy (Table 6 and Figure 2).

**TABLE 6. Tax Compliance Time for Selected Countries/Areas**

| Location | Payments (number per year) | Time (hours per year) |
|---|---|---|
| Australia | 11 | 105 |
| Belgium | 11 | 136 |
| Canada | 8 | 131 |
| France | 9 | 139 |
| Germany | 9 | 218 |
| Italy | 14 | 238 |
| Japan | 19 | 129 |
| Netherlands | 9 | 119 |
| United Kingdom | 9 | 114 |
| United States | 11 | 175 |
| East Asia & Pacific | 20.6 | 173.0 |
| Europe & Central Asia | 14.4 | 213.1 |
| Latin America & Caribbean | 28.2 | 317.1 |
| Middle East & North Africa | 16.5 | 202.6 |
| OECD high income | 10.3 | 158.8 |
| South Asia | 26.7 | 273.5 |
| Sub-Saharan Africa | 36.6 | 280.6 |

Source: World Bank (2018).

**FIGURE 2. The Tax Compliance Time of OECD Countries in 2018**




Source: World Bank (2018).

The burden of individual tax compliance in the U.S. is compared with studies from Australia, Canada, and Germany. Marcuss *et* al. (2013) analyzed the 2010 ITB survey, which used a stratified random sampling method, dividing the sample into 15 categories and further stratifying by five complexity categories. Australian data from Tran-Nam *et* al. (2014) utilized a 2011-12 sample of 517 individual taxpayers with a 13.4% response rate. The German study by Blaufus *et* al. (2019) sampled 18,196 respondents in 2019, while the Canadian study by Vaillancourt and Li (2024) used 2023 tax data from 1523 Canadian individual taxpayers. U.S. compliance costs are shown to be lower than Australia's but higher than those of Germany and Canada (Table 7).

**TABLE 7. Compliance Costs for Individual Taxpayers in Selected Countries**

| Country | Year (Data Year) | N | Resp. rate | $ per taxpayer | Cost per tax revenue | Time (hours) |
|---|---|---|---|---|---|---|
| USA | 2013 (2010) | 7,685 | 43% | 373 | - | 12.5 |
| Australia | 2014 (2011/12) | 517 | 13.4% | 774 | 4.84% | 8.3 |
| Canada | 2024 (2023) | 1,523 | N/A | 130 | 1.2% | 1.5 |
| Germany | 2019 (2015) | 18,196 | 0.54% | 96 | - | 10.6 |

Notes: Annual average exchange rates, 2015: 1 USD=1.11 Euro, 2011: 1 USD=1.03 AUD.
Sources: https://www.imf.org/external/np/fin/ert/GUI/Pages/CountryDataBase.aspx and https://data.oecd.org/conversion/exchange-rates.htm

Blaufus *et* al. (2019) focused on North Rhine-Westphalia, Germany's most populous state, and found that individual taxpayers generally need nine to ten hours to prepare their tax returns, with over 75% of this time spent collecting and sorting receipts and filing tax forms. The average total tax compliance cost ranged between €228($205) and €321($289). The study also noted a significant decrease in German income tax compliance costs from 2008 to 2016 and found these costs to represent 2.03 to 2.92% of German income tax revenues for the Tax Year 2015, which is lower than the U.S. (8.3% of tax revenue in 2000) and Australia (7.3% in 2011/12) (Table 8).

**TABLE 8. Comparison of German Individual Taxpayers Compliance Cost (2008 vs 2016)**

| | Self-Employment | | | | | |
|---|---|---|---|---|---|---|
| | No | | | Yes | | |
| | 2008 | 2016 | Sig. | 2008 | 2016 | Sig. |
| Advice | | | | | | |
| No | $N = 375$ | $N = 9,303$ | | $N = 41$ | $N = 2,247$ | |
| Total time (u.b.) | 9.88 | 7.28 | *** | 16.65 | 20.68 | |
| Monetary time (u.b.)[a] | 195.67 | 157.26 | ** | 409.11 | 507.45 | |
| Monetary expenses[b] | 31.12 | 27.99 | | 23.53 | 46.48 | * |
| Total burden (u.b.)[c] | 226.78 | 185.25 | ** | 432.64 | 553.93 | |
| Yes | $N = 129$ | $N = 1,486$ | | $N = 84$ | $N = 800$ | |
| Total time (u.b.) | 6.27 | 9.01 | *** | 43.68 | 31.19 | |
| Monetary time (u.b.)[a] | 137.44 | 202.97 | ** | 1,039.98 | 797.33 | |
| Monetary expenses[b] | 289.72 | 244.33 | | 782.89 | 1,103.96 | ** |
| Total burden (u.b.)[c] | 427.16 | 447.30 | | 1,822.87 | 1,901.28 | |

*Note:* Weighted-adjusted mean values. l.b./u.b. represent the lower and upper bounds, respectively. The values of 2008 are based on the study of Blaufus, Eichfelder, and Hundsdoerfer (2014).
[a]Adjusted for gross wage inflation 2008 to 2016 (18.9 percent; rounded values).
[b]Adjusted for inflation 2008 to 2016 (11.24 percent; rounded values).
[c]Newly calculated inflation-adjusted mean value. The estimates of 2016 are based only on the population up to the age of sixty-five to make the studies comparable (13,836 observations in total). However, the definition of self-employed deviates, as the study of Blaufus, Eichfelder, and Hundsdoerfer (2014) captures the main job, and our study defines *self-employment* as having at least income from self-employment.
*, **, and *** represent significance levels (based on two-tailed tests) of 10 percent, 5 percent, and 1 percent, respectively.

Source: Blaufus, Hechtner, and Jarzembski (2019).

Vaillancourt and Li's (2024) findings indicate that the total compliance cost and the time required for self-employed Canadians ($224) exceed the average ($130) compliance cost, which remains below the 2023 U.S. individual taxpayer compliance cost of $150 (see Table 12 in the Appendix for details). When comparing by income, the compliance cost for U.S. individuals earning over $100,000 ($670) (Table 9) is higher than their Canadian counterparts ($186) across all income levels (Table 10).

**TABLE 9. Individual Compliance Burden by AGI Strata**

| Income | N 1000s | Time (Hours) | Out-of-pocket Costs ($) | Monetized Burden ($) | Burden to AGI (%) |
|---|---|---|---|---|---|
| None | 2,577 | 26.09 | 243 | 441 | -- |
| 1 to 5,000 | 9,961 | 7.30 | 73 | 127 | 83.3 |
| 5,000 to 10,000 | 12,278 | 8.95 | 97 | 164 | 2.2 |
| 10,000 to 15,000 | 12,812 | 10.34 | 114 | 192 | 1.5 |
| 15,000 to 20,000 | 11,742 | 11.24 | 124 | 210 | 1.2 |
| 20,000 to 25,000 | 10,173 | 11.30 | 128 | 222 | 1.0 |
| 25,000 to 30,000 | 8,961 | 11.46 | 136 | 240 | 0.9 |
| 30,000 to 40,000 | 14,620 | 11.74 | 148 | 268 | 0.8 |
| 40,000 to 50,000 | 10,991 | 12.69 | 164 | 315 | 0.7 |
| 50,000 to 75,000 | 18,769 | 13.44 | 192 | 380 | 0.6 |
| 75,000 to 100,000 | 11,828 | 14.09 | 237 | 480 | 0.6 |
| 100,000 to 200,000 | 13,945 | 14.51 | 328 | 670 | 0.5 |
| Over 200,000 | 4,328 | 29.79 | 1,250 | 2,331 | 0.5 |
| Total | 142,985 | 12.54 | 198 | 373 | 6.8 |

Source: Marcuss et al. (2013)

**TABLE 10. Total Compliance Costs of Canadian Individual Taxpayers, 2023**

| Strata | Time (hours) | Monetized Time ($) | Out of pocket Costs ($) | Total Resources ($) |
|---|---|---|---|---|
| Income | | | | |
| Under $60,000 | 1 | 22 | 57 | 79 |
| $60,000 to $80,000 | 2 | 47 | 86 | 133 |
| $80,000 to $100,000 | 2 | 52 | 88 | 140 |
| Over $100,000 | 1 | 59 | 127 | 186 |
| No answer | 1 | 24 | 80 | 104 |
| Employment | | | | |
| Working | 1 | 52 | 82 | 134 |
| Self-employed | 5 | 84 | 139 | 223 |
| Not in labor force | 1 | 27 | 95 | 122 |
| Unemployed | 1 | 18 | 46 | 64 |
| No answer | 0 | 0 | 0 | 0 |
| Education | | | | |
| High school or less | 1 | 25 | 84 | 109 |
| Some College | 1 | 34 | 81 | 114 |
| College degree | 2 | 66 | 100 | 167 |
| No answer | 2 | 82 | 101 | 182 |
| Total | 2 | 42 | 88 | 130 |

Source: Vaillancourt and Li (2024).

A similar comparison for U.S. business taxpayers by Contos et al. (2012) alongside studies from the U.K. and Australia by Lignier et al. (2014), Hansford and Hasseldine (2012), and Evans et al. (2014) covering business taxpayers from Australia, Canada, South Africa, and the U.K., shows U.S. SMEs incur lower compliance costs than the U.K. and Australia. Large businesses in Australia have an average compliance cost of $1.7 million, greater than that of U.S. companies with assets over 500 million (Table 5). However, the compliance cost for small-sized companies in the selected countries is higher than that of U.S. businesses (C-corporations, S-corporations, and partnerships) with less than $10 million in assets (Table 5).

For medium-sized businesses, the compliance cost from the Australian study was AUD 54,605 ($53,014.5)[3] (see Table 15 in the Appendix), which is lower than the compliance cost of U.S. businesses with asset sizes between 10 million and 500 million (Table 5).

Overall, this review suggests that while the compliance cost for U.S. businesses, on average, is lower compared to the countries studied here, the burden for individual U.S. taxpayers is higher than Germany and Canada's average compliance cost.

[3] 1 USD=1.03 AUD (2011 average) https://www.imf.org/external/np/fin/ert/GUI/Pages/CountryDataBase.aspx

**TABLE 11. Compliance Costs of Business Taxpayer's Burden from Selected Countries**

| Country | Year | Data Year | Business Type | N | Resp. Rate | $ per Taxpayer | Cost per Tax Revenue |
|---|---|---|---|---|---|---|---|
| USA | 2012 | 2009 | All | 22,000 | 31.5% | 11,600 | - |
| Australia | 2014 | 2011 | SMEs | 682 | 7.5% | 10,684 | 14% |
| Australia | 2016 | 2011/12 | Large | 79 | 42.0% | 1,750,277 | 0.04% |
| UK | 2012 | 2011 | SMEs | 41 | <1% | 13,351 | - |
| Canada | 2014 | - | Small | 2,449 | 1.4% | 50,286 | - |
| UK | 2014 | - | Small | 4,420 | 0.9% | 36,500 | - |
| Australia | 2014 | - | Small | 3,500 | 4.5% | 34,640 | - |

Notes: 1 USD = 1.03 AUD (2011 average), 1 USD = 1.6 pound (2011)
Source: https://www.imf.org/external/np/fin/ert/GUI/Pages/CountryDataBase.aspx

## 5. Conclusion

The primary purpose of this study is to conduct a comparative analysis of tax compliance costs across countries and over time. The study explores the concepts and challenges in methodology, summarizes the findings of selected studies focusing on the structure and composition of tax compliance costs, and conducts a case study analysis.

The study identifies several challenges in tax compliance research, including data limitations, non-response bias, variations in the method of valuing compliance time, and questionnaire framing. Specifically, several studies acknowledged the presence of non-response bias and employed various mitigation techniques, including wave analysis, and attaching weights. The main message of these challenges is the importance of caution when comparing tax compliance costs from different studies.

Moreover, two multi-country studies, the World Bank Doing Business data and the European Union Standard Cost Model (SCM), are reviewed in this study. The main strength of the SCM is its suitability for impact assessment and cross-border comparison, as well as its relevance to all forms of taxes. Lack of representativeness is the weakness of SCM. However, the World Bank Data offers consistency and tax expert insights. The downside of the World Bank data is its lack of data by business size, which hampers exploring the effect of business size on tax compliance costs. The IRS data collected from individual and business taxpayers is representative and employs a robust methodology.

The empirical studies reviewed show that tax compliance costs are determined by firm size, income, and tax code complexity. The case studies offered valuable insights into the compliance costs faced by individual and business taxpayers across different countries. From the United States to Germany, Australia to Canada, the analysis reveals compliance costs' variability and regressive nature, influenced by income levels, business size, and tax code complexity.

The key findings of this research can be summarized as follows: Firstly, tax compliance studies face numerous challenges, including data scarcity, non-response bias, and variability in the valuation of tax compliance time. Consequently, comparisons between tax compliance studies should be approached with caution. Secondly, the study indicates that tax compliance costs exhibit a regressive pattern, with firm size and income negatively correlated with compliance burdens. Thirdly, it is observed that individual taxpayers in the United States shoulder higher tax compliance costs compared to the countries examined in this study (Germany and Canada). Conversely, compliance costs for small businesses in the United States are lower than those in the United Kingdom.

In all the studies reviewed, capturing the change over time and across observations is impossible due to the lack of panel data. As noted by Hsiao (2007) and Hsaio (2022), using panel data and exploring alternative estimation techniques will enable us to capture the heterogeneity of taxpayers.

## References


Adhikari, B., Alm, J., & Harris, T. F. (2020, May). Information reporting and tax compliance. In *AEA Papers and Proceedings* (Vol. 110, pp. 162–166). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.

Benzarti, Y. (2020). How taxing is tax filing? Using revealed preferences to estimate compliance costs. *American Economic Journal: Economic Policy*, *12*(4), 38–57.

Berger, D., Bryant, V., Guyton, J., & Langetieg, P. (2017). Estimating the effects of tax reform on compliance burdens. *IRS Research Bulletin,* 179–190.

Blaufus, K., Eichfelder, S., & Hundsdoerfer, J. (2011). *The hidden burden of the income tax: Compliance costs of German individuals* (No. 2011/6). Diskussionsbeiträge.

Blaufus, K., Eichfelder, S., & Hundsdoerfer, J. (2014). Income tax compliance costs of working individuals: Empirical evidence from Germany. *Public Finance Review*, *42*(6), 800–829.

Blaufus, K., Hechtner, F., & Jarzembski, J. K. (2019). The income tax compliance costs of private households: Empirical evidence from Germany. *Public Finance Review*, *47*(5), 925–966. Accessed March 27, 2024. https://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=1790929&site=ehost-live

Brick, J. M., Contos, G., Masken, K., & Nord, R. (2010). Response Mode and Bias Analysis in the IRS Individual Taxpayer Burden Survey. *Survey Practice*, *3*(5). https://doi.org/10.29115/SP-2010-0025.

Contos, G., Eftekharzadeh, A., Guyton, J., Erard, B., & Stilmar, S. (2009, December). Econometric simulation of the income tax compliance process for small businesses. In *Proceedings of the 2009 Winter Simulation Conference (WSC)* (pp. 2902–2914). IEEE.

Contos, G., Guyton, J., Langetieg, P., & Nelson, S. (2009). Taxpayer compliance costs for small businesses: Evidence from corporations, partnerships, and sole proprietorships. In *Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association* (Vol. 102, pp. 50–59). National Tax Association.

Contos, G., Guyton, J., Langetieg, P., & Vigil, M. (2010). Individual Taxpayer Compliance Burden: The Role of Assisted Methods in Taxpayer Response to Increasing Complexity. *Recent Research on Tax Administration and Compliance*.

Contos, G., Guyton, J., Langetieg, P., Lerman, A. H., & Nelson, S. (2012). Taxpayer compliance costs for corporations and partnerships: A new look. In *IRS Research Bulletin: Proceedings of the 2012 IRS Research Conference* (pp. 4–18).

D'Andria, D., & Heinemann, M. (2023). Overview on the tax compliance costs faced by European enterprises—with a focus on SMEs.

DeLuca, D., Stilmar, S., Guyton, J., Lee, W. L., & O'Hare, J. (2007). Aggregate estimates of small business taxpayer compliance burden. In *The IRS Research Bulletin—Proceedings of the 2007 IRS Research Conference* (pp. 147–184).

Eichfelder, S., & Hechtner, F. (2018). Tax compliance costs: Cost burden and cost reliability. *Public Finance Review*, *46*(5), 764–792.

Eichfelder, S., & Vaillancourt, F. (2014). Tax compliance costs: A review of cost burdens and cost structures. *Available at SSRN 2535664*.

Evans, C., Hansford, A., Hasseldine, J., Lignier, P., Smulders, S., & Vaillancourt, F. (2014). Small business and tax compliance costs: A cross-country study of managerial benefits and tax concessions. *eJTR*, *12(2)*, 453–482.

Evans, C., Lignier, P., & Tran-Nam, B. (2013). Tax compliance costs for the small and medium enterprise business sector: Recent evidence from Australia. *Tax Administration Research Centre University of EXETER Discussion Paper*, 003–13.

Evans, C., Lignier, P., & Tran-Nam, B. (2016). The tax compliance costs of large corporations: An empirical inquiry and comparative analysis. Canadian Tax Journal., *64*, 751.

Guyton, J. L., O'Hare, J. F., Stavrianos, M. P., & Toder, E. J. (2003). Estimating the compliance cost of the US individual income tax. *National Tax Journal*, *56*(3), 673–688.

Guyton,J., Langetiege, P., Rose, P., Schafer, B., Edelman, S., Garcia, A., Stasko,M. (2023). Taxpayer Compliance Burden. Publication 5743(Rev. 4–2023) Department of Treasury IRS. https://www.irs.gov/pub/irs-pdf/p5743.pdf

Hansford, A., & Hasseldine, J. (2012). Tax compliance costs for small and medium sized enterprises: the case of the UK. *eJTR*, *10*, 288–303.

Hsiao, C. (2007). Panel data analysis—advantages and challenges. *Test*, *16*(1), 1-22. Hsiao, C. (2022). *Analysis of panel data* (No. 64). Cambridge university press.

Lavic, V., & Hadziahmetovic, A. (2020). Corporate income tax burden for SMES—the case of Bosnia and Herzegovina. In *BH Ekonomski forum* (Vol. 13, No. 2, pp. 151–165).

Lazos, G., Pazarskis, M., Karagiorgos, A., & Koutoupis, A. (2022). the Tax Compliance Cost for Businesses and Its Key Determinants: Evidence from Greek Businesses. *Journal of Tax Administration*, *7*(1), 39–56.

Legge, A.D., Ceccanti,D., Foronda, F. H., Németh, K., & Csonka, M. (2022). Tax compliance costs for SMEs: An update and a complement, European Innovation Council and SMEs Executive Agency (EISMEA) Innovation ecosystems, SMP / Entrepreneurship and Consumers. Luxembourg: Publications Office of the European Union. doi:10.2873/180570

Lignier, P., Evans, C., & Tran-Nam, B. (2014). Tangled up in tape: The continuing tax compliance plight of the small and medium enterprise business sector. *Australian Tax Forum.*, *29*, 217–247.

Machen, R. C., Jones, M. T., Varghese, G. P., & Stark, E. L. (2021). Investigation of Data Irregularities in Doing Business 2018 and Doing Business 2020: Investigation Findings and Report to the Board of Executive Directors. Wilmer-Hale. https://thedocs.worldbank.org/en/doc/84a922cc9273b7b120d49ad3b9e9d3f9-00900120 , 21.

Mathieu, L., Waddams Price, C., & Antwi, F. (2010). The distribution of UK personal income tax compliance costs. *Applied Economics*, *42*(3), 351–368.

Pedersen, H. S., Moerup, C., Andersen, C., Findsen, L., Nielsen, J. R., & Lang, T. C. (2013). A review and Evaluation of Methodologies to calculate tax compliance costs. *Taxation Papers, European Commission's Directorate-General for Taxation and Customs Union*.

Sapiei, N. S., Abdullah, M., & Sulaiman, N. A. (2014). Regressivity of the corporate taxpayers' compliance costs. *Procedia-Social and Behavioral Sciences*, *164*, 26–31.

Sapiei, N. S., Kasipillai, J., & Eze, U. C. (2014). Determinants of Tax Compliance Behavior of Corporate Taxpayers in Malaysia. eJournal of Tax Research, 12 (2), 383–409.

Schoonjans, B., Van Cauwenberge, P., Reekmans, C., & Simoens, G. (2011). A survey of tax compliance costs of Flemish SMEs: magnitude and determinants. *Environment and Planning C: Government and Policy*, *29*(4), 605–621.

Slemrod, J. B., & Venkatesh, V. (2002). The income tax compliance cost of large and mid-size businesses. *Ross School of Business Paper*, (914).

Smulders, S., Sitglingh, M., Franzen, R., & Fletcher, L. (2012). Tax compliance costs for the small business sector in South Africa: Establishing a baseline. *eJTR*, *10(2)*, 184–226.

Stamatopoulos, I., Hadjidema, S., & Eleftheriou, K. (2017). Corporate income tax compliance costs and their determinants: Evidence from Greece. In *Advances in Taxation* (pp. 233–270). Emerald Publishing Limited.

Stark, K., & Smulders, S. (2019). Compliance Costs Matter—The Case of South African Individual Taxpayers. *eJTR*, *16*, 801.

Tran-Nam, B., Evans, C., & Lignier, P. (2014). Personal taxpayer compliance costs: Recent evidence from Australia. *Austl. Tax F.*, *29*, 137.

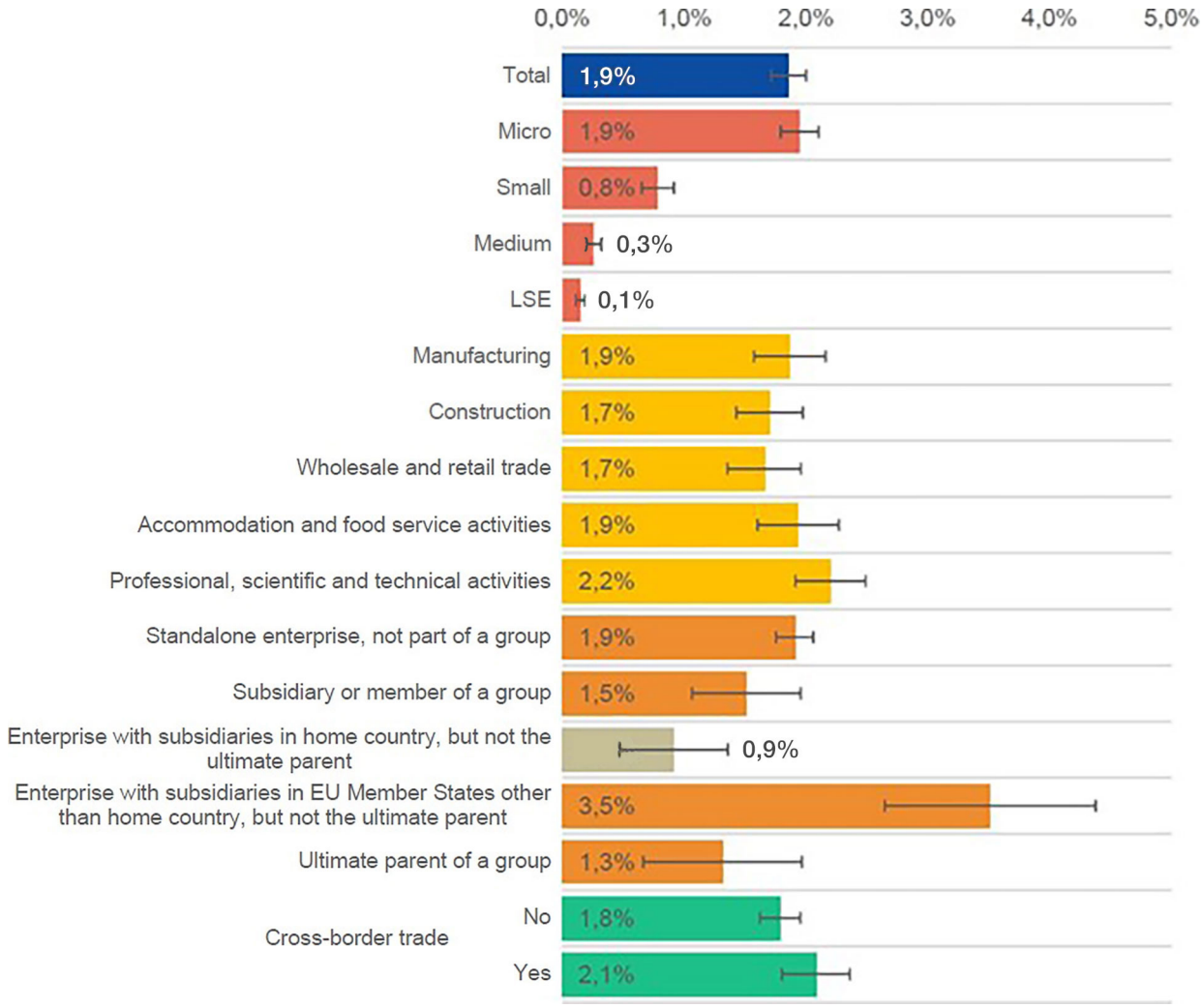
Tran-Nam, B., Evans, C., Walpole, M., & Ritchie, K. (2000). Tax compliance costs: Research methodology and empirical evidence from Australia. *National tax journal*, *53*(2), 229–252.

Vaillancourt, F., & Li, N. (2024). Personal Income Tax Compliance for Canadians: How and at What Cost?

Yesegat, W. A., Coolidge, J., & Corthay, L. O. (2017). Tax compliance costs in developing countries: Evidence from Ethiopia. *eJTR*, *15(1)*, 77–104.

# Appendix

**FIGURE 3. Mean Compliance Costs for Businesses as Percentage of Turnover**




Source: VVA/KPMG (2022), based on 2,479 sampled firms.

**TABLE 12. Compliance Costs for U.S. Individual and Business Taxpayers, 2010 to 2023**

| | Total time | | | Average cost ($) | | |
|---|---|---|---|---|---|---|
| | Non Business | Business | All | Non Business | Business | All |
| 2010 | 12 | 32 | 18 | 160 | 410 | 240 |
| 2011 | 12 | 32 | 18 | 150 | 410 | 230 |
| 2012 | 8 | 23 | 13 | 120 | 420 | 210 |
| 2013 | 7 | 24 | 12 | 120 | 430 | 210 |
| 2014 | 8 | 24 | 13 | 110 | 410 | 200 |
| 2015 | 8 | 22 | 13 | 110 | 410 | 200 |
| 2016 | 9 | 22 | 13 | 120 | 430 | 210 |
| 2017 | 8 | 21 | 12 | 120 | 410 | 210 |
| 2018 | 7 | 19 | 11 | 110 | 400 | 200 |
| 2019 | 7 | 20 | 11 | 130 | 410 | 210 |
| 2020 | 8 | 21 | 12 | 140 | 440 | 230 |
| 2021 | 9 | 22 | 13 | 160 | 470 | 240 |
| 2022 | 8 | 25 | 13 | 140 | 530 | 250 |
| 2023 | 9 | 24 | 13 | 150 | 560 | 270 |

Notes: Details may not add to total time due to rounding. Dollars rounded to the nearest $10. Business filers are those that file one or more of the following: Schedule C, E, or F or Form 2106. You are considered a "non-business" filer if you do not file any of those schedules or forms with Form 1040 or 1040-SR.
Source: Compiled from 1040 instructions https://www.irs.gov/pub/irs-pdf

**TABLE 13. Time Spent on Various Internal Compliance Tasks for all Taxes, Percent**

| Country | Australia | Canada | South Africa | UK |
|---|---|---|---|---|
| Recording information needed for taxes | 66 | 45 | 52 | 66 |
| Calculating, filling forms, and paying taxes | 15 | 21 | 17 | 11 |
| Dealing with the tax office | 1 | 5 | 9 | 4 |
| Tax planning and advice | 4 | 6 | 5 | 4 |
| Dealing with external advisers | 8 | 10 | 8 | 7 |
| Learning about taxes | 5 | 12 | 9 | 8 |
| Other activities | 1 | 1 | 0 | 0 |
| Total | 100 | 100 | 100 | 100 |

Source: Evans et al. (2014)

### TABLE 14. U.S. Individual Taxpayers Time Allocation by Activities (in percent) (2010–2023)

| Year | Recordkeeping | Tax Planning | Form Completion and Submission | All other time |
|---|---|---|---|---|
| 2010 | 41.7 | 16.7 | 33.3 | 16.7 |
| 2011 | 41.7 | 16.7 | 33.3 | 16.7 |
| 2012 | 37.5 | 12.5 | 37.5 | 12.5 |
| 2013 | 42.9 | 14.3 | 42.9 | 14.3 |
| 2014 | 37.5 | 12.5 | 37.5 | 12.5 |
| 2015 | 37.5 | 12.5 | 37.5 | 12.5 |
| 2016 | 33.3 | 11.1 | 33.3 | 11.1 |
| 2017 | 37.5 | 12.5 | 37.5 | 12.5 |
| 2018 | 28.6 | 14.3 | 42.9 | 14.3 |
| 2019 | 28.6 | 14.3 | 42.9 | 14.3 |
| 2020 | 37.5 | 12.5 | 37.5 | 12.5 |
| 2021 | 33.3 | 11.1 | 33.3 | 11.1 |
| 2022 | 37.5 | 12.5 | 37.5 | 12.5 |
| 2023 | 33.3 | 11.1 | 33.3 | 11.1 |
| Average | 36.3 | 13.2 | 37.2 | 13.2 |

Source: Compiled from 1040 instructions https://www.irs.gov/pub/irs-pdf

### TABLE 15. Total Compliance Cost by Business Size (Australia)

| | Micro | Small | Medium | All |
|---|---|---|---|---|
| External costs (adjusted) | 1,049 | 3,871 | 16,300 | 3,425 |
| Value of internal time | 2,343 | 8,298 | 38,305 | 7,579 |
| Total | 3,392 | 12,169 | 54,605 | 11,004 |

Notes: Average calculated based on population weightings for different size categories.
Source: Lignier et al. (2014)

# 2



## Discovering the Art of Avoidance

**Stuntz ◆ Udell**

**Pierson**

**Collins ◆ Wilson ◆ Miller ◆ Payne**
**Roh ◆ Sun ◆ Turk**

# Using a Gravity Model to Predict Cross-Border Tax Avoidance[1*]

*Lori Stuntz and Michael Udell (IRS, RAAS)*

## 1. Introduction

International trade economists use gravity models to explain cross-border flows of goods and services between countries. These models include measures that encourage trade (mass or size) or discourage trade (physical distance). We liken cross-border tax avoidance to a type of cross-border trade and adopt a gravity model to measure the attractiveness of moving various financial flows across borders for tax avoidance.

Our approach recasts the gravity model mass concept as a gradient measure of tax rates between countries. The component measures of this gradient include a withholding tax rate (WHT) on payments between countries and any minimum ownership requirements associated with each WHT plus capital gains tax rates between the source and destination countries. Lower tax rates for this gradient measure increases cross border tax avoidance. We recast the physical distance measure in gravity models as a measure of information transparency across borders as well as an indicator of regulator quality to proxy for the riskiness of having money in a country. Less information transparency and increased regulator quality lead to more attractive cross-border tax avoidance just as less distance between two economies increases trade.

Unlike the classical gravity equation, our model of cross-border tax avoidance is not limited to attraction across a single border. Instead, we develop a framework to measure the attraction across multiple borders as a sequence of border crossings. An important feature of our model is that the order of countries in a sequence matters for tax avoidance.

We create a database of treaty dividend WHT and associated ownership percentages for qualified corporate dividends across 230 countries using treaty information from the International Bureau of Fiscal Documentation (IBFD). For each possible country pair, we record up to 4 different dividend withholding rates and required company ownership percentages, for a total of 59,018 possible bi-lateral cross-border dividend withholding rates.

To generate sequences of countries, we begin with all possible dividend withholding rates and associated ownership requirements for dividends between each of the 230 countries in our database (Country A–Country B) and we join all the dividend treaty WHTs for each of the 59,018 Country Bs with the rest of the world to create over 15 million potential 3-country sequences. We calculate the gravity equation for tax avoidance for each sequence with weights estimated using a measure of foreign financial investment flows.

We develop a framework to chain gravity indexes and evaluate if it is advantageous to add an additional country to a sequence. The gravity model for cross-border tax avoidance is flexible and can be generalized so that sequences may originate from any of the 230 countries in our dataset.

Cross-border tax avoidance is the movement of taxable income from a higher tax rate country to a lower tax rate country for the purpose of reducing tax liability. Tax avoidance is a legal undertaking. Tax treaties can enable cross-border tax avoidance because they reduce WHTs on outbound payments of dividends, interest, and royalty income to encourage trade and promote inbound investment from trading partners.[2] Cross-border income flows that are undertaken to reduce tax liability below domestic tax rates may serve a tax avoidance purpose.[3] Although there is a lack of consensus on the size of cross-border tax avoidance in the literature, there is consensus on its existence. For example, Beer, de Mooij, and Liu (2020) perform a meta-analysis of 37 papers, with data spanning from 1982 through 2012, and find that corporate profits

---

[1] *This paper does not represent any official views or opinions of the Internal Revenue Service, United States Treasury. We thank Danielle Sockin and Karl Nichols for valuable research assistance. Emails: Lori.Stuntz@irs.gov, Michael.A.Udell@irs.gov.

[2] The 12 countries with 0% withholding tax rates on outbound dividend payments to a foreign owner of a U.S. corporation are also some of the largest U.S. trading partners. These countries are Australia, Belgium, Denmark, Finland, Germany, Japan, Mexico, New Zealand, Netherlands, Spain, Sweden, and the United Kingdom.

[3] To be clear, investor motives to undertake a cross-border investment might be to access a foreign market or foreign manufacturing expertise, or foreign resource availability, and when these investments result in outbound flows of dividends, interest, or royalty payments, reduced withholding rates are not prima facie evidence of a tax avoidance motive.

decrease, on average, by 1.59% for each 1 percentage point increase in domestic corporate tax rates.[4] Lejour (2021) finds annual worldwide corporate tax revenue losses due to avoidance range from $123 billion to $180 billion at 2015 levels of corporate profits.[5] For individual income tax avoidance, Johanessen et al. (2023) estimate that US household wealth held in tax haven countries was approximately $2 trillion (2.5% of all US household wealth) in 2018.[6]

We develop a model to identify and rank sequences of countries that could facilitate tax avoidance with respect to financial flows that originate from a source country. Our model is conceptually like gravity models of international trade, which model trade as the result of economic forces of attraction between two or more economies in the numerator and forces that attenuate that attraction, such as transportation costs that increase with physical distance, in the denominator. Unlike traditional gravity equations, our model does not use mass or GDP in the numerator or physical distance in the denominator. Instead, we measure the cross-border attraction of a financial flow with treaty WHTs across each border crossing plus country level domestic tax rates at the source and destination countries. The denominator includes measures of information transparency at each border crossing via participation in exchange of information programs as well as a World Bank index for regulator quality. An attractive country sequence for an investor has both low information transparency and high regulator quality. A sequence with lower regulator quality is riskier to the investor.

Our model identifies country sequences that may be attractive for either cross-border tax avoidance or tax evasion. Because tax evasion is a legal determination, our statistical model does not shed light on whether such activity is legal tax avoidance or illegal tax evasion.

Any number of border crossings can be specified in our model as a sequence. Our primary focus is on the global footprint of tax administration parameters such as tax treaty withholding tax ("WHT") rates and country level information reporting agreements. Cross-border reductions in WHTs may not coincide with reductions in transparency with the source country tax administration and so multiple border crossings may be necessary to both avoid tax liability and tax administration transparency of a source country. For example, one country might have favorable WHTs on cross-border income from a source country but also be highly transparent for tax administration purposes, while another country might have unfavorable WHTs with a source country but also lack transparency with the tax administration of that country. Directing a flow of taxable income from a source country across these two countries in sequence might achieve both low WHTs and low tax administration transparency with respect to the source country. Countries in the "middle" of the border crossings between a source country and a destination country are called conduit countries.[7] Conduit countries, and sequences of conduit countries, can specialize in both reducing cross-border tax liability and cross-border tax transparency. Because multiple border crossings can improve the attractiveness of tax avoidance, the gravity model we propose estimates best paths across multiple border crossings for the purpose of tax avoidance. We avoid the term "tax haven" because it is subjective. A country can be a source, a conduit, or a destination depending upon how it contributes to a country sequence for tax avoidance.

Our gravity model for cross-border tax avoidance ranks the attractiveness of each sequence of countries for the purpose of tax avoidance by calculating an index number for each sequence of countries beginning with a source country, travelling through conduit countries, and ending with a destination country. The greater the value of the index number the more attractive a sequence would be for tax avoidance.

The gravity equation for cross-border tax avoidance is structural in the sense that the variables used—tax rates, information exchange agreements and tax transparency agreements—are policy parameters of each country's tax system. Economic agents with a tax avoidance motive react to these parameters in predictable ways and these observable parameters describe a worldwide system of cross-border toll charges for moving income. Entities with a tax avoidance motive

[4] Beer, S., de Mooij, R. and Liu, Li; (2020). International Corporate Tax Avoidance: A Review of the Channels, Magnitudes, and Blind Spots. Journal of Economic Surveys, Vol. 34, No. 3, pp 660-688. The authors list the main channels of corporate tax avoidance as: mispriced transfer prices, location of intellectual property in low/no tax jurisdictions, treaty shopping, risk-transfer through contracts (for the U.S. this means cost-sharing agreements), avoidance of permanent establishment in high-tax jurisdictions, and locating assets sales in low/no tax jurisdictions.

[5] Arjan Lejour, (2021). *The Role of Conduit Countries and Tax Havens in Corporate Tax Avoidance*. (CentER Discussion Paper; Vol. 2021-014). Center for Economic Research, Tilburg University.

[6] Johannesen, N., Reck, D., Risch, M., Slemrod, J., Guyton, J., and Langetieg, P., (2023). The Offshore World According to FATCA: New Evidence on the Foreign Wealth of U.S. Households. NBER Working Paper 31055.

[7] The size of a country does not distinguish conduit countries. Some conduit countries have small populations and GDP while others are very large with trillion-dollar economies.

might seek to minimize these tolls. An important feature of our model is path dependence across a sequence. A cross-border path through countries A, B, and C may have a different gravity index than a path through countries A, C, and B. As countries alter their tax administration policy parameters, the gravity equation calculation also changes with respect to the attractiveness of each sequence for tax avoidance. Another important feature of the gravity model for cross-border tax avoidance is its reliance on country features and tax parameters that change over time. In the future, the gravity model will be able to identify sequences that become more (or less) attractive for tax avoidance as the underlying variables also change. In this paper, we present a model based upon tax treaties, country tax rates, and information exchange agreements in force during 2017.

This paper proceeds as follows: Section 2 briefly discusses the general use of gravity models for studying international trade. In Section 3, we modify the gravity equation to explain tax avoidance. Section 4 details the various data sources. Section 5 describes our approach to combine countries into sequences of border crossings. In section 6, we present results and gravity model predictions, and then in section 7, we discuss use-cases and applications as well as future expansions for this model.

## 2. Gravity models in the literature

### *2.1 Newton to Tinbergen*

Gravity models are used to explain forces of attraction between bodies. Newton first formulated the gravitational force between two bodies as,

$$F = G\frac{M_1 M_2}{D^2} \tag{1}$$

where F is the gravitational force between two bodies, 1 and 2, measured in a unit of force applied over a certain distance over time. In this specification, force is an attractive force between the masses,  and  of two bodies attenuated by distance between them squared, , plus a constant, , the gravitational constant.

In 1962, Jan Tinbergen adapted this specification to explain the amount of bilateral trade between two countries.[8] The trade equation replaces Netwon's force, F, with the amount of bilateral trade between country A and country B, replaces the mass of the two bodies, $M_1$ and $M_2$ , with the GDP of the two countries, and maintains the same concept of distance, although rather than measured in meters it is measured in miles or kilometers as a proportional relationship:

$$Trade_{AB} \propto \frac{GDP_A^{\alpha} GDP_B^{\beta}}{Distance_{AB}^{\gamma}} \tag{2}$$

Head and Mayer (2014) show that across hundreds of published papers applying the gravity equation to bilateral trade there is a remarkable stability in the parameter estimates of α, β, and γ.[9]

### *2.2 Gravity models in cross-border tax avoidance*

Gravity models have struggled to gain traction for estimating cross-border tax avoidance. An early attempt by John Walker and Brigitte Unger (2009) posited cross-border money laundering as a type of international trade.[10] Unlike most analyses of international trade where cross-border flows of goods are well documented, cross-border money laundering is rarely observed. The inherent inability to observe of cross-border crime has been the fly-in-the-ointment to practical use of the

[8] Tinbergen, J. (1962). An Analysis of World Trade Flows. In "*Shaping the World Economy*". New York, Twentieth Century Fund.

[9] Head, K., and Mayer., T. (2014). Gravity Equations: Workhorse, Toolkit, and Cookbook. In "*Handbook of International Economics,* vol. 4, edited by Gita Gopinath and K. Rogoff. Amsterdam, Elsevier.

[10] Walker, John and Brigitte Unger. (2009). "Measuring Global Money Laundering: 'The Walker Gravity Model'". Review of Law and Economics.

international trade/gravity model framework. The authors do not resolve this issue but instead hypothesize that a gravity model *could be considered* for estimating the amount of cross-border money laundering.

Ferwerda, van Saase, Unger, and Getzner (2020) partially overcome the unobservability of cross-border money laundering by using a special data set of regulatory filings of suspicious transaction reports (STRs) filed by financial institutions – similar to the suspicious activity reports (SARs) that financial institutions in the U.S. are required to file – covering 2009 through 2018 in the Netherlands.[11] These reports identify cross-border flows that have indicia of being suspicious for bank regulatory and money-laundering purposes. Some of the STRs identify cross-border money-laundering, and some do not identify any criminal activity. The STRs report suspicious money flows for both inbound to and outbound from the Netherlands. The authors use country-pairs of cross-border money-laundering with the Netherlands either as the source country or as the destination country as the dependent variable in their gravity equation.[12] They show that traditional international trade gravity model variables such as the GDP of each country and the geographic distance between each country can explain almost half of the variation in amounts reported on the STRs.[13] As with gravity models of international trade, cross-border money laundering increases with GDP for the destination country and decreases with geographic distance between countries. Both papers use a standard gravity equation of international trade and add explanatory variables that might indicate more crime or less crime. We take a different approach. We reformulate the standard gravity equation as a tax avoidance model. In this model, reduced tax withholding rates at border crossings and reduced tax administration transparency with the source country increase cross border financial flows for tax avoidance.

## 3. Constructing a Gravity model for Tax Avoidance

### *3.1 Don't try to observe the unobservable*

For our reformulated gravity equation, the ideal dependent variable would be a measure of tax avoidance across borders, but this is mostly unobservable. Instead, we use a widely reported cross-border measure of inward foreign direct investment (FDI). This has several advantages discussed further below. As explanatory variables in the gravity equation, we include tax rates in each country, treaty WHTs between countries, regulator quality, and measures of transparency with tax administration.

Our model makes three important contributions: 1) the model uses readily available country level measures that can be updated annually; 2) we modify the gravity equation to apply to multiple borders; and 3) we exploit the double counting inherent in global inward FDI measurement to identify the role of conduit countries. Conduit countries are key actors in cross-border tax avoidance. This model allows us to consider sequences of countries with any number of border crossings. It provides distinct measures of attractiveness for each sequence where the order of countries matters, which we refer to as the gravity index. For example, the country sequence A->B->C->D could have a different gravity index than country sequence A->C->B->D.

We use the gravity model indexes to identify best potential conduit or destination countries given a set of countries. A conduit country facilitates financial flows into and out of a country with the least cost, and in some instances, the least transparency for tax administration. A destination has low (or no) taxes and the least tax transparency or information sharing with a source country. For example, we might observe a set of countries associated with a taxpayer on a tax return. We can then use the gravity model to identify the most likely set of additional conduits and destination countries, for the purpose of tax avoidance, that we do not observe on the tax return.

[11] Ferwerda, J., van Saase, A., Unger, B., and Getzner, M., (2020). *Estimating money laundering flows with a gravity model-based simulation*. Scientific Reports, https://doi.org/10.1038/s41598-020-75653-x.

[12] In addition, within country money laundering would also be reported in the STRs. The authors exploit the relationship between "in-country" and "cross-border" money laundering to examine the shares of money laundering that remain domestic and that move across borders.

[13] Other variables in their gravity specification include whether the two countries share a common language (yes), currency, background, religion (yes), are listed as a tax haven (no), or membership in the Egmont group. This last variable is positive and significant for the destination country, which might indicate that it is endogenous with the STR filings.

### *3.3 A gravity equation specification for cross-border tax avoidance*

We modify the Newton-Tinbergen gravity equation by replacing mass or GDP measures in the numerator with measures of attraction between countries including the withholding tax (WHT) rate on payments between countries and ownership requirements associated with each WHT, and capital gains taxes imposed by the source and destination countries. In the denominator, the physical distance term is recast as measures of tax administration transparency across each border with less transparency increasing the attractiveness for cross-border tax avoidance and measures of the regulator quality across the sequence.[14] Unlike the Newton-Tinbergen gravity equations, the gravity equation for cross border tax avoidance is not limited to a single border crossing. We adapt this gravity equation to measure the attraction across multiple borders as a sequence of cross-border withholding rates, investment ownership requirements, country regulator quality, and tax transparency measures.

Our basic gravity equation for cross-border tax avoidance is represented by equation (3),

$$TA = \frac{DIV.OWN.path^{\boldsymbol{\beta_1}}(1 - CG.ratio)^{\boldsymbol{\beta_2}}}{\frac{1}{RQ.path}^{\boldsymbol{\beta_3}}(1 + EOI.ratio)^{\boldsymbol{\beta_4}}} \tag{3}$$

where  is a measure of dividend withholding taxes and ownership requirements across all countries in a sequence,  is the ratio of capital gains tax rates in the destination country over the capital gains tax rate in the source county,  is an index that uses the World Bank regulator quality index across each country in the sequence, and  is a constructed index that ranks changes in tax transparency across countries in a sequence.[15] While we develop this model around cross-border dividend payments, it is easily extendable to cross-border interest, royalty, and other types of income payments as these are all identified in tax treaties. Variable sources are discussed in section 4, and variable construction are discussed in detail in section 5.

If reducing taxes on cross-border income with the least amount of transparency with respect to the source country and with the greatest degree of confidence in the safety of the border crossings are indicia of a tax avoidance motive, then this specification can rank sequences of border crossings that are consistent with that motive.

The coefficient on each variable in the gravity index identifies the contribution of tax system parameters (and the World Bank regulator quality) to cross border FDI that may contribute to tax avoidance. We can estimate these importance weights by taking a log transformation of the gravity equation:

$$\begin{aligned}\log(TA) = \beta_0 + \beta_1 \log(DIV.OWN.path) + \beta_2(1 - CG.ratio) + \beta_3 \log(RQ.path) \\ + \beta_4 log \frac{1}{(1 + EOI.ratio)} + \varepsilon\end{aligned} \tag{4}$$

In a perfect world, we would estimate these weights across all possible country sequences using a measure of observed tax avoidance across each country sequence. But tax avoidance is largely unobservable. Instead, we ask a slightly different question about a related variable, foreign direct investment (FDI), that is widely observed. We ask, "what portion of inbound (to a source country) FDI is consistent with a tax avoidance motive, and which cross-border sequences of outbound income (from a source country) supportive this motive?" How does FDI relate to cross-border tax avoidance? FDI is an inbound flow of investment that gives rise to an outbound flow of dividends, interest, royalties, and capital gains.

[14] Unlike the Newton and Tinbergen gravity models where an increase in physical distance squared attenuates forces of attraction, our model does not attenuate the cross-border forces of attraction by an increase in tax administration transparency squared. Tax administration transparency measures are not physical distance concepts but rules-based concepts whose measure can change as countries enter or exit various information exchange agreements. The impact of our rules-based concept on forces of attraction is an empirical question at this stage of model development.

[15] The World Bank Regulator Quality measure "reflect the perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development." We use these measures as a proxy for the degree of confidence than an investor can have in moving money across a sequence of countries. Greater confidence should enhance the attractiveness of a sequence. See Kaufmann, Daniel, Aart Kraay and Massimo Mastruzzi (2010). "The Worldwide Governance Indicators: Methodology and Analytical Issues". World Bank Policy Research Working Paper No. 5430 (http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1682130).

To the extent that inbound FDI is used to relocate otherwise taxable income of a domestic enterprise outbound for a tax avoidance purpose, the gravity equation will capture that effect.[16]

An unrelated literature on the double counting in FDI statistics unintentionally puts a spotlight on conduit countries. Double counting of FDI occurs when one country receives inward FDI and then sends similar amounts of FDI outward to other countries. This is because FDI data is surveyed at the country level and not on a worldwide consolidated basis. Damgaard et al. (2019) estimate that for 2015 approximately $15 trillion of the nearly $40 trillion in global FDI identified in Coordinated Direct Investment Survey (CDIS) could be double counted.[17] While double counting of FDI poses a significant challenge to the efficacy of FDI statistics—because it confounds real investment with flow-through investment—it helps identify conduit countries. First, double counting of FDI is prima facie evidence of the role of conduit countries. A conduit country, on net, balances inward investment with outward investment.[18] On an annual basis, the net international investment position of investment flows into and out of a conduit country is near zero. Second, the flow of FDI from one country to another sets up an opposite flow in the form of dividends, interest, royalty, and capital gains income. These cross-border income flows are important components of (annual) tax avoidance and ultimately are what we want to measure.

## 4. Data Sources

### *4.1 International Bureau of Fiscal Documentation (IBFD)*

We obtain country level tax and treaty data for 230 countries from the International Bureau of Fiscal Documentation (IBFD)[19] using their historical "Country Tax Guides" and "Country Treaty Tables" for calendar year 2017. Each Country Tax Guide contains data on tax features such as tax rates on income and capital gains for corporations and individuals. We use the capital gains tax rate for individuals in the source country and nonresident individuals in destination countries in the gravity index.

Each country has its own Country Treaty Table on IBFD that shows its treaty negotiated WHTs for dividends, interest, and royalties, as well as the WHT in place for countries when there is no treaty. We record the dividend WHT for each country pair as well as the required minimum ownership percentages from the Country Treaty Tables.[20] For each country pair, we code up to four dividend WHT rates and required ownership percentages that are labeled LOW (the lowest possible treaty dividend WHT), HIGH (the highest treaty dividend WHT), EU (accounts for special rates under an EU parent-subsidiary directive[21]), and DEFAULT (the withholding rate in effect when there is no treaty). Each of the up to 4 WHTs also has an associated ownership percentage. DEFAULT rates have no ownership minimum. For countries with treaty rates, the ownership percentage to obtain the LOW rate is generally higher than the ownership percentage needed to obtain the HIGH rate. For the 230 countries in the IBFD data, this yields 59,143 cross-border country pairs of dividend WHTs.

### *4.2 Worldwide Governance Indicators (WGI)*

The World Bank Worldwide Governance Indicator (WGI)[22] project reports governance indicators for six dimensions of governance including: Voice and Accountability, Political Stability and Absence of Violence / Terrorism, Government

[16] The numerator will capture both the "normal" and "excess" returns of inward FDI back to foreign investors. Whether the return to investors is "normal" or "excess" does not matter for tax avoidance. Instead, what matters is the path that foreign investors use across countries to remove these profits from the source country.

[17] Damgaard, J., Elkjaer, T. and Johannesen, N. (2019*). What is Real and What is Not in the Global FDI Network*?, IMF Working Paper WP/19/274. The Coordinated Direct Investment Survey is performed by the IMF each year. See http://data.imf.org/CDIS.

[18] Conduit countries are market makers for international capital flows.

[19] https://research.ibfd.org. Data obtained via a paid subscription that includes access to historical archive tables.

[20] The required minimum ownership percentages are generally stored in a series of footnotes on each Country Treaty Table. Significant human time was involved in properly coding each of the WHT/ownership pairs.

[21] The EU parent-subsidiary directive exempts dividends paid by subsidiary companies to their parent company from withholding taxes. A parent company is a company from an EU member state with a minimum of 10% ownership in a company from another EU member state. https://taxation-customs.ec.europa.eu/taxation-1/company-taxation/parent-companies-and-their-subsidiaries-european-union_en

[22] https://info.worldbank.org/governance/wgi/

Effectiveness, Regulator Quality, Rule of Law, and Control of Corruption. We use Regulator Quality (RQ) in the gravity index as a proxy measure of the legal stability of moving money across multiple border crossings. Index values for each country range from -2.5 to 2.5 and are normalized so that 0 represents the average. We re-standardize these measures to lie between 0 and 1. The WGI includes data on 209 countries for RQ in 2017. For perspective, 15 countries have an RQ indicator that is better than the United States and 194 have an RQ that is worse than the Unites States.

We impute values for countries with missing RQ via ordinary least squares regression using FATCA and AEOI participation, log (GDP per capita), dummy variables for membership in a variety of multinational treaty agreements, and dummy variables for whether the country is a territory of France, the Netherlands, the UK or the US. The adjusted-$R^2$ of this regression is 0.739. Imputation regression coefficients and results can be found in Appendix Tables A1 and A2. We are unable to impute RQ for Saint Barthelemy or Saint Martin (French) as we do not have GDP data for either of these French Departments. We impute values for Curacao, Gibraltar, Monaco, Guadeloupe, Sint Maarten (Netherlands), San Marino, Bonaire, Isle of Man, Faroe Islands, British Virgin Islands, Guernsey, Turks and Caicos, New Caledonia, Falkland Islands, Northern Mariana Islands, French Polynesia, Cook Islands, Montserrat, and Niue.

### *4.3 Exchange of Information Indicators*

We consider four exchange of information (EOI) variables, each coded as an indicator equal to 1 if the country is a participant with the U.S., and 0 if the country is not a participant. These four indicators are FATCA, EOIR (Exchange of Information upon Request), AEOI (Automatic Exchange of Information) and KYC (Know Your Customer anti-money laundering rules).[23] Automatic exchange of information agreements are in place with respect to specific income items while exchange of information agreements upon request can be more broadly based.

Foreign Account Tax Compliance (FATCA) is an automatic exchange of information agreement that requires foreign financial institutions (FFIs) to report to the IRS information about financial accounts held by U.S. taxpayers, or by foreign entities in which the U.S. taxpayers hold a substantial ownership interest.[24] EOIR denotes countries with which the U.S. has an income tax treaty or other convention or bilateral agreement relating to the exchange of information.[25] AEOI represents countries with which the Treasury Department, the IRS, and another country have determined that the automatic exchange of deposit information is appropriate.[26] Know Your Customer (KYC) countries participate in agreements with the Treasury Department that require FFIs to obtain identity documents from clients.[27]

There is a great deal of variation among countries with the four exchange of information indicators as shown in Figure 1. Each panel is a tabulation of the EOI variable in the legend against each of the other EOI variables. For example, the top left panel depicts a crosstab for FATCA participation with each of the other three EOI variables. All 45 AEOI countries also participate in FATCA, 64 out of 92 EOIR countries have FATCA participation, and 58 KYC countries have FACTA while 14 do not. All 45 countries with AEOI also have FATCA and EOIR participation.

[23] These four EOI variables are specific to information sharing with the Unites States tax administration. The concept is easily adapted to other EOI variables such as the OECD Common Reporting Standard ("CRS") or country-by-country reporting.

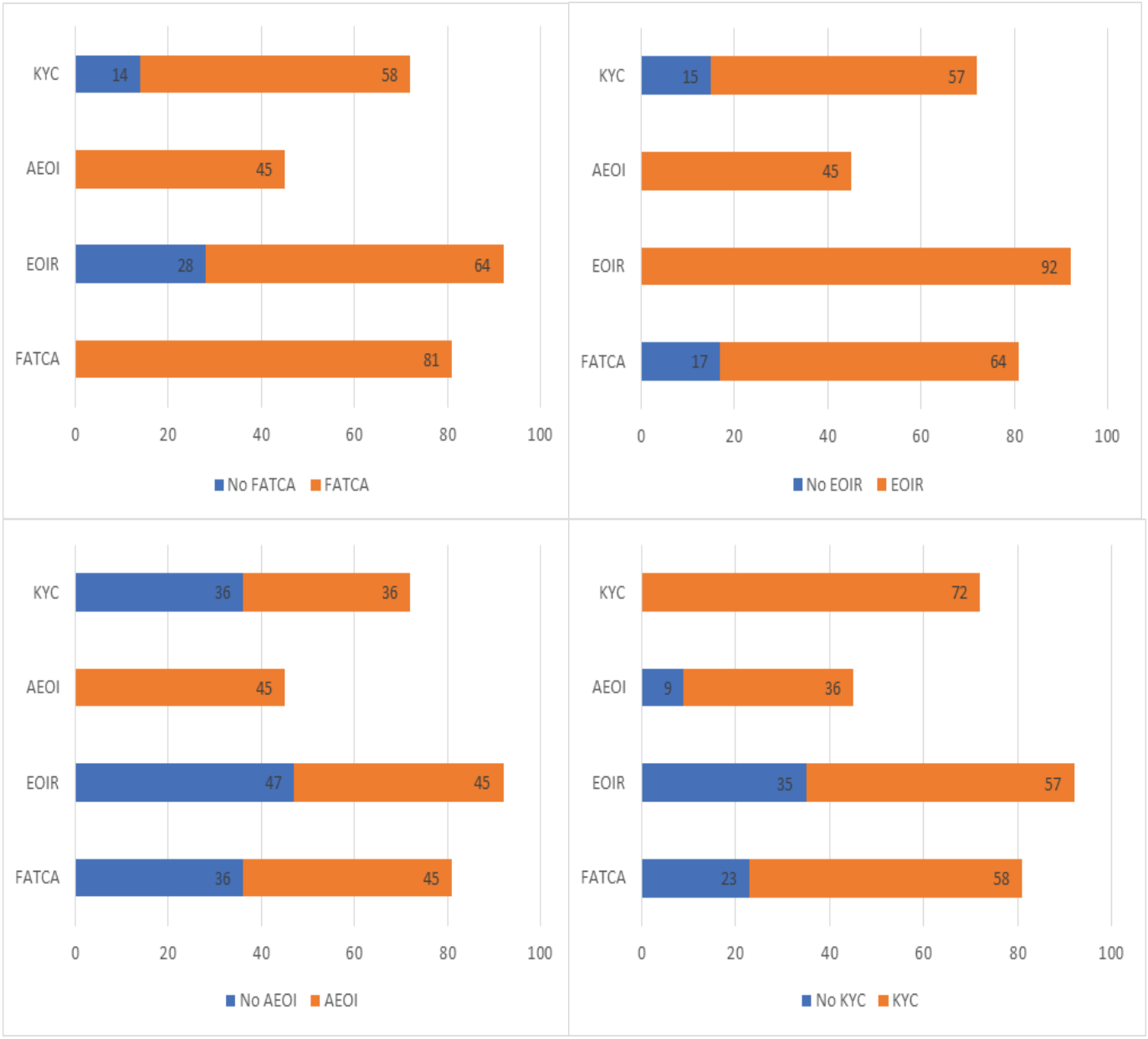
[24] https://home.treasury.gov/policy-issues/tax-policy/foreign-account-tax-compliance-act

[25] Rev. Proc. 2021-32, Section 3 (page 3) https://www.irs.gov/pub/irs-drop/rp-21-23.pdf

[26] Rev. Proc. 2021-32, Section 4 (page 6) https://www.irs.gov/pub/irs-drop/rp-21-23.pdf

[27] https://www.irs.gov/businesses/international-businesses/list-of-approved-kyc-rules

**FIGURE 1. Exchange of Information Indicators**

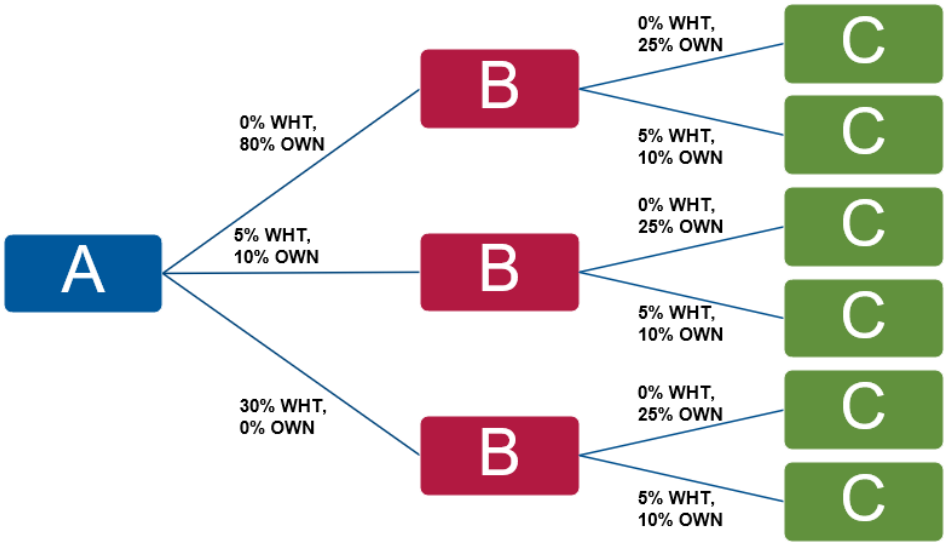


## 5. Sequence and Index Variable Construction

### *5.1 Sequence Construction*

The dataset contains 59,143 country pairs of treaty dividend withholding rates and associated ownership rates. We use this country pair dataset to create country sequences by linking together treaty rates. For example, consider three countries: A, B, and C. Dividend WHTs (DIV) between countries A and B depend on the ownership percentage (OWN) of the taxpayer in country B. In this example, there are 3 possible DIVs between countries A and B: 0% with 80% ownership, 5% with 10% ownership, and 30% with no minimum ownership requirement. Further suppose there are two possible DIV rates between countries B and C: 5% with 10% ownership and 0% with 25% ownership. When we link these WHTs across country pairs, we find six possible paths across the 3-country sequence: A-B-C. Figure 2 illustrates this construction and the six different WHT paths for the 3-country sequence.

**FIGURE 2. Stylized Example of Three Country Sequence: A-B-C**




After linking all combinations of the country pairs, we create 15,178,094 possible 3-country sequences that can originate from any country in the world. We call these "worldwide 3-country sequences". Of these, 80,564 are sequences that originate in the United States. Table 1 summarizes the available date for the 14,837,452 sequences with complete data for the right-hand side of the gravity equation.

The Dividend WHT and Ownership Rate data are for a pair of countries. Country 1-2 represents the average treaty WHT a taxpayer in Country 2 would pay for a dividend that originated in Country 1 and Country 2-3 is the average WHT a taxpayer in Country 3 would pay for a dividend that originated in Country 2. All other variables are presented for each of the three countries.

**TABLE 1. Gravity Model Summary Data for Worldwide 3-Country Sequences, 2017**

| Variable | | Count | Mean | Median | Min | Max |
|---|---|---|---|---|---|---|
| Dividend Withholding Tax | Country 1-2 | 14,837,452 | 11.229 | 10.0 | 0.0 | 36.0 |
| | Country 2-3 | 14,837,452 | 11.454 | 10.0 | 0.0 | 36.0 |
| Ownership Requirements | Country 1-2 | 14,837,452 | 2.039 | 0.0 | 0.0 | 100.0 |
| | Country 2-3 | 14,837,452 | 2.060 | 0.0 | 0.0 | 100.0 |
| Capital Gains Tax | Country 1 | 14,837,452 | 14.398 | 12.0 | 0.0 | 60.0 |
| | Country 2 | 14,837,452 | 14.549 | 12.5 | 0.0 | 60.0 |
| | Country 3 | 14,837,452 | 14.168 | 12.0 | 0.0 | 60.0 |
| Regulator Quality RQ (with imputations) | Country 1 | 14,837,452 | 0.526 | 0.506 | 0.027 | 0.931 |
| | Country 2 | 14,837,452 | 0.537 | 0.516 | 0.027 | 0.931 |
| | Country 3 | 14,837,452 | 0.522 | 0.502 | 0.027 | 0.931 |
| FATCA | Country 1 | 14,837,452 | 0.382 | 0.0 | 0.0 | 1.0 |
| | Country 2 | 14,837,452 | 0.413 | 0.0 | 0.0 | 1.0 |
| | Country 3 | 14,837,452 | 0.381 | 0.0 | 0.0 | 1.0 |
| EOIR | Country 1 | 14,837,452 | 0.447 | 0.0 | 0.0 | 1.0 |
| | Country 2 | 14,837,452 | 0.478 | 0.0 | 0.0 | 1.0 |
| | Country 3 | 14,837,452 | 0.434 | 0.0 | 0.0 | 1.0 |
| AEOI | Country 1 | 14,837,452 | 0.231 | 0.0 | 0.0 | 1.0 |
| | Country 2 | 14,837,452 | 0.257 | 0.0 | 0.0 | 1.0 |
| | Country 3 | 14,837,452 | 0.225 | 0.0 | 0.0 | 1.0 |
| KYC | Country 1 | 14,837,452 | 0.346 | 0.0 | 0.0 | 1.0 |
| | Country 2 | 14,837,452 | 0.371 | 0.0 | 0.0 | 1.0 |
| | Country 3 | 14,837,452 | 0.343 | 0.0 | 0.0 | 1.0 |

## *5.2 Index Variable Construction*

Equation 3 specifies the gravity model equation for cross-border tax avoidance which depends on four variables: DIV.OWN.path, cg.ratio, EOI.ratio, and RQ.path. In this section, we discuss how we construct each of the index variables. Directionality is a feature of the first three index variables, meaning the index variable is dependent on the order of the countries in the sequence. This means that sequences A-B-C, A-C-B, B-C-A, B-A-C, C-A-B, and C-B-A could each have different values for these variables, and therefore different estimated gravity indexes, despite all being sequences made up of the same three countries.

DIV.OWN.path is constructed across a path by multiplying (1–DIV.path) times (1–OWN.path), as shown in Table 2. Consider a 3-country sequence, A-B-C, with 9 possible sets of dividend withholding rates across the full sequence. Columns [1]–[4] contain dividend withholding tax and required minimum ownership rates for each of the 2 country pairs in the sequence, A-B and B-C. Column [5] constructs the first part of DIV.OWN.path, (1–DIV) as the product of 1 minus each dividend WHT across the sequence. For the second term, (1–OWN.path), we first need to calculate the implied ownership of an investor in Country C into Country A. That is, for a person in Country C who owns a portion of a company in Country B that in turn owns a share of a company in Country A, what is the minimum required ownership for the Country C investor in the Country A company.[28] We calculate this in column [6] by multiplying the minimum owner-

[28] All tax treaties with differentials in withholding rates contain percentage ownership criteria. We use the percentage ownership variable as a weight that provides variation among country paths. For example, some countries will require a large ownership stake in an entity to receive cross-border dividend payments free of withholding tax. The ownership variable reflects the "price of admission" to the lower withholding tax rate. It restricts the pool of investors who can qualify for the lowest withholding tax rates. Of the 12 U.S. tax treaties with a 0% withholding tax on outbound dividends, 11 require an 80% ownership stake, and 1, with Japan, requires a 50% ownership stake. In tax treaty data, the highest withholding tax rates generally have no percentage ownership criteria.

ship for an investment from Country C in Country B, column [4], times the minimum ownership for an investment from Country B in Country A, column [2]. Column [7] constructs the second term as the product of (1–ownership between B and C) and (1–implied ownership between C and A), or (1–column [4]) times (1–column [6]). Note that the effect from including a conduit country in this sequence is to reduce the required minimum ownership of an investor in the country A entity from 80% (row 1 column 2) to 4% (row 1 column 6). This can expand the pool of investors eligible for 0% withholding rates on cross-border income.

Solely based on these treaty rates, our model would call the top row the "BEST" option out of these 9 and the bottom row would be deemed the "WORST". A "BEST" path is one with the lowest possible WHT and minimal ownership requirements across the sequence. A "WORST" path is one with the largest combination of withholding tax and ownership rates.

**TABLE 2. Stylized Example of Construction**

| DIV A-B | OWN A–B | DIV B C | OWN B C | (1 - DIV) path | Implied OWN A–C | ( 1 - OWN) path | DIV.OWN. path |
|---|---|---|---|---|---|---|---|
| [1] | [2] | [3] | [4] | [5] | [6] | [7] | [8] |
| | | | | (1 - [1]) * (1 - [3]) | [4] * [2] | (1- [4]) * (1- [6]) | [5] * [7] |
| 0 | .80 | 0 | .05 | 1 | .04 | 0.912 | 0.912 |
| 0 | .80 | 0 | .10 | 1 | .08 | 0.828 | 0.828 |
| 0 | .80 | .15 | 0 | 0.85 | 0 | 1 | 0.850 |
| .05 | .10 | 0 | .05 | 0.95 | 0.005 | 0.94525 | 0.899 |
| .05 | .10 | 0 | .10 | 0.95 | .01 | 0.891 | 0.846 |
| .05 | .10 | .15 | 0 | 0.8075 | 0 | 1 | 0.808 |
| .30 | 0 | 0 | .05 | 0.7 | 0 | 0.95 | 0.665 |
| .30 | 0 | 0 | .10 | 0.7 | 0 | 0.9 | 0.630 |
| .30 | 0 | .15 | 0 | 0.595 | 0 | 1 | 0.595 |

Variable *CG.ratio* measures the ratio of capital gains taxes in the destination country to the capital gains taxes in the origin country. We obtain the country level capital gains tax rate on shares (as opposed to the capital gains tax rate on immovable property) for nonresident individuals from IBFD. Directionality is an important feature of this variable. A capital gains tax rate of 0% in the destination country is more attractive to a person leaving a country with a high capital gains tax rate than it is a person leaving a country that also has a 0% capital gains tax rate. To capture this directional feature, we calculate *CG.ratio* as:

$$CG.ratio = \frac{(1 - CG_3)}{(1 - CG_1)} \tag{5}$$

where $CG_3$ is the capital gains tax rate in the destination country and $CG_1$ is the capital gains tax rate in the origin country. A value greater than 1 indicates improvement in the capital gains tax rate, while a value less than 1 indicates that the taxpayer is worse off along this dimension. A value equal to 1 indicates no change in the capital gains tax rate between the destination and origin countries.

The third variable with direction is *EOI.ratio*. The EOI variable equals 1 when a country participates in an exchange of information program and 0 when not a participant. For 3-country sequences, there are 8 possible combinations of EOI variables. Table 3 shows these eight combinations and how we construct *EOI.ratio*. First, *EOI.path* is based on the average of EOI for the three countries in the sequence. It is defined as 1 / (1 + average(EOI)). We add 1 in the denominator to

create a number that is larger than 0 so that we can use the measure in a log regression. One thing to notice about *EOI.path* is the lack of direction. Any sequence where only 1 country participates in an EOI program has EOI.path equal to 0.75, and it doesn't matter if that is the origin, conduit, or destination country. That is, (0,0,1), (1,0,0), and (0,1,0) all have the same value but clearly (1,0,0) should be preferred to (0,0,1) because a good destination country—the third number in each triple—would not be transparent with tax administration in other countries. To add directionality, we first take the ratio of EOI in the destination to EOI in the origin country, or $\frac{(1+EOI_C)}{(1+EOI_A)}$. Next, we divide *EOI.path* by this ratio to arrive at *EOI.ratio* in the final column. The largest value is the most attractive for tax avoidance: leaving a country with EOI participation and moving through two countries with no EOI participation. The smallest value is the least attractive: starting in a country with no EOI and moving through two countries with EOI participation.

The final index variable, *RQ.path*, is the simple average for *RQ* across each country in a sequence (See footnote 14).

**TABLE 3. EOI.Ratio Construction**

| Country A | Country B | Country C | EOI.path | $(1 + EOI_C) / (1 + EOI_A)$ | EOI.ratio |
|---|---|---|---|---|---|
| 0 | 1 | 1 | 0.60 | 2 | 0.30 |
| 0 | 0 | 1 | 0.75 | 2 | 0.375 |
| 1 | 1 | 1 | 0.50 | 1 | 0.50 |
| 1 | 0 | 1 | 0.60 | 1 | 0.60 |
| 0 | 1 | 0 | 0.75 | 1 | 0.75 |
| 0 | 0 | 0 | 1 | 1 | 1 |
| 1 | 1 | 0 | 0.60 | 0.5 | 1.2 |
| 1 | 0 | 0 | 0.75 | 0.5 | 1.5 |

### *5.3 Dependent Variable Construction*

We cannot observe worldwide cross-border tax avoidance. Instead, we estimate the gravity index parameters using a constructed measure of financial flows that we call *Inward.Adjust*. We obtain Foreign Direct Investment from the IMF Coordinated Direct Investment Survey (CDIS).[29] The CDIS provides data on *Inward FDI* and *Inward Derived FDI* between country pairs. For countries who did not report any FDI, the CDIS contains *Inward Derived FDI* determined when a partner country reports *Outward FDI* to that country. For example, consider two countries A and B. Country A reports $50 outward FDI to Country B but country B did not report receiving inward FDI from country A. The CDIS would show that Country B has $50 derived Inward FDI from Country A. We use reported *Inward FDI* whenever available and fill in missing values with *Inward Derived FDI* when that measure is available.

We construct a measure of Adjusted FDI for each 3-country sequence that is the portion of Inward FDI from Country C into Country B that could possibly make it into Country A. Table 4 contains a stylized example. Suppose $100 of Inward FDI in Country A comes from Country B (Inward.1) and that total Inward FDI into Country A is $1,000 (Inward.Total.1). Inward.2 is the amount of Inward FDI from the country listed in the third column into Country B and this amount totals $735 (Inward.Total.2). We know that only a maximum of $100 out of this $735 of Inward FDI into Country B is invested into Country A.

We adjust all amounts proportionally by the ratio of Inward.1 to Inward.Total.2 (adjust) and multiply this factor times the amounts in Inward.2 to derive what we are calling *Inward.Adjust*. This is the maximum amount of Inward FDI from each country into Country B that could eventually become Inward FDI into Country A if all countries C, D, E, F, G, H invest in Country A through their investment in Country B. Notice that *Inward.Adjust* sums up to 100, and for the first sequence it is $27.[30]

[29] IMF Coordinated Direct Investment Survey data available for download here: https://data.imf.org/?sk=40313609-f037-48c1-84b1-e1f1ce54d6d5

[30] The construction of Inward.Adjust in table 4 assumes that country B does not invest in country A. In other words, country B acts as a pure conduit by bundling investment from countries C to H and sending some of the $735 bundle onto country A in the amount of $100. It also assumes that country A investment in country B does not "round trip" back to country A. There is evidence of round tripping in FDI data. In subsequent versions we will relax these assumptions.

**TABLE 4. Dependent Variable Construction**

| Country | | | Inward.1 | Inward.2 | Inward. Total.1 | Inward. Total.2 | adjust | Inward. Adjust |
|---|---|---|---|---|---|---|---|---|
| A | B | C | 100 | 200 | 1000 | 735 | 0.136054 | 27 |
| A | B | D | 100 | 50 | 1000 | 735 | 0.136054 | 7 |
| A | B | E | 100 | 25 | 1000 | 735 | 0.136054 | 3 |
| A | B | F | 100 | 300 | 1000 | 735 | 0.136054 | 41 |
| A | B | G | 100 | 10 | 1000 | 735 | 0.136054 | 1 |
| A | B | H | 100 | 150 | 1000 | 735 | 0.136054 | 20 |

## 6. Estimating Gravity Model Weights

Table 5 presents mean values of these constructed index variables for all worldwide sequences and several subsamples. To estimate the gravity index weights, only the BEST paths for each country sequence is used (reminder, BEST paths are those sequences with the most advantageous treaty withholding rates and ownership requirements as determined by *DIV.OWN.path*). To use the full range of treaty withholding rates for dividends we could have up to 12 different withholding rate paths for a given 3-country sequence. But we do not have a dependent variable that can identify each of these 12 different paths. We could have 12 different withholding paths for the same sequence of countries all associated with a single value of the dependent variable and this would introduce a lot of noise. Instead, we only include BEST paths in the regression. Country sequences with no treaty rates have only one set of withholding taxes and ownership requirements and do not face this problem. They are all considered BEST paths.

We have complete data for 14,837,452 sequences 3-country sequences, and of these, 11,696,856 represent the BEST path within a sequence. Approximately 4 million BEST paths have enough Inward FDI data to construct *Inward.Adjust*. And finally, only 670,281 paths are used in the regression where the log-specification removes any values less than or equal to 0.[31]

**TABLE 5. Constructed Index Variables for Worldwide Sequences, 2017**

| Variable or Stat | ALL | BEST | With FDI | No FDI | Reg Sample |
|---|---|---|---|---|---|
| Path: (1 - DIV WHT) | 0.786 | 0.796 | 0.796 | 0.795 | 0.828 |
| Path: (1 - OWN) | 0.979 | 0.995 | 0.992 | 0.997 | 0.982 |
| DIV.OWN | 0.768 | 0.791 | 0.790 | 0.792 | 0.812 |
| CG ratio | 1.036 | 1.032 | 1.034 | 1.031 | 1.040 |
| Path: WB RQ | 0.532 | 0.514 | 0.543 | 0.498 | 0.600 |
| FATCA w/ direction | 0.828 | 0.849 | 0.804 | 0.873 | 0.691 |
| EOIR w/ direction | 0.809 | 0.828 | 0.771 | 0.859 | 0.659 |
| AEOI w/ direction | 0.905 | 0.920 | 0.875 | 0.859 | 0.789 |
| KYC w/ direction | 0.851 | 0.868 | 0.834 | 0.885 | 0.729 |
| **Count:** | **14,837,452** | **11,696,856** | **4,075,933** | **7,620,923** | **670,281** |
| **Count w/ FDI** | **6,067,599** | **4,075,933** | **4,075,933** | **0** | **670,281** |
| **Avg Adjusted FDI ($B)** | **24.1** | **8.1** | **8.1** | **---** | **49.8** |

We estimate the gravity index parameters according to equation (4) as shown in Table 6. We consider each of the four EOI indicators separately and all together in the final specification. Our preferred specification includes all four EOI indicators jointly because it shows variation between the indicators.

[31] Our dependent variable *inward.adjust* can have negative values. This happens when a country removes investment made in prior years from another country. Often this is the result of a sale of the invested assets in the country where for accounting purposes, the "return" of capital from the sale is recorded as negative FDI. As a result, when a sequence of countries has a negative inward FDI, it is excluded from estimating the gravity equation when we apply a log transformation to the data. We address negative values of inward FDI in subsequent versions of the model.

**TABLE 6. Gravity Index Weight Estimation (Dependent Variable = *Inward.Adjust)***

| | FATCA | EOIR | AEOI | KYC | ALL |
|---|---|---|---|---|---|
| Constant | 1.322*** | 1.502*** | 1.646*** | 1.255*** | 0.892*** |
| | [62.820] | [72.419] | [81.953] | [60.600] | [40.889] |
| log DIV.OWN.path | 3.704*** | 3.606*** | 3.666*** | 3.672*** | 3.719*** |
| | [88.327] | [85.888] | [87.272] | [87.754] | [89.047] |
| log cg_ratio | 0.153*** | 0.278*** | 0.552*** | -0.0026 | 0.232*** |
| | [6.543] | [11.784] | [22.911] | [-0.115] | [9.539] |
| log RQ.path | 11.554*** | 11.907*** | 11.793*** | 11.436*** | 11.191*** |
| | [364.916] | [383.541] | [376.242] | [361.825] | [350.519] |
| log FATCA.ratio | -1.146*** | | | | -0.339*** |
| | [-98.966] | | | | [-20.389] |
| log EOIR ratio | | -1.032*** | | | -0.306*** |
| | | [-87.114] | | | [-20.556] |
| log AEOI ratio | | | -1.037*** | | -0.279*** |
| | | | [-86.033] | | [-17.194] |
| log KYC ratio | | | | -1.274*** | -0.845*** |
| | | | | [-112.249] | [-60.142] |
| Observations | 670,281 | 670,281 | 670,281 | 670,281 | 670,281 |
| R2 | 0.225 | 0.2225 | 0.2223 | 0.2282 | 0.2316 |
| Adjusted-R2 | 0.225 | 0.2225 | 0.2223 | 0.2282 | 0.2316 |
| F statistic | 48,669.05 | 47,967.80 | 47,908.29 | 49,561.12 | 28,866.36 |

Notes: Standard errors in brackets: *** p < 0.001; ** p < 0.01; * p < 0.05.

These estimates are intuitive. For 3-country sequences, lower dividend withholding rates and lower ownership rates increase inward FDI from destination countries through conduit countries to a source country. Lower capital gain taxes in the destination country relative to the source country also increase inward FDI to the source country. Greater regulator quality across the entire sequence increases inward FDI to the source country. Finally, a greater presence of tax administration information exchange agreements (FATCA, EOIR, AEOI and KYC) across a sequence of countries reduces inward FDI to a source country.

We use the estimated coefficients from our preferred specification on Table 6 (ALL) and plug them into our gravity model equation, which with some rearranging looks like this:

$$DIV.OWN.path^{\beta_1}(1 - CG.ratio)^{\beta_2} RQ.path^{\beta_3}(1 + EOI.path)^{-\beta_4} \quad (7)$$

Equation (7) allows us to weight the gravity index for all 3-country sequences ("triplet") with complete data (11,696,852 BEST sequences).[32]

Because the number of potential country sequences expands exponentially with the length of the sequence, we create a stopping rule to reduce computational complexity. To do this we generate a weighted index for each 2-country sequence of countries ("pair") using the same estimated weights. If the weighted index for the 3-country sequence is larger than the weighted index for the 2-country sequence, then move on to Country 3. Otherwise, stay in Country 2. This rule substantially reduces the set of potential triplets down to 4 million and will make it computationally possible to construct longer sequences.

[32] We are aware that the 670,281 3-country sequences used in estimation had average FDI of $49.8 B while the 11,696,852 BEST sequences had average FDI of $8.1 B. In subsequent estimation we will explicitly control for this discrepancy when calculating the gravity index for the many small FDI sequences in BEST that are not in the estimation.

To create longer sequences, notice that a 4-country sequence is comprised of two 3-country sequences. We chain sequences by multiplying the index for each triplet and implementing a test to see if it's better to move on to the fourth country or to remain in the third country. Figure 3 illustrates this.

**FIGURE 3. Chaining Indexes to Create Longer Sequences**


Consider two triplets (A-B-C) and (B-C-D) with weighted gravity indexes Index1 and Index2. We link these two indexes together and construct *Chained.Index = Index1 * Index2*. We then compare *Chained.Index* to *Index1 * Index1*. If , then the gravity model predicts that it is advantageous to move on to country D. If *Chained.Index* ≤ *Index* $1^2$, then it is best to remain in Country C. Using the weighted gravity indexes for all possible country triplets, we can link countries indefinitely. The general formula for the chained index is then: $Chained.Index = \prod_{i=1}^{n} Index_i$, where *n* represents the number of triplets in the sequence. And the stopping rule comes from comparing *Chained.Index* to $Index_1 n$.

## 7. Applications, use cases, and next steps

The weighted gravity model indexes can be used to predict the most likely conduits and destinations for any set of countries. Given a source country, we use the model to look at sequences with the largest index values to find the most attractive conduits and destinations. Any of 230 countries that we have data for can be a source country.

The use of FDI to construct our dependent variable is a compromise. FDI is antecedent to cross-border flows of income as dividends, interest, and royalties. As such, it is not the measure that we would prefer, which would be actual amounts of these cross-border income flows. Ultimately, we would like to replace FDI with other highly observable financial flows for which the character of the income can be ascertained. This will more closely align the dependent variable with the richness of tax treaty data that can be fully expressed in the gravity equation specification, but which we cannot fully exploit.

This model is based on data for 2017. As we acquire treaty data for additional years and update the model to run for each year going forward, we anticipate observing changes in FDI sequences that reflect the changing landscape of tax treaties and tax administration transparency.

Our first version of this model is based on WHTs for cross-border dividend payments. We anticipate extending this model to cross-border interest and royalty payments as well. It is also possible that an outbound flow of dividends from a source country could morph into another income character type as it crosses multiple borders. This is because each country accepting an inbound flow of income might have different tax rules that incentivize certain types of income as outbound payments.

# Appendix 1. Imputation for missing World Bank Regulator Quality

**TABLE A1. Regulator Quality Imputation**

| | Variable | Coefficient |
|---|---|---|
| | Intercept | -4.165*** |
| | | [-14.192][1] |
| | | 0.456*** |
| | | [12.978] |
| Information Sharing Agreements | FATCA2017 | 0.414*** |
| | | [3.546] |
| | AEOI2017 | 0.219 |
| | | [1.621] |
| Multilateral Treaty Participation | European Economic Area (EEA) | 0.082 |
| | | [0.607] |
| | East African Community (EAC) | 0.600** |
| | | [2.686] |
| | Caribbean Community (CARICOM) | -0.169 |
| | | [-1.037] |
| | WAEMU | 0.460* |
| | | [2.371] |
| | Arab Maghreb Union | -0.356 |
| | | [-1.492] |
| | CEMAC | -0.475* |
| | | [-2.197] |
| | Arab Economic Union Council | -0.492** |
| | | [-3.224] |
| Territory Status | French Department | 3.346*** |
| | | [8.537] |
| | Netherlands Constituent Country | 0.667 |
| | | [1.285] |
| | UK Territory | -0.279 |
| | | [-0.896] |
| | UK Dependency | -0.744 |
| | | [-1.415] |
| | US Territory | -0.074 |
| | | [-0.276] |
| | Observations | 209 |
| | R2 | 0.739 |
| | F-statistic | 40.281 |

Notes: t-statistics reported in brackets: *** p < 0.001; ** p < 0.01; * p < 0.05. CEMAC= Economic and Monetary Community of Central Africa. WAEMU= West African Economic and Monetary Union

**TABLE A2. Regulator Quality Imputation Results**

| ISOCode | RQ | RQ.imputed | ISOCode | RQ | RQ.imputed | ISOCode | RQ | RQ.imputed | ISOCode | RQ | RQ.imputed |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HKG | 2.167 | 2.167 | JPN | 1.377 | 1.377 | BES | NA | 1.101 | NCL | NA | 0.595 |
| SGP | 2.118 | 2.118 | GIB | NA | 1.376 | CYP | 1.032 | 1.032 | SVN | 0.58 | 0.58 |
| NZL | 2.092 | 2.092 | TWN | 1.372 | 1.372 | MUS | 1.029 | 1.029 | FLK | NA | 0.525 |
| NLD | 2.051 | 2.051 | CHL | 1.35 | 1.35 | ARE | 1.015 | 1.015 | BRB | 0.495 | 0.495 |
| AUS | 1.933 | 1.933 | GRL | 1.324 | 1.324 | ESP | 0.945 | 0.945 | ROU | 0.488 | 0.488 |
| CAN | 1.89 | 1.89 | MCO | NA | 1.322 | PRT | 0.911 | 0.911 | URY | 0.476 | 0.476 |
| CHE | 1.887 | 1.887 | MLT | 1.285 | 1.285 | IMN | NA | 0.894 | MNP | NA | 0.425 |
| FIN | 1.823 | 1.823 | GUF | 1.282 | 1.282 | POL | 0.881 | 0.881 | HRV | 0.424 | 0.424 |
| NOR | 1.816 | 1.816 | ISR | 1.274 | 1.274 | PRI | 0.872 | 0.872 | OMN | 0.423 | 0.423 |
| SWE | 1.801 | 1.801 | GLP | NA | 1.252 | FRO | NA | 0.865 | QAT | 0.42 | 0.42 |
| DEU | 1.786 | 1.786 | BEL | 1.247 | 1.247 | VGB | NA | 0.85 | BHR | 0.416 | 0.416 |
| MAC | 1.76 | 1.76 | CZE | 1.235 | 1.235 | BMU | 0.844 | 0.844 | PAN | 0.388 | 0.388 |
| GBR | 1.717 | 1.717 | MTQ | 1.21 | 1.21 | VIR | 0.844 | 0.844 | PYF | NA | 0.378 |
| LUX | 1.694 | 1.694 | REU | 1.21 | 1.21 | SVK | 0.826 | 0.826 | COK | NA | 0.35 |
| EST | 1.645 | 1.645 | AND | 1.21 | 1.21 | GGY | NA | 0.768 | COL | 0.341 | 0.341 |
| USA | 1.631 | 1.631 | SXM | NA | 1.195 | CYM | 0.756 | 0.756 | LCA | 0.307 | 0.307 |
| DNK | 1.624 | 1.624 | ABW | 1.194 | 1.194 | BRN | 0.718 | 0.718 | KNA | 0.293 | 0.293 |
| IRL | 1.588 | 1.588 | FRA | 1.16 | 1.16 | ITA | 0.706 | 0.706 | MSR | NA | 0.284 |
| LIE | 1.497 | 1.497 | LTU | 1.159 | 1.159 | JEY | 0.683 | 0.683 | MEX | 0.279 | 0.279 |
| AUT | 1.44 | 1.44 | LVA | 1.157 | 1.157 | HUN | 0.652 | 0.652 | GRC | 0.24 | 0.24 |
| ISL | 1.435 | 1.435 | SMR | NA | 1.141 | TCA | NA | 0.634 | ZAF | 0.234 | 0.234 |
| CUW | NA | 1.405 | KOR | 1.108 | 1.108 | BGR | 0.626 | 0.626 | NIU | NA | 0.228 |

Notes: Shaded countries are imputed. Countries with smaller values of RQ than NIU are not shown as none of those countries had missing RQ.

# Art in the Age of Tax Avoidance

*Matthew Pierson[1] (WRDS, University of Pennsylvania)*

## 1. Introduction

The United States utilizes fewer offshore tax havens relative to other developed countries, at a rate of about 6% of GDP compared to 14% in 2022 (EUTO Offshore Atlas (2023)). One reason for this difference is that the U.S. offers many of the same offshore opportunities for domestic tax avoidance (Hemel (2022); Alstadsæter, Johannesen, and Zucman (2018)). Prior work suggests that the tax deductibility of charitable donations makes offshore tax evasion less valuable, with evidence that the sensitivity of charitable donations is especially sensitive to tax changes in the U.S. (see Duqette (2016); Duqette (2019); Meer and Priday (2020a, 2020b); Brounstein (2023); Ring and Thoresen (2023); and Fack and Landais (2016)).

Yet the nonprofit sector is unquestionably engaged in the private provision of public goods. Individuals donated $500 billion in 2022 to nonprofits, with charitable causes ranging from food banks to higher education, and total assets worth $16.8 trillion. An extensive and rich literature has emerged studying nonprofits and their importance as an alternate provider of public goods (see List (2011) or Gee and Meer (2019) for reviews). Donations to nonprofits allow individuals to socially signal, (Glazer and Konrad (1996)) or gain personal satisfaction, which, inclusive of the tax advantages, may maximize social welfare (Diamond (2006); Saez (2004)).

In this paper, I seek to resolve the tension between these two aspects of U.S. nonprofits—assistance of tax avoidance and charitable provision. First, I construct a comprehensive sample of all digitized IRS Form 990 tax-exempt organization annual filings from 2011–2022. To my knowledge, this is the first study examining fine art donated to nonprofit organizations in the United States. This sample allows me to examine the extent of tax-deductible donations of art, and to ascertain when nonprofits choose to record, value, and revalue art. This sample is extensive, totaling 5.4 million nonprofit years, which allows observation of significant variation in the type, behavior, and fundamentals of nonprofits.

Second, this sample allows me to establish a number of new key facts. Only 1.2% of Form 990 filing organization-years identify holding art, with only 0.2% choosing to recognize these donations as revenue and value these assets on their balance sheets. Despite this small subsample, observed art holdings are large-worth around $6 billion in 2022, with annual donations worth $300 million. Donations are numerous, and typically quite small, with 2022 total count of donated items around 200,000, and an average value of a little over $1,500. Furthermore, art donations to nonprofits are associated with both seemingly typical, charitable activities as well as potential hallmarks of secrecy and agency problems.

Among these stylized facts is that the choice to value art, conditional on accepting it, is related to the charitable organization type. Due to ethics concerns for museums and other nonprofits, and perhaps counterintuitively, organizations that are more likely to hold permanent "collections" of art are also those least likely to provide valuations of them. Simultaneously, within these organization types, there is still significant heterogeneity the stated usage of art. This heterogeneity follows a similar pattern, with art held for public exhibit among the least likely to be valued, and art held for research, loan, or other purposes among the most.

Moreover, the size of the donation matters greatly for its likelihood of overvaluation. Below the $5,000 threshold that requires qualified appraisal, there is significant bunching of average yearly donation values and donations are much more likely to be donated at values that the organization subsequently writes down. Donation valuation methods provided by donors are systematically more likely to be overvalued as well, but again, below this mandatory qualified appraisal threshold. This behavior allows donors to book tax deductions at inflated amounts, while the nonprofit does not violate any IRS requirement to obtain outside, accurate valuations.

Third, the choice to disclose, value, and revalue art donations is dependent on the likelihood of IRS audit. Nonprofits respond to behavior that engages in potential audit flags by being more likely to disclose art, more likely to value it

[1] Matthew Pierson: Wharton Research Data Services (WRDS), The Wharton School, University of Pennsylvania, Email: mpiers@wharton.upenn.edu. I thank Will Boning, Aart Gerritsen, Cristi Gleason, Jim Hines, Stacie K. Laplante, Juliana Londoño-Vélez, Jennifer Mayo, Rainer Niemann, Jake Thornock, and participants of the EIASM Conference on Current Research in Taxation 2023, the IIPF 2023 Annual Congress, and the 2024 IRS-TPC Joint Conference on Taxation for helpful comments and discussions. I thank Jakob Brounstein for graciously sharing nonprofit family foundation data and for his discussions. I am responsible for all remaining errors and omissions. I have nothing to disclose.

conditional on filings, and, conditional on valuation, more likely to revalue these assets down. From this response, I generate estimates of tax losses. Art is donated at inflated valuations, leading to about $734 million in tax losses over our sample period. Extrapolating these losses based on audit flag rates, potentially up to $5.5 billion in taxes have been lost due to inflated art donations from 2011 to 2022.

The setting of art donations to nonprofits is useful for several reasons. Art has been historically understood as used for money laundering and tax avoidance purposes. From anecdotal cases and focused settings in both popular press and research (e.g., ICIJ (2022); Oliver (2021); U.S. Senate (2020); Ang (2020); Harrington (2016); and De Simone, Lester, and Markle (2020)), little is comprehensively documented about the use of art as a tax avoidance tool. Several features of art markets—illiquidity, subjective value, asset portability,[2] little public information on transaction history, and lack of regulation[3]—can create substantial differences in transaction prices for any single asset, allowing for dislocations from its "true" value. A nascent literature has begun to document this. De Simone, Lester, and Markle (2020) find spillover effects from financial asset tax enforcement efforts into art assets, as international responses to Foreign Account Tax Compliance Act (FATCA) led to increased art holdings in the Geneva freeport. Londoño-Vélez and Ávila-Mahecha (2023) find that Colombian individuals respond to wealth taxes by shifting their wealth composition into harder to value assets like art, while simultaneously under-stating this value to avoid taxes. In this setting, I provide the first comprehensive evidence of fine art charitable donations for tax avoidance in the U.S.

Further, by focusing on a narrow type of donation, I can more finely isolate tax avoiding behavior in a narrow context, void of potentially confounding factors. Many studies have documented the relationship between the tax-deductibility of charitable giving and taxes. Recently, Ring and Thoresen (2023) find that households subject to a wealth tax in Norway reduce charitable giving in response to it. Fack and Landais (2016) survey this historical relationship across several countries. Detailed donations data, paired with tax year-end valuation of assets by type allow for disentangling the role of donor and nonprofit, the effects of charitable and tax avoiding behavior, and the effects of potential audits on all these actors and their behavior. The IRS identifies distinct categories of art[4] that, when compiled, illustrate a comprehensive picture of the types of assets entities may hold or seek to acquire. This classification as well as the public disclosure of nonprofit organization tax filings creates a unique laboratory to study the role of nonprofit organizations in facilitating tax avoidance.

Finally, there is an open question about the role of nonprofits in public good provision. The theory that involuntary public good provision crowds out voluntary provision has sprung out of Bergstrom, Blume, and Varian, 1986, (e.g., Cornes and Itaya (2010); Villanacci and Zenginobuz (2007, 2012)). Empirical tests of this theory (Hungerman (2005); Andreoni and Payne (2003); Andreoni and Payne (2011)) often find crowd out, but it is imperfect. There are many explanations for crowd out or lack thereof, but one potential aspect that has not been well studied is the role of tax avoidance and asset protection in the sizable assets held by nonprofits. That is, the lack of direct charitable motive for many nonprofits may provide some role in resolving this question. Non-charitable motives may explain the lack of crowd out that I observe empirically due to non-charitable purposes.

U.S. charitable donations of art assets are substantial, worth one% of 2022 U.S. art sales (McAndrew (2023)). Across organizations, there is significant variation in the determinants of holding art, providing detailed estimates of its value, and updating these estimates over time. These determinants are associated with public good provision, as well as facets of tax avoidance—secrecy, asset protection, and poor compliance. Due to variation in reporting requirements, of some 5,364,313 organization-years in my sample only 1.2% report receiving a donation of fine art at any point.[5] However, of this 1.2%, only 16% supply valuations for donations and holdings. This narrows the sample of observable holdings to 9,801 organization-years or 0.2% of nonprofit organizations. However, despite potentially sizable assets and donations, much of

[2] See De Simone, Lester, and Markle (2020) for a discussion of the use of freeports in allowing storage and viewing of art while obtaining tax advantages.

[3] Dealers of antiquities, as of March 2021, are required to file Suspicious Activity Reports (SARs) to FinCEN regarding suspected use of antiquities in the financing of terrorism or the facilitation of money laundering. Any fine art that is not classified as an antiquity is otherwise largely devoid of regulation on its sale. See: https://www.fincen.gov/sites/default/files/2021-03/FinCEN%20Notice%20on%20Antiquities%20and%20Art508C.pdf

[4] These categories are titled Art Assets, Fractional Interest of Art, Historical Treasures, Historical Antiquities, Works of Art, and Other Qualified Contributions of Art.

[5] Two checkbox items on Form 990 indicate the presence of art holdings. Filers who check either of these boxes are required to fill out Schedule D and/or Schedule M, detailing the value of these assets as well as the value of donations in that tax year.

the value of art is potentially left unreported. There is clear selection by nonprofit organizations in the decision to record complete and accurate tax filings.

Art donations are associated with both typical organizations—museums, libraries, and educational institutions—and other organization types less commonly associated—private and family foundations or medicals institutions. However, the heterogeneity in accepting art donations becomes clearer with the choice to value art donations. Nonprofits may choose to accept art donations and classify them as a "collection" which must follow specific IRS guidelines. An organization can then choose not to "capitalize"—or value—a collection, following the Financial Accounting Standards Board (FASB)'s Accounting Standards Code (ASC) 958-360-25. This is, in fact, suggested for certain organization types like museums. I find evidence for this, as libraries (2.5% provide valuations), religious organizations (5%), and museums, private foundations, and other organizations (9%) are less likely to value art collections, while those that are not typically associated with collections like medical (21%) and educational (36%) organizations are much more likely to value art.

There is substantial variation in valuing art not just across organization types but within organization types by the stated usage of art assets. Art used for public exhibit (16%) and preservation (14%) is less likely to be valued, while that held for research (26%), loan (25%), or other (28%) purposes is more likely to be valued. Organization types are correlated with stated use categories, but imperfectly so, and the probability of valuation rises with organization type and stated use together.

Conditional on valuing art, organizations often choose to write down it's value across my sample. Doing so reveals a wedge between the inflated value of art donations, and its true value to the organization as an asset. Donations of art and other non-monetary contributions are required to be valued by a qualified appraisal if they claim over $5,000 in income tax deductions for the donor. This threshold creates a discontinuity, with bunching of average donation values just below it. Donations that are below this threshold are 33% (8 percentage points) more likely to have their value written down later (which I refer to as "overvalued"). Therefore, there is significant strategic donations of art by individuals to nonprofits, with donations bunching below the required appraisal threshold, and these donations significantly more likely to be overvalued.

Moreover, nonprofits may identify the valuation method of art donations. Donation methods that are influenced by donors, like those providing valuations from auction purchase prices or supplying their own valuation, are significantly more likely to be overvalued. Separating these valuation methods below and above the required qualified appraisal threshold, all valuation methods lead to significantly more overvaluation below the threshold than above it. Nonprofits appear to accept overvalued donations largely from a lack of due diligence, and correctly value art donations only after they have accepted them at inflated, tax deductible values from donors. Why do nonprofits revalue art? While nonprofits more consistently correctly value art when required to, the choice to revalue these assets is also a matter of compliance. There is a consistent factor across the choice to accept art donations, record the value of those donations and holdings, and revalue them post-donation. Audit flags are consistently related to the decision to properly disclose art on the extensive and intensive margins. Examining this more deeply, I find that nonprofits respond to activities that increase IRS audit probability are 221% more likely to be associated with disclosure of art, conditional on filing art, 87% more likely to be associated with disclosing its value, and conditional on valuing art, 12% more likely to be associated with re-valuing it correctly. Nonprofits respond to increased scrutiny by more accurately filing their Form 990s.

The threat posed by tax authorities leads nonprofits to conduct the due diligence of correctly valuing their art holdings. This leads to writing down the value of art holdings.[6] The threat for nonprofits includes non-filing penalties for each Form 8282, that is, each donation, for any donations that should have been appraised but were not. Further, consistent misconduct may lead to revocation of 501(c)(3) status. With these findings, I disentangle the effect of art donations for charitable purposes versus those for tax avoidance. A growing literature documents taxpayers moving wealth into assets or investment types that remain largely unobserved by tax authorities (See Caruana-Galizia and Caruana-Galizia (2016); De Simone, Lester, and Markle (2020); Huizinga and Nicodème (2004); Johannesen (2014); Johannesen and Zucman (2014);

[6] Discussions with nonprofits revealed a possible explanation for this behavior, as well as for the acceptance of donor valuations and overvalued art. Nonprofits with plainly charitable purposes may often accept art donations not as a direct means of revenue, but for goodwill with a donor to secure possible further future monetary donations. In exchange, organizations may display the art prominently, or may, in the purposes of my study, accept the donor's valuation at face value.

Langenmayr and Zyska (2021); Leenders, Lejour, Rabate, and van't Riet (2023); and Omartian (2017)). I contribute to this literature, showing that donations of art are substantially driven by tax avoidance.

Considering traditional models of tax compliance in the framework of Allingham and Sandmo (1972), Kleven, Kreiner, and Saez (2016), Kleven, Knudsen, Kreiner, Pedersen, and Saez (2011), among others, I can consider an informal model in a setting with nonprofits rather than individuals. Nonprofits respond to an increase in audit probability by reducing their evasion in equilibrium by correctly filing tax forms. I find that this increased compliance has further effects on the governance of a nonprofit. Nonprofits respond to an audit flag by spending more on compliance in the subsequent year, with a 1.2% increase in the establishment of an audit committee, and an increase in accounting and legal fees, evaluated at the yearly mean, by $2,450 and $2,700 respectively.

However, nonprofits can also respond to an increase in audit probability by attempting to reduce the probability directly. Kleven, et al. (2011), in documenting the importance of third-party reporting in tax compliance, show that individuals who can self-report income are more likely to evade taxes. Nonprofits with assets below $10 million may paper file their 990s, while those above this threshold are required to e-file. Nonprofits are known to bunch around these filing thresholds (Marx (2018)). I find that nonprofits above this threshold who are required to e-file respond to audit flags by correctly filing. However, those that are below this threshold are no more likely to correctly file correctly file, and paper filers are less likely to disclose art, value it, or revalue it on subsequent filings. Paper filings, then, may act as a method for reducing the probability of an audit by the IRS through added delay and complication. As predicted in other settings, nonprofits which cannot reduce their audit probability evade less, while those that can, do.

The rest of my study focuses on quantifying the tax losses of these fine art donations. Nonprofits act as a convenient vehicle for those looking to maximize tax avoidance with works of art. Within the United States, nonprofits, family foundations, and trusts can be used in conjunction to avoid both estate and capital gains taxation, preserving intergenerational wealth and exacerbating income inequality (Boserup, Kopczuk, and Kreiner (2016)). I capitalize the cumulative donated value of art to a nonprofit and compare this to the tax-year-end value of art assets. This decomposition allows us to separate donated value from the true value of the art as established by the organization. This measure becomes positive when a nonprofit revalues art lower than the sum of net donations. This may occur in the year of donation, or it may occur in response to expected increase in audit probability from engaging in an audit flag. Regardless, this measure can quantify when, and by how much, nonprofits overvalue art donations compared to its true value to the organization.

Utilizing this measure to tease out overvaluation, donations of art are unconditionally overvalued 30% of the time, avoiding[7] $734M in income taxes in 2022 dollars throughout my sample period.[8] I use this binary measure to focus specifically on the case of overvaluation, and not a continuous measure that would capture all valuation types. Extrapolating the amount of overvaluation to total valued art and utilizing the rate of audit flags for nonprofits which record but do not value art, I create estimates of tax avoidance for all donations of art to nonprofits in the United States. I find $5.5B in tax loss from these donations in 2022 dollars over the 12 years of my sample. Putting these numbers in context, these tax losses can account for 0.2% of US individual income tax receipts in 2022 (U.S. Treasury, 2024). While tax avoidance of this type may not be considerable by itself in revenue terms, it acts in complement, or even substitution, with other forms of tax avoidance. To relate this sum to other methods of tax avoidance, this amount is 1.3% of IRS estimates of the 2014 to 2016 net tax gap (Internal Revenue Service, 2022), 16.7% of missing U.S. multinational tax receipts from offshoring in 2020 (Tørsløv, Wier, and Zucman, 2022, Alstadsæter et al., 2023), or 22.8% of the ongoing capital gains tax losses from U.S. offshore tax evasion (Hanlon, Maydew, and Thornock, 2015) throughout my sample.

The rest of the paper proceeds as follows. Section 2 summarizes institutional detail, methodology, and data. Section 3 presents summary results of the determinants of art donations. Section 4 examines the role of audit flags in nonprofit filing compliance, organization responses, and tax loss estimates. Section 5 concludes.

[7] I document the use of these donations and the taxes they avoid. I am agnostic as to the donor's purpose—a donation to a nonprofit organization may contain both altruistic and pecuniary motives. I therefore define tax avoidance as an outcome, rather than an intention, in this paper.

[8] Charitable donations offer income tax deductions of between 30 and 50% of adjusted gross income, though this can be carried forward over 5 years. I assume that the entirety of donations in my sample are both itemized and income tax deductible. Without linking individual tax returns to non-monetary donations on Form 8823 and nonprofits' list of donors on Schedule B, this assumption cannot be empirically tested.

## 2. Data and Institutional Setting

### *2.1 Institutional Setting and Data*

Tax-exempt organizations in the United States must file one of 4 tax forms each tax year: Form 990, 990-N, 990-EZ, and 990-PF. Unlike most tax filings, these filings are then made available to the public by the IRS, including some digitized versions of paper forms.[9] These raw data contain over 900 unique variables, which the researcher must then organize based on the tax filing year and technical specifications. I parse these filings, standardize, clean, and organize them into a panel spanning all available nonprofit organizations in the United States. This consists of data from 858,531 organizations from Tax Years 2008 to 2023. However, complete data coverage only begins in 2011 and ends in 2022. Therefore, my study is limited to the 2011 to 2022 period. The IRS notes further data issues in providing filings after 2019. In un-tabulated analyses, I confirm that these issues do not drive my results by repeating each analysis in the paper for the 2011 to 2019 sample period.

This forms the core of my sample, in which I observe the filing type (990, 990-EZ, or 990-PF. 990-N and 990-T are not presented in a machine readable format by the IRS.), total assets, total revenue, total expenditure on salaries, total value of contributions received in the tax year (what I refer to as donations to the nonprofit organization) percentage of revenue from donations, an indicator whether the organization held any art assets or donations during the tax year, total art asset book value, and the stated use of these assets.

I consider 6 categories of art donations to nonprofit organizations: art assets, fractional interest of art, historical treasures, historical antiquities, works of art, and other qualified contributions of art. I combine these categories to form total art donations for each organization in each tax year, noting that while fractional interest donations may be a notable tax avoiding method known to practitioners, I do not observe any donations of this type in my sample. For each of these categories I observe the number of donations of art received, the total value of these donations, and the valuation method used to determine this.

Valuation method, and the choice to value art by nonprofits, is determined by several FASB and IRS rules. First, following FASB ASC 958-360, nonprofits may choose not to value or capitalize collections of art. Holdings of art qualify as a collection if they "are held for public exhibition, education, or research in furtherance of public service rather than financial gain", "are protected, kept unencumbered, cared for, and preserved", and "are subject to an organizational policy that requires the use of proceeds from items that are sold to be for the acquisitions of new collection items, the direct care of existing collections, or both" (FASB ASC 958-360-20). However, some nonprofits do capitalize their collections. Nonprofits provide valuation as a function both of whether their art holdings are a collection and as a function of whether they choose to capitalize that collection.

Upon receipt of a non-cash donation, nonprofits must sign Form 8283, Noncash Charitable Contributions, if the donation is greater than $500, and provide a qualified appraisal if the claimed value is greater than $5,000.[10] These forms must be signed by both the donor and the nonprofit, indicating that mis-valuation is a joint decision. Based on my discussions with employees at nonprofits, there are often incentives for nonprofits to enable misreporting of these non-cash contributions to secure larger cash contributions in the future.

However, there are substantial penalties for misfiling. Failure to file an accurate Form 8283 is viewed by the IRS as a failure to file the correct form, potentially with intentional disregard. Failure to file penalties for nonprofits, then, can be range from $250 to $660 per return. While this penalty appears low, with an average number of donations of individual pieces of art of around 400 thousand per year in my sample, intentional disregard non-filing total penalties may rise quickly. Finally, for intentional disregard penalties, there is no maximum penalty.

From this parsing, cleaning, and organizing process I remove all firms that never file a Form 990. Unfortunately, only IRS Form 990 requires disclosure of art asset and donation values. What remains is the full sample for this study,

[9] These data may be found at https://www.irs.gov/charities-nonprofits/tax-exempt-organization-search-bulk-data-downloads. The author is unaware of the criteria for digitization of 990s and conversion into XML, but simply note that some paper filings, as indicated by IRS index files, also appear in the machine-readable data.

[10] For more detail, see the instructions for Form 8283 (https://www.irs.gov/instructions/i8283)

consisting of 5,364,313 organization-years. I document summary statistics about this unconditional sample in Table 1. I obtain data on audit statistics from the IRS Data Book (years 2011 to 2020), as well as the Charity Navigator API for information on nonprofit scores. Mayo (2023a, 2023b) provide extensive detail on these data from Charity Navigator, describing the breadth and impact of these ratings on donations to charitable organizations.

**TABLE 1. Summary Statistics**

| | Mean | Std. Dev. | 25th % | 50th % | 75th % |
|---|---|---|---|---|---|
| **Organization Characteristics** | | | | | |
| Art Filing | 0.012 | 0.107 | 0.000 | 0.000 | 0.000 |
| Art Value | 0.002 | 0.044 | 0.000 | 0.000 | 0.000 |
| Audit Trigger | 0.080 | 0.271 | 0.000 | 0.000 | 0.000 |
| Charity Nav. Rating | 0.366 | 0.482 | 0.000 | 0.000 | 1.000 |
| Charity Nav. Stars | 1.061 | 1.548 | 0.000 | 0.000 | 3.000 |
| Foreign Operated | 0.004 | 0.064 | 0.000 | 0.000 | 0.000 |
| Family Foundation | 0.342 | 0.474 | 0.000 | 0.000 | 1.000 |
| log(Total Assets) | 12.996 | 2.693 | 11.180 | 12.963 | 14.739 |
| Total Revenue (millions) | 10.203 | 241.852 | 0.008 | 0.126 | 0.950 |
| Salary Expense (millions) | 0.367 | 7.285 | 0.000 | 0.000 | 0.019 |
| Contributions/Total Revenue | 0.398 | 20.184 | 0.000 | 0.003 | 0.963 |
| **Organization Type** | | | | | |
| Private Foundation | 0.176 | 0.381 | 0.000 | 0.000 | 0.000 |
| Education | 0.061 | 0.239 | 0.000 | 0.000 | 0.000 |
| Religious | 0.014 | 0.117 | 0.000 | 0.000 | 0.000 |
| Library | 0.005 | 0.071 | 0.000 | 0.000 | 0.000 |
| Museum | 0.005 | 0.072 | 0.000 | 0.000 | 0.000 |
| Medical | 0.017 | 0.130 | 0.000 | 0.000 | 0.000 |
| Other | 0.722 | 0.448 | 0.000 | 1.000 | 1.000 |

Notes: N=5,364,313. This table presents summary statistics for the characteristics and types of nonprofit organization in our sample. All variables are defined in the Appendix.

In some tests, I condition this sample on organizations that indicate art holdings or donations. This sample includes all organizations that indicate on their Form 990 the presence of art assets or donations during that tax year. However, these filings are often left incomplete. 1.2% of my sample identifies that the organization held some kind of work of art or antiquities, yet only 17% of these filings, or 0.2% of the full sample, provide values for assets or donations within that filing. Put another way, 83% of all nonprofit organization art holdings incompletely record art holdings and donations with the IRS.

## *2.2 Methodology*

I construct measures of tax avoidance as the difference between art donation value in a tax year compared to its book value for the nonprofit organization. This approximates the value of fine art the donor claims on their tax returns, compared to the value that the nonprofit organization determines. To do this, I capitalize yearly donations of art, net of sales of these assets, capitalizing net flows into stocks of assets over the year. I then deflate these donations by the book value of art assets at the end of the tax year. Overvaluation is the sum of art assets in 2011 and donations net of sales for each year since 2011. That is, the overvaluation for nonprofit  in year  is

$$V_{i,T} = A_{i,2011} + \left( \sum_{t=2011}^{T} D_{i,t} - S_{i,t} \right) - A_{i,T}, \quad (1)$$

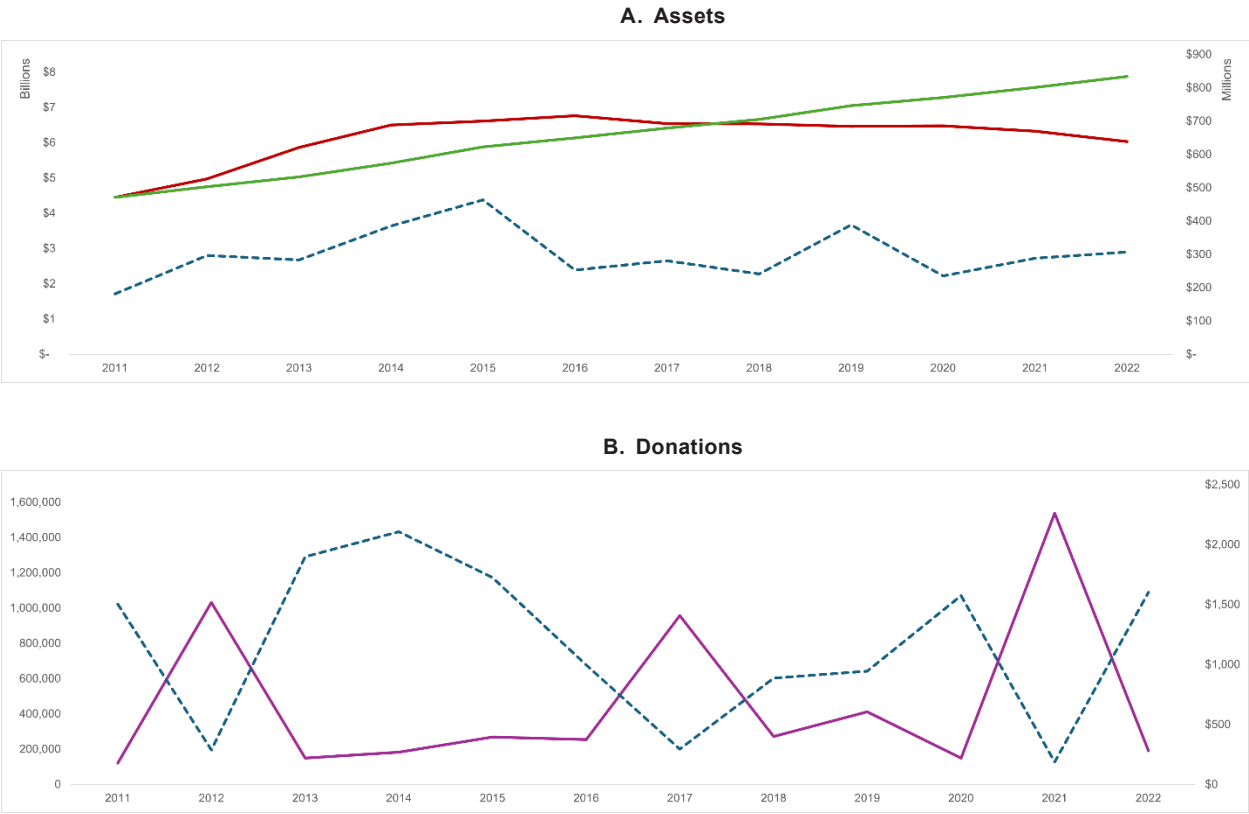
where $D_{i,t}$ is donations of art, $S_{i,t}$ is sales of art, $A_{i,t}$ is art assets for nonprofit in year $t$. This method hinges on several key assumptions. In the case that capitalized donations are less than the book value of art assets, it is possible that these assets appreciate substantially over time, especially in the case where an organization does not receive substantial donations per year. This method also assumes no growth in art value over time, simply comparing net inflows of art to stocks, which biases against my findings unless nonprofits actively write down the value of art over time below market value.

## 3. Results

### *3.1 Nonprofit Organizations and Art Holdings*

I begin my analyses by examining the determinants of art holdings and donations among nonprofits. Figure 1 presents the time series of Art Assets, Art Donations, and Cumulative Art Donations. Art assets across nonprofits in my sample are substantial, worth $6 billion in 2022. In 2022 dollars, art assets have risen steadily over my sample, from roughly $4.5B to $6B over the 12-year period. Cumulative Art Donations have risen by $1.8B more to roughly $7.8B. Donations have risen steadily, while art valuations have fluctuated over time, reaching a peak in 2016 before declining in value to 2022. Art Donations themselves have remained steady, varying between $180 to a peak of $460 million.

**FIGURE 1. Time Series of Art Assets and Donations**




Notes: Figure A presents the sum across all nonprofit organizations within our testing sample, for each tax year, of real (2022 dollar) value of Art Assets (red line, left Y axis), the Cumulative Art Donations (green line, left Y axis), and total Art Donations (blue dash line, right Y axis). Figure B presents, across all nonprofit organizations within our testing sample for each tax year, the sum of the number of Art Donations (purple line, left Y axis) and, in real 2022 dollar terms, the Average Donation Value (blue dashed line, right axis), created as the sum of all Art Donations in that tax year deflated by the total number of Art Donations in that tax year.

Yet despite the significant value of art donations each year, the average size of art donations is small. Figure 2 shows the time series average of the number of pieces of art donated in each year and the average (real) value of a donation. The number of art donations varies greatly each year, between 200,000 to 1.4 million in total number of pieces of art. With this high number of donations, the average donation value each year varies between roughly $500 and $2,000. There is great dispersion in the value of donations, but the average donation of art in my sample is quite small. On average, donations in my sample are not multimillion-dollar pieces of artwork, but in fact thousands of small donations that may be itemized in small amounts.

Examining broad impressions from my sample, I provide summary statistics in Table 1. Despite the size of assets and donations and the sheer number of pieces of art, the number of organizations that list receiving a donation is only 1.2% of organization-years in my sample. Organizations that provide value for both donations and assets consist of a much further reduced 0.2%. To summarize, $6 billion in Art Assets, $200M in Art Donations are provided by only 1/6th of organizations that hold art, leaving 5/6 nonprofits with unobserved art assets and donations.

When conditioning my sample of nonprofits to those that disclose their art holdings and donation value, there are several key disclosure issues to note. Part IV, Question 8 of a Form 990 asks "Did the organization maintain collections of works of art, historical treasures, or other similar assets?", with Question 30 asking "Did the organization receive contributions of art, historical treasures, or other similar assets, or qualified conservation contributions?". Checking of these boxes indicates the presence of art assets or donations, respectively, and requires filling out a Schedule D, with information on art assets, or Schedule M for donations. Checking one of these boxes indicates an art filing in my sample. However, as previously noted, it is apparent that, despite an obligation to file the relevant schedules after checking these boxes, only 17% do so. These organizations claiming art assets take advantage of FASB ASC 958 and do not report these assets on their balance sheet. Organizations that identify the need to file a Schedule M but do not identify these donations in their reported revenue avoid filing this form. It is clear, then, that the choice of valuation of art assets and donations for nonprofits is an issue of filing completeness. Should these nonprofits experience a shock in their need to thoroughly file their Form 990s, provided valuations would increase.

Unconditionally, nonprofit organizations hold a substantial amount of assets, worth $16.8 trillion in 2022. In context, these assets make it the third largest industry in the United States in the Fama-French 49 industry classification. In revenue terms, nonprofits have an average total revenue of $10.2 million in my sample. In 2022, after removing donations, this leads to a median ROA of 4.8%, which, setting aside its unique features, would make nonprofits more profitable than 20% of public industries. In Appendix Table A4, I utilize 3-character NTEE classifications, while in my primary analyses I focus on several broad, exclusive categories of nonprofits. I do so to provide ease of intuitive understanding of what types of nonprofit organizations are involved in my analyses, rather than granular, and potentially difficult to interpret, classifications. These categories include private foundations, educational institutions, religious organizations, libraries, museums, medical facilities, and all others. I focus on these categories for organization types as they both provide a broad breakdown of general charitable categories separate from simple public assistance types and these categories are derived directly from Form 990 filings themselves, and therefore more tightly align with nonprofit selected disclosure choices. Comparisons of the Art Assets of NTEE and organization types is shown in Appendix Figures A1 and A2. Inclusive among these groups are several other categories of organizations. Among these are family foundations, which I construct based on data from Form 990 directors, substantial contributors, shareholder managers, and contributing managers,[11] and foreign operated organizations, which I identify as having an address for the principal officer or an operating address listed as outside the U.S. In the case of missing data, I assume the organization is U.S.-based. I also collect data from Charity Navigator, a private charity rating service that rates the quality of a nonprofit. Previous work by Mayo (2023a, 2023b) has shown the importance of Charity Navigator in influencing donations to nonprofits. 36.6% of nonprofits in the sample are rated by Charity Navigator, with an average rating of 1.06 stars out of 4.[12] Finally, I identify nonprofits that engage in behavior that is a likely flag for a possible IRS audit.

[11] Specifically, I identify a family foundation as an organization that has 3 or more (2 or more if one is a substantial contributor) individuals with the same last name among these types of individuals, an individual whose last name is part of the name of the organization, or if the organization has the words "family" and "foundation" as part of its name.

[12] Missing data are imputed as zero stars to not artificially limit the sample. I include a rating indicator dummy to correct for this missing data issue, differentiating between zero from missing ratings and zero-star ratings. The average Charity Navigator star rating of firms that are rated is 2.899.

Moving from summary statistics to comparing organizations conditional on art filings, organizations that accept art donations are substantially different from those that do not. Table 2 presents statistics comparing organization types for those that accept art donations, value them, or revalue them and those that do not. Organizations that accept art are often associated with typical charitable activities. Among these, art accepting nonprofits are more likely to be educational institutions, libraries, and museums. However, in valuation and re-valuation decisions, only educational and religious institutions remain the same sign across all 3 categories. I investigate further the role in organization type and how organizations that may be more likely to accept art donations may also be less likely to value them.

**TABLE 2. Distribution of Organization Types by Art-Related Activities**

| Organization Type | With | Without | Difference |
|---|---|---|---|
| | | **A. Art Filing** | |
| Private Foundation | 0.021 | 0.178 | -0.157 |
| Education | 0.198 | 0.059 | 0.139 |
| Religious | 0.002 | 0.014 | -0.012 |
| Library | 0.021 | 0.005 | 0.016 |
| Museum | 0.160 | 0.003 | 0.157 |
| Medical | 0.025 | 0.017 | 0.008 |
| Other | 0.573 | 0.724 | -0.150 |
| | | **B. Art Value** | |
| Private Foundation | 0.012 | 0.022 | -0.010 |
| Education | 0.487 | 0.145 | 0.342 |
| Religious | 0.001 | 0.002 | -0.001 |
| Library | 0.003 | 0.024 | -0.021 |
| Museum | 0.097 | 0.172 | -0.075 |
| Medical | 0.037 | 0.023 | 0.014 |
| Other | 0.363 | 0.612 | -0.250 |
| | | **C. Overvalued Donation** | |
| Private Foundation | 0.016 | 0.011 | 0.005 |
| Education | 0.521 | 0.472 | 0.049 |
| Religious | 0.000 | 0.001 | -0.001 |
| Library | 0.007 | 0.002 | 0.004 |
| Museum | 0.091 | 0.099 | -0.008 |
| Medical | 0.034 | 0.039 | -0.005 |
| Other | 0.331 | 0.376 | -0.045 |

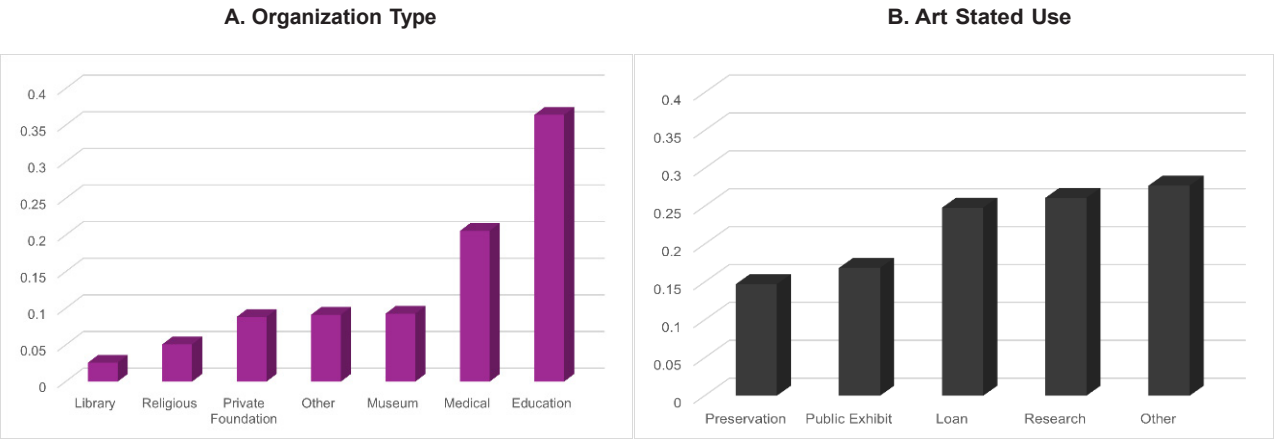
Notes: This table presents differences in pooled average organization type and characteristic measures between nonprofits with Art Filings, Art Value, or Overvalued Art versus those without. The two-sample t-tests for the differences are presented. All variables are defined in the Appendix.

### *3.2 Valuation Choice, Organization Type, and Stated Use*

Organizations that choose to hold collections of art may be, intuitively, the most likely to accept art donations. However, museums and other organizations that hold collections may also be the least likely to capitalize these collections, providing values on their balance sheet. This relationship is explained by a long-standing tradition of not capitalizing collections as the ethical standard among museums. Organizations like the Association of Art Museum Directors, the American Alliance of Museums, International Council of Museums, and the American Association for State and Local History all state that collections should not be capitalized. Other organization types may also follow this behavior, as FASB standards not requiring capitalization of collections extends not just to museums but to all nonprofits with standing collections.

The choice to value art assets and donations is likely determined by whether the nonprofit typically offers collections of art as part of its activities. Panel A of Figure 2 shows the percentage of nonprofits that provide valuations for art assets by organization type. Organization types that are typically associated with collections are much less likely to value art than those that are not. Among these, libraries, religious organizations, museums, private foundations, and other organizations are much less likely to value art, at a rate of 2.5%, 5% and 9% for the remainder, respectively. Meanwhile, organizations like medical and education institutions, which have a stated goal often differing from the presentation of a collection of art, are much more likely to do so, at a rate of 21% and 36%, respectively.

**FIGURE 2. Valuation Choice by Organization Type and by Art Stated Use**

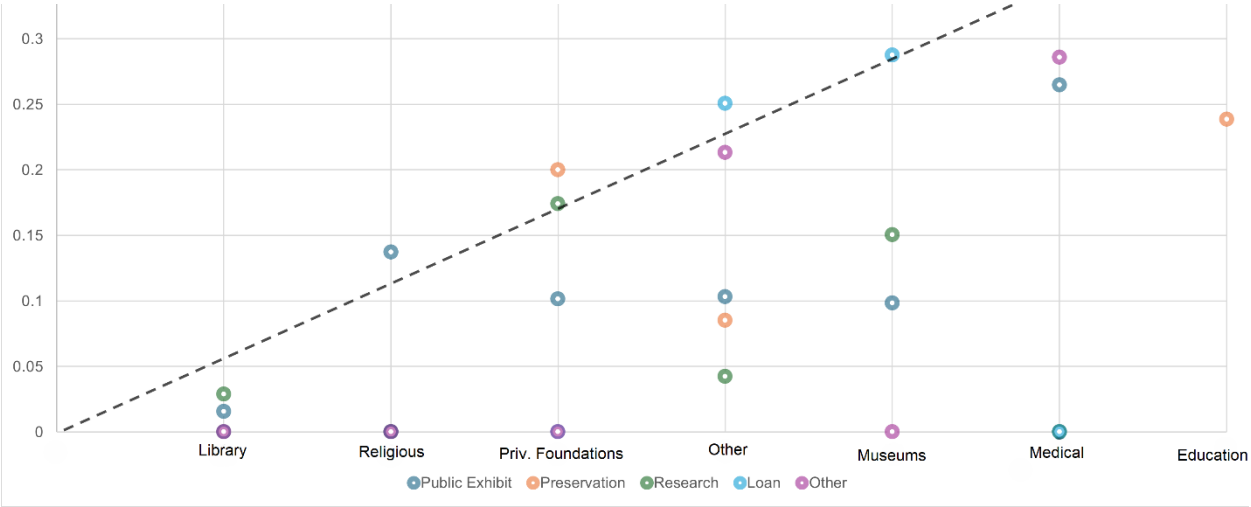



Notes: Figure A presents the average percentage of organization-years across my sample that provide valuations for art assets and contributions by type of nonprofit organization. Figure B presents the same average percentage by the organization's stated use of art assets.

While organizations themselves may be more likely to record art as a collection and subsequently choose not to capitalize it, organizations must identify the stated usage of their art holdings. The categories for the stated use of art among nonprofits include preservation, public exhibit, research, loan and other usage. Figure 2 Panel B displays the percentage of nonprofits that provide valuations by the stated usage of their holdings. Categories associated with collections are less likely to capitalize art assets. Art used for public exhibit (16%) and preservation (14%) are least likely to be valued, while that held for research (26%), loan (25%), or other (28%) purposes is more likely to be valued.

Across both organization types and stated use categories, there is a negative relationship between types and categories associated with art collections and the likelihood of valuing art. However, these two categories are not perfectly determined by the other. Examining stated use categories by organization types, there is substantial variation in the likelihood of valuing art. I present these results in Figure 3. Across organizations, which are ordered by those least likely to value art to the greatest from Figure 2 Panel A, those that use art for public exhibit and preservation are, generally, less likely to value art than those that hold art for the purpose of research, loan, or other or unknown purposes.

**FIGURE 3. Impact of Organization Type and Art Stated Use on Valuation Choice**




Notes: This figure presents the average percentage of organization-years across my sample that chose to value art assets and contributions, grouped by the stated use of art and by the organization type. The X axis presents organization type, with line 1 indicating Libraries, line 2 indicating Religious organizations, line 3 indicating Private Foundations, line 4 indicating Other organizations, line 5 indicating Museums, line 6 indicating Medical organizations, and line 7 indication Educational organizations.

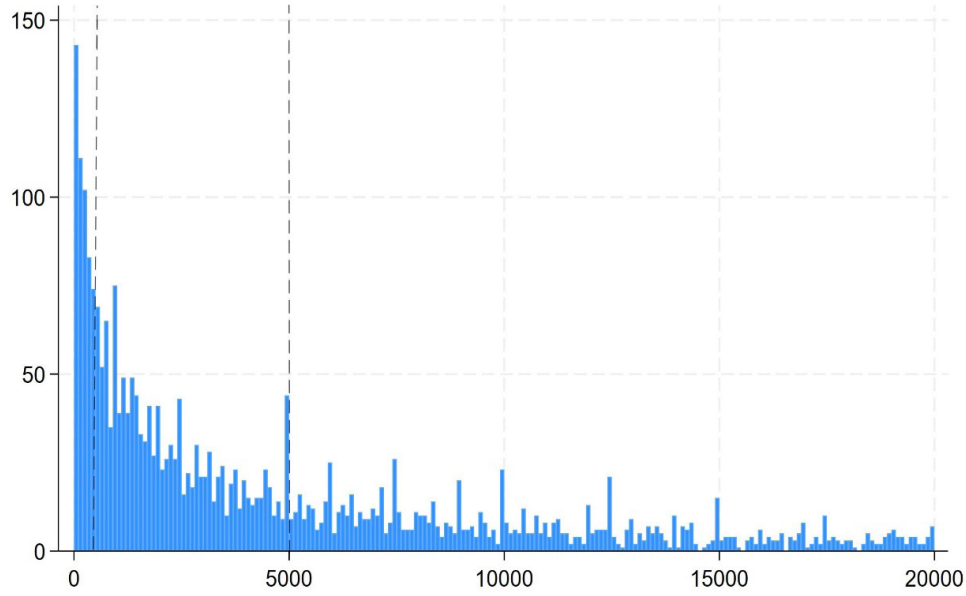
Organization types associated with owning collections, and those that choose to use art within FASB parameters that allows for their designation as a collection are both, incrementally, associated with choosing not to value art. Taken together, I interpret this relationship as primarily determined by the role of collections driving the choice of art valuation, and the decision whether to value art is determined substantially by the likelihood of an organization recognizing its art as a collection.

### *3.3 Donation Value, Qualified Appraisal, and Donation Valuation Methods*

Next, conditional on providing valuations, re-valuing donations is also driven by choices made by nonprofits. However, these decisions, unlike those to value art, are made in conjunction with an art donor. Non-cash contributions of art to nonprofits is required to be valued by a qualified appraisal when claiming an income tax deduction for the donor of over $5,000. This appraisal must follow IRS rules and come attached to the Form 8283 that a donor files to claim their charitable contribution's income tax deduction. This valuation threshold creates an opportunity, where donations claimed for less than $5,000 do not require third party valuation, while those over this threshold do. I expect, following Kleven, et al. (2011) on tax reporting and evasion, third party reporting should greatly increase valuation accuracy for art donations. Downward re-valuation of these donations by nonprofits reveals tax avoiding contributions by donors. Therefore, the choice to revalue art by nonprofits should be more common for donations under $5,000 and decrease for those donations over this amount.

Nonprofits do not directly report individual contribution amounts for art, but it is possible to approximate this for each nonprofit year by calculating the average donation value from the number of pieces divided by the total value of art donations. Using this estimate, in Figure 4 I plot the frequency of average donation values for all nonprofit years in my sample. As expected, if donors work to avoid third-party qualified appraisal, then there is a clear discontinuity at the $5,000 average donation value point, with large bunching of donations just below this threshold. This bunching indicates that donors seek to avoid third party verification of art donation values the claim on their taxes.

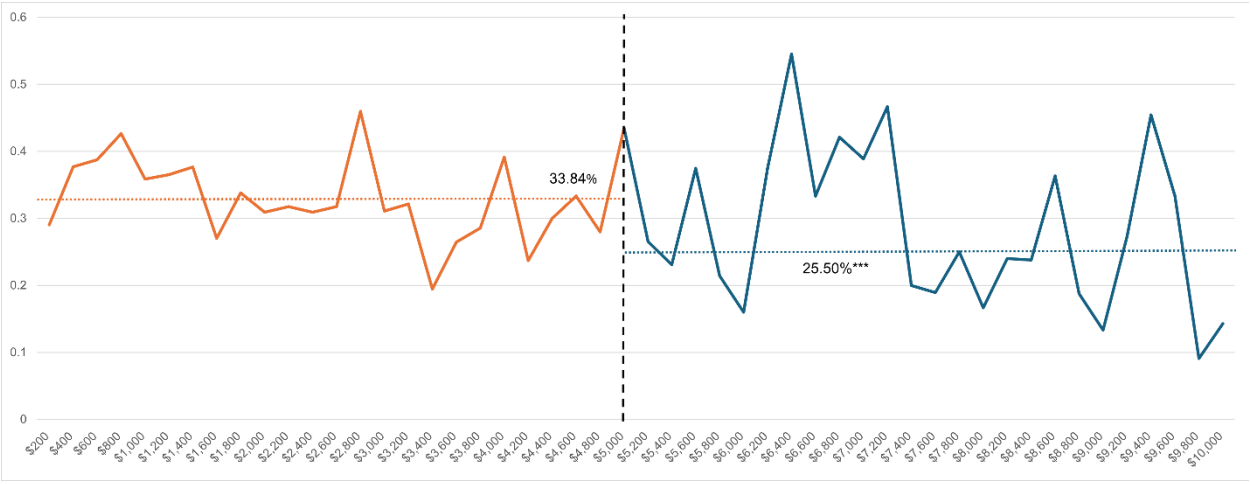
**FIGURE 4. Frequency of Average Donation Values**




Notes: This figure presents the frequency of organization-years in the sample by average donation value for each piece of art donated to that organization. The X axis is in nominal US dollars, and the Y axis represents the number of organization-years. The first dotted line is located at $500, the mandatory reporting requirement threshold for Form 8283 Section A, while the second dotted line is located at $5,000, which is the mandatory reporting requirement threshold form Form 8283 Section B, that also requires an attached qualified appraisal for art contributions.

Avoiding third party appraisal does not necessarily indicate illicit behavior on its own. Nonprofit organizations may value their net art donations differently as book assets, and re-valuing these donations reveals a wedge between the value of these donations as booked-as-income tax deductions by donors versus what they are worth to the nonprofit. Donations that are specifically written down in value are referred to as "overvalued" throughout. To observe the effect of the appraisal threshold on overvalued art donations, I plot the average overvaluation amount by average donation values in $100 bins in Figure 5.

**FIGURE 5. Overvaluation and Donation Appraisal Threshold**

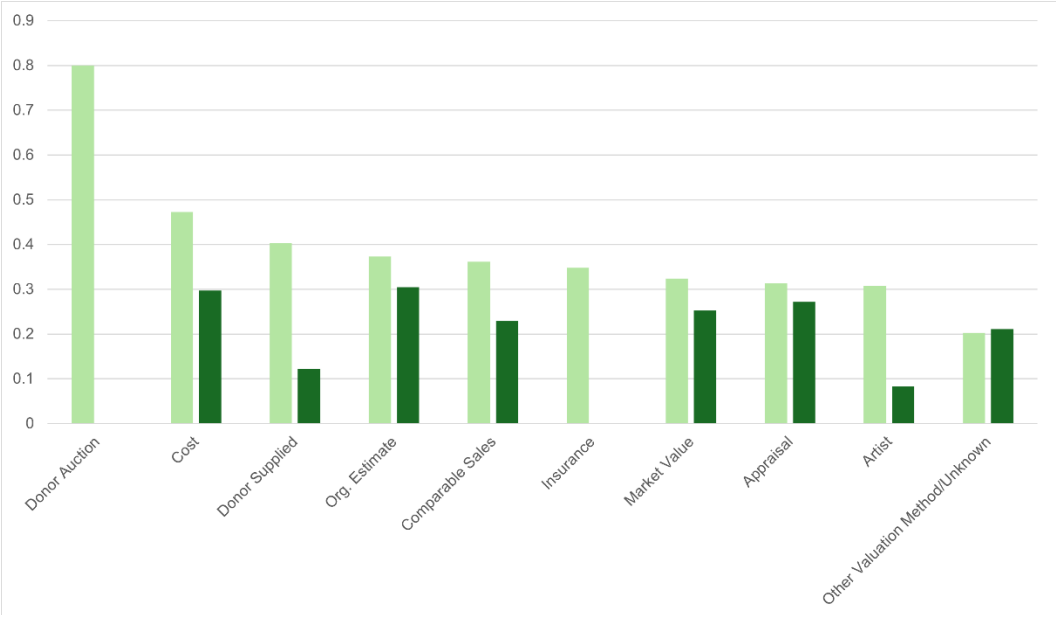



Notes: This figure presents the average percentage of overvalued art donations in the sample (Y axis) by the average donation value (X axis) across all nonprofit organizations within my sample (solid orange and blue lines). The horizontal orange dotted line indicates the average percentage of overvalued art donations for all average donation values less than $5,000 (indicated by the black vertical dashed line), while the horizontal blue dotted line indicates this same figure for all average donation values greater than $5,000 in the sample.

Donations below the appraisal threshold are significantly more likely to be overvalued, at a rate of roughly 34 percentage points of these donations compared to 26 percentage points above it, a difference of 8 percentage points Donations are more likely to bunch below the required appraisal threshold, and when they do, they are also more likely to be overvalued. Similar to the stated use of art collections for nonprofit organizations, Schedule M, Part I, column (d) of a Form 990 requires organizations to identify the "method of determining noncash contribution amounts". Using keywords, I break out these free form responses into 10 different, exclusive categories. The categories consist of auction (which often describes the donor's quoted value amount and not necessarily a completed sale), comparable sale, cost (which includes both de minimis values as well as the cost of acquisition), donor supplied, organization estimate, FMV (which simply identifies a reporting of "market value" by the organization, often without exposition), insurance, appraisal, artist, and finally other/unknown (which includes nonsense fields as well as donations with no supplied valuation method).

With non-cash donations above $5,000 requiring a qualified appraisal, and donation valuation methods describing these methods listed on the nonprofit's Form 990, there may be heterogeneity between these methods and the appraisal threshold. Investigating this relationship, I organize the average overvaluation likelihood by donation valuation methods above and below the qualified appraisal threshold in Figure 6. Consistent with the results from Figure 5, overvaluation of art donations is greater across all donation valuation methods below the appraisal threshold, except in the unknown or other valuation category. Further, donation valuation methods more likely to be supplied by the donor themselves, rather than the organization, are more likely to be overvalued. Those taken from a donor's auction or supplied by the donor themselves are among the top three categories for highest likelihood of overvaluation, conditional on average donation values below the appraisal threshold, while those supplied by the artist or appraisal are among the three least likely. Finally, while there is significant differences across valuation types below the appraisal threshold, varying between overvaluation likelihoods of 80 to 20 percentage points, above the appraisal threshold there is significantly less variation—between 30 and 10 percentage points

**FIGURE 6. Overvaluation, Donation Valuation Method, and Donation Appraisal Threshold**




Notes: This figure presents the average percentage of overvalued art donations by donation valuation method (X axis) and grouped by average donation value below the mandatory appraisal threshold (under $5,000) and by average donation value above this threshold.

Overvaluation of art donations is revealed, in part, by the presence of required appraisal thresholds based on donation value. There is significant bunching behavior below this threshold, indicating donors attempting to select into this group. Beyond this bunching and avoidance of third-party reporting issues, donor-provided valuation methods further increase

overvaluation. Overall, qualified appraisal thresholds for non-cash contributions allows for a substantial wedge between donations overvalued by their donors and their true value, using art donations to avoid income taxes. However, there are further issues with disclosure that reveal tax avoidance.

## 4. Audit Flags

Audit flags are nonprofit behaviors that increase the probability of triggering an IRS audit, which have been identified by firms specializing in nonprofit assistance (Carr, Riggs, and Ingram (2023); Foundation Group (2023)). These categories include: 1) a diversion of assets, 2) acknowledging prohibited political activity, 3) unrelated business income, 4) excess benefit transactions or loans to disqualified persons, 5) excess compensation, 6) foreign grant activity, and 7) income and expense discrepancies from fundraising events. Eight% of organization-year observations are flagged with one of these practices.

Much of the prior descriptive evidence on art filings, holdings, and overvaluation is conditioned on the unobserved propensity for nonprofit organizations to accurately and thoroughly file their Form 990 and attached schedules. In initial tests, however, I document a substantial amount of evidence that these filings are not complete or accurate. This includes obvious examples, like the lack of documenting donation valuation methods or the stated purpose of a collection, and it extends to the increase in size and revenue with disclosure accuracy, a stylized fact common in other settings, like publicly listed, for-profit firms (e.g., Engel, Hayes, and Wang (2007)). One piece of evidence that persists across all types of disclosure, however, is the increase in the propensity for an audit flag.

Audit flags, by definition, should increase the probability of an IRS audit for a nonprofit. However, nonprofit audits are exceedingly rare, with an audit rate of 0.2% in 2022. With a low base rate, any increase in probability, no matter how small, may have outsized effects on nonprofit behavior. Canonical tax evasion models, like Allingham and Sandmo (1972), focus on individuals who decide whether to report or evade based on an observed audit probability. In this setting, nonprofits may report or comply with tax filings. This has the potential to carry a penalty like the revocation of nonprofit status, but this penalty is likely to be low in the case of misfiling tax forms rather than outright evasion, which can be determined through both the tax rate in the original model as well as the fine factor. In this model, the nonprofit is also affected by the probability of an audit, and in a sense, the probability of detection.

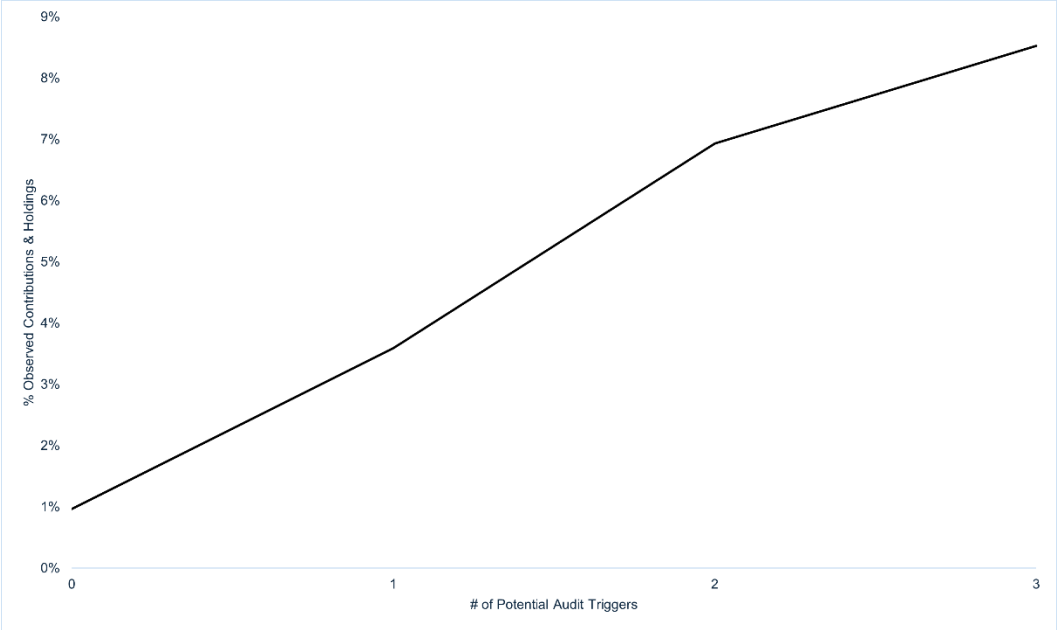
I can derive several lines of analysis from this. With a low probability of audit, filing compliance is likely to be empirically very low. Kleven et al. (2011) has documented that, despite low audit rates, third party reporting is a mitigating factor for widespread evasion by individuals. But non-monetary charitable gifts are self-reported by individuals on Form 8283, and unlikely to be valued on Schedule M or D by the nonprofit. Donations of art to nonprofits take advantage of tax savings that are difficult to value correctly in the best circumstances, and empirically a free-for-all between a donor and a willing nonprofit. Theoretically, the way to improve both the use of art in tax evasion as well as tax compliance by nonprofits is to more effectively target audits, or to raise the audit rate across the board.

Second, should nonprofits more accurately file due to a response to an increase of audit probability, nonprofits should mechanically spend greater amounts on compliance and filing costs. This may include ongoing or one-time fees or governance structures. And finally, nonprofits may attempt to attenuate the effects of an increase in audit probability by undertaking behavior that reduces audit probability. I conduct tests on each of these hypotheses in the sections that follow.

To begin, I confirm that audit flags are indeed negatively correlated with the IRS' previous year audit rate, with a correlation of -5.45%***. This correlation is likely low due to the well-documented funding issues (Boning et al., 2023) and low priority of nonprofit audits over the course of the sample. Figure 3 demonstrates this negative correlation graphically. I examine the time series of changes in audit rates with changes in audit flag rates. Nonprofits may observe the prior tax years' audit rate and allow audit flag behavior if they judge the next tax year's rate to be sufficiently low. For the purposes of this study, while this may be reducing audits on average, the specific targeting of audits and increased audit probability for an individual organization is the variation of interest. While audit flags are associated with an increase in audit propensity, and therefore substitute in response to prior audit rates, nonprofits may not respond to this change in propensity. I examine this association in Figure 7. I show that the likelihood of disclosing an art donation or collection and the likelihood of valuing art both increase monotonically with the number of audit flags a firm encountered in the previous year. Audit

flags, then, are associated with the propensity to properly document art by nonprofits.

**FIGURE 7. Observed Valuations to Disclosure**




Notes: This figure presents the average across all nonprofit organizations within our testing sample, for each year, of the percentage of organizations with observed art contributions and holdings (art value), plotted against the number of audit flags.

What are the main drivers of audit flags and their association with audit rates? Investigating this, I decompose the components of audit flags into each category, and explore their association with art filings, disclosed art value, and overvalued art donations in the subsequent tax year. Across all categories, Unrelated Business Income and Fundraising Income and Expense Discrepancies appears to be the primary drivers of disclosure responses to potential audit flags. Other categories, including loans to disqualified persons, foreign grant activity, and political activities are either unrelated to or negatively related to increases in the disclosure of art, the valuation of art, or the acceptance of overvalued donations.

While audit flags are associated with improved disclosure in univariate tests, I conduct further analyses on their improved disclosure conditional on other known factors that I have previously examined. For most regression specifications, I use linear probability models (LPM) rather than logit or probit. In Appendix Table A5, I also demonstrate that my results remain qualitatively and quantitatively similar when using fixed effect logit estimators. I choose LPMs for several reasons. First, as these analyses are exploratory, with many binary indicators and multiple fixed effects types across specifications, using organization and year fixed effects in a logit or probit model may constrain the setting, and absorb all other variation. LPMs with fixed effects allow for easy comparison across various specifications that may or may not be valid when using other models. Second, while LPMs are not bound to the unit interval, I am not using these models to predict and am only interested in aggregate effects. Chen, Martin, and Wooldridge (2023) show that LPMs recover similar effects as nonlinear, binary choice models under multivariate normality assumptions for the covariates of the model. Finally, while LPMs introduce heteroskedasticity in their estimation, all specifications use robust or clustered standard errors to correct this issue.

I present these results in Table 4. I conduct pooled, fixed effects linear probability model (LPM) regressions using controls for total assets, revenue, salary expense, and contributions as a percentage of revenue, with year, and for art valuation and overvaluation panels, collection stated use, and donation valuation method fixed effects. Within all models but model 11, I control for an individual organization type, and in model 11 of each panel, I control for all organization types

as well as these other controls and fixed effects. Across all models in Panel A, an audit flag in a prior tax year is associated with an increased likelihood of 0.3 to 0.4% to record the presence of art holdings or donations. Unconditionally, this percentage is small but does reveal some nonprofit response to increased audit probability. In Panel B, an audit flag in the prior tax year leads to an increase of 2.9 to 3.5% in the likelihood to value art, conditional on reporting holding or receiving a donation. This roughly 3% increase in art valuations provided in response to audit flags indicates an abandonment of previous accounting practices, and a recognition of donations as part of revenues and art assets as part of total assets. Nonprofits in this circumstance are not simply re-valuing assets and donations but altering revenue recognition and total assets. Finally, conditioning on valuing art, organizations respond to an audit flag by re-valuing their art assets downward 4.6 to 4.8% of the time.

**TABLE 4. Disclosure Responses to Audit Flags**

| | (1) Art Filing | (2) Art Value | (3) Overvalued Art |
|---|---|---|---|
| Audit Flag(t-1) | 1.806*** | 2.941*** | 4.660** |
| | (0.067) | (0.836) | (2.211) |
| Family Foundation | 0.094*** | 1.752** | -1.676 |
| | (0.028) | (0.889) | (2.433) |
| Private Foundation | -0.196*** | 1.347 | 11.746 |
| | (0.019) | (2.102) | (11.725) |
| Medical | -1.926*** | 8.116** | -1.360 |
| | (0.150) | (3.167) | (7.150) |
| Educational | 2.084*** | 20.778*** | 0.547 |
| | (0.098) | (1.545) | (3.369) |
| Religious | -0.092** | 0.620 | -31.681*** |
| | (0.042) | (6.495) | (4.844) |
| Library | 3.523*** | -7.872*** | 29.732 |
| | (0.374) | (1.379) | (20.401) |
| Museum | 34.225*** | -1.080 | 1.247 |
| | (0.806) | (0.954) | (3.874) |
| Foreign Operated | 4.455*** | 4.321 | 9.306** |
| | (0.449) | (2.752) | (4.627) |
| Charity Nav. Rating | -0.263*** | 1.691 | -1.753 |
| | (0.084) | (1.491) | (7.225) |
| Charity Nav. Stars | 0.636*** | 0.407 | 2.146 |
| | (0.030) | (0.375) | (1.688) |
| log(Total Assets)(t-1) | 0.332*** | 2.184*** | -0.521 |
| | (0.007) | (0.219) | (0.742) |
| Total Revenue(t-1) | -0.440* | -1.324** | -2.327 |
| | (0.227) | (0.616) | (3.304) |
| Salary Expense(t-1) | 0.102*** | 0.003 | 0.038 |
| | (0.013) | (0.018) | (0.061) |
| Contributions/Total Revenue(t-1) | 0.024 | 239.761** | -558.022* |
| | (0.016) | (96.425) | (310.944) |
| Year F.E. | Yes | Yes | Yes |

| | (1) Art Filing | (2) Art Value | (3) Overvalued Art |
|---|---|---|---|
| Art Use & Don Val F.E. | - | Yes | Yes |
| Observations | 5,364,313 | 62,541 | 9,801 |
| R-squared | 0.090 | 0.125 | 0.028 |

Notes: This table presents panel linear probability model (LPM) regression estimates of the relation between nonprofit audit flags and art filing outcomes. In column (1), art filing outcomes are measured by the disclosure of art donations or holdings. In column (2), conditioning on disclosure, art filing outcomes are measured by the valuation of art donations and holdings. In column (3), conditioning on valuation, art filing outcomes are measured by the overvaluation of art donations and holdings. Prior year controls include log of total assets, total revenue, salary expense, and contributions/total revenue. Other controls include charity navigator stars, and dummies for charity navigator rating, family foundation, private foundation, medical organization, educational, religious, library, museum, foreign, art use, donation valuation, and year fixed effects. All variables are defined in the Appendix. Standard errors are reported in parentheses and are clustered by the organization. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively. Coefficients multiplied by 100 for readability.

I would expect that organizations known for their charitable purposes—medical, educational, religious institutions and libraries and museums—would have better compliance with IRS filing requirement than those that do not—private foundations, family foundations, and foreign operated organizations. I confirm this in Appendix C Table A1, which uses LPM models focusing on each organization type. This makes theoretical sense. Organization types and mission are important in for-profit firms, but theory has identified nonprofit organizations as particularly sensitive to governance capture by mission-oriented employees (Besley and Ghatak, 2005, Glaeser, 2002). Nonprofits that focus on charitable activity are more likely to be effectively focused toward providing this public good. At the same time, those organization types bound by general motivations will likely have weaker governance and less focused activity. For instance, organizations bound by family ties or are privately held and controlled are more likely to have employment and donation activities in line with these motives; this is less conducive to providing charitable goods and services. Therefore, I include these variables as controls, as well as previously mentioned variables.

I also conduct this analysis using a logit fixed effects model for robustness. I present these results in Appendix C Table A5. Using this model, with more restrictive organization and year-level fixed effects, yields remarkably similar results in average marginal effects. An audit flag associated with an increased likelihood of 0.18% of recording art in the next tax year, conditioning on art filing, an increased likelihood of 2.09% of valuing art, and conditioning on art valuations, an increased likelihood of 4.676% of overvalued art. Art donations, the decision whether to value those donations, and the re-valuation of those assets to market are all more accurately disclosed following a nonprofit engaging in behavior which increases the propensity for an IRS audit. Further, there are potential issues in recoding the ordinal treatment of audit flags into a binary indicator. In unconditional panel regressions, presented in Appendix C Table A6, I show that recoding this count variable into a binary one is valid, as much of the variation in audit flags comes from the presence of one category, and little is gained from multiple. This is intuitive, as an IRS audit is increased in likelihood by any audit flag, but with low base rates, it is not substantially increased any more beyond this likelihood with the presence of other, related flagged behavior. Put differently, my estimation is valid as it focuses on extensive margin compliers only, as documented by Rose and Shem-Tov (2024).

These results imply that nonprofits that accept and value art allow it to be donated at more than its true value. As this behavior expands, at every level, when nonprofits are more likely to be audited, the true extent of this behavior is impossible to observe. Unconditionally, nonprofits' disclosure behavior is driven primarily by audit flags and their increasing theoretical audit propensity. The interplay of accepting art donations, valuing those donations, and accepting donations at inflated values is complex, and is driven by organization type and the stated use of art valuation behavior, as well as the $5,000 qualified appraisal threshold and donation valuation methods for overvaluation behavior. Across all these categories, however, these determinants have broad implications for tax avoidance by nonprofits and the non-monetary donations they accept.

### *4.1 Mechanism and Responses*

While audit flags are significantly associated with increases in disclosed art filings, valuation, and overvaluation, I should further observe a mechanical increase in compliance costs in response, as filing and valuation are not costless. For the nonprofit, the selection that drives audit flag choices is likely partly driven by increased disclosure costs. I confirm that the

increase of filings does indeed incur compliance costs. I confirm the mechanism of compliance increases in response to proper filing in Table 5. Specifically, I observe an increased likelihood of a nonprofit having an internal audit committee of 1.2%, an increase in accounting fees worth $2,450 and an increase in legal fees worth $2,700 in a given tax year. These compliance costs are important for organizations, as Marx (2018) documents nonprofits bunching at filing kinks to avoid increased disclosure requirements. Overall, I interpret these results as evidence for the relationship between audit flags and increased filing compliance spilling over into increased nonprofit costs.

**TABLE 5. OLS Estimates of Nonprofit Compliance Outcomes**

| | Audit Committee (1) | log(Accounting Fees) (2) | log(Legal Fees) (3) |
|---|---|---|---|
| Audit Flag (t-1) | 0.012*** | 0.317*** | 0.379*** |
| | (0.002) | (0.006) | (0.012) |
| Observations | 1,547,312 | 2,084,595 | 897,598 |
| R-squared | 0.085 | 0.429 | 0.297 |

Notes: This table presents OLS panel regression estimates of the relation between nonprofit audit flags and compliance outcomes. In specification (1), Audit Committee is an indicator variable set as one (Yes) for a nonprofit with an audit committee, and zero (No) otherwise. In specification (2), log(Accounting Fees) is the natural logarithm of total accounting fees for a nonprofit in that tax year. In specification (3), log(Legal Fees) is the natural logarithm of total legal fees for a nonprofit in that tax year. Controls include log(Total Assets), Total Revenue, Salary Expense, Contributions/Total Revenue, all in the prior tax year, and dummies for Family Foundation, Private Foundation, Medical Organization, Educational, Religious, Library, Museum, Foreign. Art Use, Donation Valuation, and Year fixed effects are included. All variables are defined in the Appendix. Standard errors are reported in parentheses and are clustered by the organization. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively.

While nonprofits may respond to an increased audit probability by more accurately disclosing, and enabling tax avoidance less, they actively decide to use audit flags. Therefore, they can observe their audit flags and prepare responses that may lower their audit propensity in similar amounts that audit flags raise it. I test this hypothesis with data on paper filings in addition to e-filing data used in the majority of my tests. Nonprofits must file Form 990s electronically if they have total assets over $10 million, while assets below this level allow organizations to use either paper or electronic tax filings. Past work has documented nonprofits bunching around arbitrary thresholds. Marx (2018) documents nonprofits bunching around the threshold for 990-EZ filings, while Mayo (2023a) documents nonprofits manipulating Charity Navigator star ratings to bunch at above kinks in the rating system. In un-tabulated results, I find that nonprofits do bunch around the $10M total asset threshold, however a comparison of kinks for nonprofits with audit flags in the year prior or not is neither statistically nor economically significant.

Combining audit flag and art filing, valuation, and overvaluation data from electronic filings with data on paper filings for the same nonprofits, I examine whether nonprofits below the e-filing threshold use paper filings to reduce their ex ante audit probability simultaneously with audit flag behavior. Paper filings may offer several advantages in reducing the probability of an audit. Paper filings are known to be processed slower, and processing is more costly than electronic filings. These facets may disguise audit flags from identification for a longer period or increase the cost of an audit before it begins. I present results on these tests in Table 6. In Panel A, for nonprofits above the $10M threshold, audit flags in the prior tax year are significantly associated with an increase in the disclosure of art donations, the valuation of art, and the overvaluation of art. The magnitude of these estimates is larger for this subsample than for the unconditional tests—with an increase in likelihood of 0.3% vs 7.2%, 3.0% vs 8.9%, and 4.7% vs 6.1%. However, for nonprofits that can choose between paper or electronic filings, those that choose to e-file have either insignificant or severely attenuated responses to audit flags in the prior tax year. I interpret these tests as suggesting that the electronic filing threshold for nonprofits allows many of them that may hold art assets to leave them undisclosed. Put differently, for small nonprofits, the choice to electronically file is a high-quality signal of their prior filings.

**TABLE 6. OLS Estimates of Filing Responses to Audit Flags**

**Panel A**

| Dependent Variable | Any Art Filing | | Filed Art Value () | | Overvalued Art | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Total assets | >$10M | <$10M | >$10M | <$10M | >$10M | <$10M |
| Audit Flag (t-1) | 7.204*** | -0.081 | 8.934*** | 1.409*** | 6.088*** | -1.718 |
| | (0.114) | (0.051) | (0.355) | (0.414) | (0.874) | (1.523) |
| Observations | 461,847 | 1,696,638 | 51,977 | 52,842 | 9,162 | 5,146 |
| R-squared | 0.009 | 0.000 | 0.012 | 0.000 | 0.005 | 0.000 |

**Panel B**

| Dependent Variable | Paper Filing | Any Art Filing | | Paper Filing | Any Art Value | | Paper Filing | Overvalued Art | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| Audit Flag (t-1) | 2.51*** | | | 4.53*** | | | 0.93 | | |
| | (0.11) | | | (0.62) | | | (2.52) | | |
| Paper Filing (t-1) | | -0.04 | | | -0.47 | | | -2.57** | |
| | | (0.04) | | | (0.29) | | | (1.02) | |
| Audit Flag (t-2) Paper Filing (t-1) | | | 0.13 | | | 0.76 | | | -0.89 |
| | | | (0.11) | | | (0.89) | | | (2.91) |
| Observations | 1,696,638 | | | 52,842 | | | 5,146 | | |
| R-squared | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 |

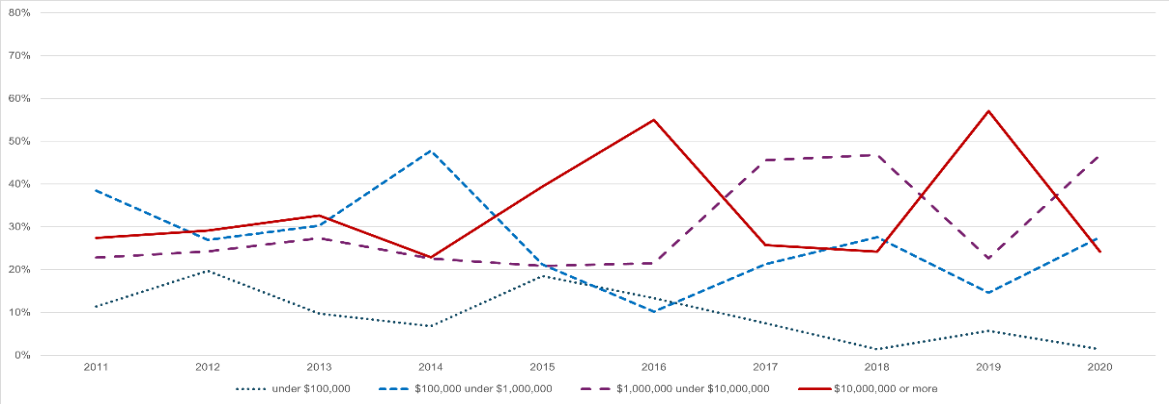
Notes: This table presents OLS estimates of the relationship between audit flags and paper return filing. Panel A, column (1) shows estimates on filing for nonprofits above $10M in total assets in the prior year and column (2) shows estimates for nonprofits below $10M in total assets. Specifications (3) and (4) present the effect of audit flags on art value, conditioning on art filings by the nonprofit, for nonprofits above and below $10M in Total Assets in the prior year. Specifications (5) and (6) present the effect of audit flags on overvalued art, conditioning on art valued by the nonprofit, for nonprofits above and below $10M in Total Assets in the prior year. In Panel B, specification (1) presents the impact of audit flags in the prior year on the likelihood of paper filing by a nonprofit, with specifications (2) and (3) presenting the impact of paper filings in the prior year on art filings and the impact of paper filings in the prior year with audit flags in the year before that. Specifications (4), (5), and (6) follow the same pattern, conditioning on art filings for nonprofits and presenting impacts for art value, with specifications (7), (8), and (9) conditioning on art value for nonprofits and presenting impacts for overvalued art. All variables are defined in the Appendix. Standard errors are reported in parentheses and are clustered by the organization. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively. Coefficients multiplied by 100 for readability.

While this evidence is suggestive, it is not confirming. Examining this paper filing sample, I show that for nonprofits with an audit flag in the prior tax year (that was electronically filed) are 2.5%, 4.5%, or, though statistically insignificant, 1% more likely to paper file in the next tax year for those in the unconditional, art filing, and art valuing samples, respectively. Nonprofits use paper filings as a method of avoiding IRS scrutiny. Next, examining art filing behavior after a paper filing, there is no positive or significant association with paper filings in the prior tax year and subsequent art disclosures, valuations, or re-valuations. Therefore, when a nonprofit files a paper filing after an audit flag occurs, there is no response to paper filings by more accurately disclosing, valuing, or re-valuing art in future electronic filings. Focusing on the exact relationship where audit flags lead to paper filing responses which are then followed by a return to electronic filings, I find no significant association with an increase in art disclosure, valuation, or overvaluation. Together, these results indicate that, after a paper filing, the increase in expected audit probability dissipates, and no future electronic filings need to be thorough. In summary, nonprofits below the electronic filing threshold, when faced with the increased audit probability brought on by an audit flag, may respond by either more accurately filing and disclosing, valuing, and re-valuing art, or may respond by paper filing, which appears to ex post decrease the nonprofit's expected audit probability. Relating this back to an informal theoretical framework, some nonprofits engage in behavior that increases their audit probability in conjunction with behavior that further decreases it when the opportunity is available, while others respond to an increase in audit probability by filing more accurately, revealing tax avoidance.

### *4.2 Tax Loss Estimates*

Finally, I conclude my analyses by conducting some back-of-the-envelope calculations of tax losses due to the re-valuation of art donations I observe. I collect data from the IRS Statistics of Income and Data Book on donations of non-monetary gifts to nonprofit organizations. I first present, in Figure 8, a summary of the fraction of art donations by AGI group over my sample period. Most art donations to nonprofits come from the highest end of the income distribution, with incomes over $10 million making up between 30 to 50% of the fraction of all donations.

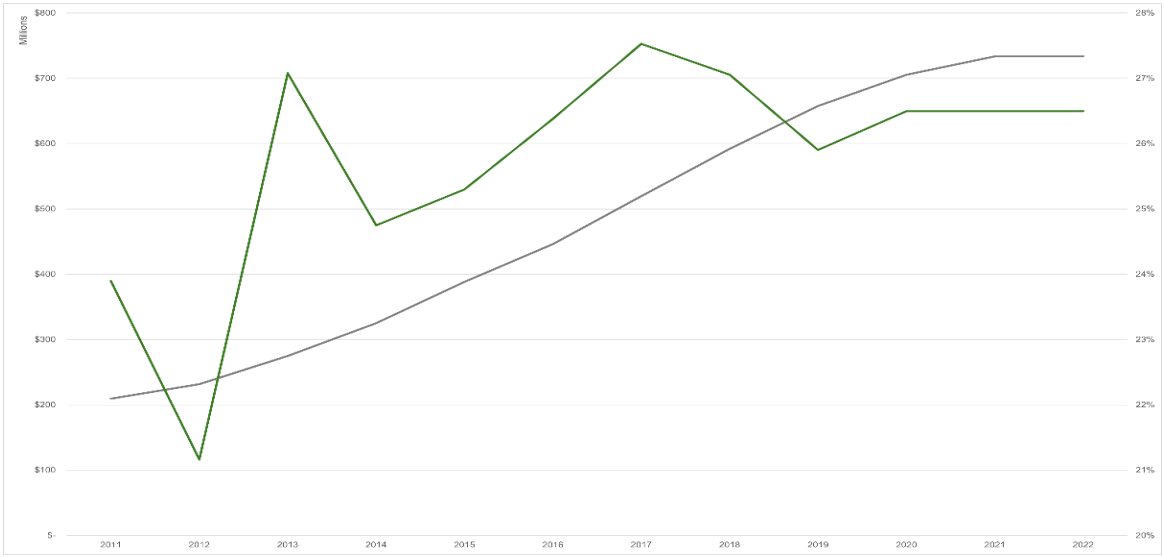
**FIGURE 8. Time Series of Art Donation Fraction by AGI**




Notes: This figure presents the fraction of art donations each tax year by AGI, taken from the IRS' Statistics of Income data. Donations are collected into four AGI buckets—$10M or more, $1M to $10M, $100K to $1M, and under $100K. Data coverage for these statistics does not extend beyond 2020.

This explains the weighted average effective income tax rates I calculate in Figure 9, which vary between 21 and 28 over the course of the sample. I calculate these rates by taking the effective tax rate (ETR) for each AGI bucket each year and weighting this ETR by the percentage of art donations given by that AGI bucket in that year.

**FIGURE 9. Time Series of Real Tax Loss and Weighted Average ETR**




Notes: This figure presents estimates of the time-series of real tax loss for overvalued donations by organizations that list art value using estimated weighted average effective tax rates (green line), with numbers in real 2022 dollars on the left axis. Charted on the right axis are estimates of weighted average effective tax rates across all art donating individuals, weighted by art donation amount per AGI group (grey line).

Unconditional tax loss estimates are significant, totaling roughly $733 million in 2022 dollars across my sample. I present these results in Table 7. I generate these estimates by taking the total amount of overvaluation from a nonprofit's write-down-of-art asset value for each year, multiplied by the weighted average ETR for that year. I also extrapolate tax losses from the sample of art that is valued for all art filings, using the ratio of written down value to the rate of audit flags for the number of valued art organizations as a baseline. Using this ratio, I project a written-down value of art for all art filings based on the rate of audit flags and number of nonprofits for that sample. This exercise, intended to provide a conservative estimate, is meant to capture the selection differences between the two groups, as the rate of audit flags is significantly different, as well as the amount of hidden assets. As I have previously indicated, many nonprofits do not recognize donations in their revenue or value art assets within their total assets by accounting convention. However, these organizations do hold art. Therefore, under the very strong—though conservative in estimate magnitude—assumption that external IRS audit probability is the primary factor driving disclosure, I estimate tax losses worth $5.5 billion in 2022 dollars across the 12 years of my sample.

**TABLE 7. Estimates of Tax Avoidance**

| | Organizations with Art Value | | | Organizations with Art Filing | | |
|---|---|---|---|---|---|---|
| | Audit Flag Rate (1) | Write down Value (2) | Estimated Tax Loss (3) | Audit Flag Rate (4) | Predicted Write down Value (5) | Predicted Tax Loss (6) |
| 2011 | 8.56% | $876.30 | $209.40 | 6.74% | $5,218.91 | $1,247.13 |
| 2012 | 8.16% | $105.69 | $22.37 | 6.13% | $834.70 | $176.66 |
| 2013 | 16.48% | $159.22 | $43.11 | 12.36% | $1,582.60 | $428.53 |
| 2014 | 47.54% | $203.05 | $50.25 | 28.73% | $1,462.59 | $361.99 |
| 2015 | 47.30% | $249.47 | $63.11 | 29.43% | $1,876.16 | $474.65 |
| 2016 | 48.30% | $219.37 | $57.87 | 29.29% | $1,795.03 | $473.52 |
| 2017 | 49.00% | $266.92 | $73.47 | 29.00% | $2,022.97 | $556.85 |
| 2018 | 51.86% | $267.93 | $72.49 | 31.27% | $2,078.04 | $562.21 |
| 2019 | 48.49% | $253.13 | $65.57 | 28.34% | $2,222.06 | $575.56 |
| 2020 | 47.99% | $180.14 | $47.73 | 26.26% | $1,587.36 | $420.59 |
| 2021 | 48.65% | $106.00 | $28.09 | 25.15% | $946.41 | $250.76 |
| 2022 | 48.78% | $0.34 | $0.09 | 27.41% | $3.56 | $0.94 |
| **Total** | | **$2,887.56** | **$733.56** | | **$21,630.40** | **$5,529.40** |

Notes: This table presents time-series estimates of effective income tax loss through art donations to nonprofit organizations. Audit Flag Rate is the mean audit flag across all organizations in each year. Write-down Value indicates the cumulative net art donations above art assets across all nonprofits in the respective tax year. Estimated Tax Loss uses the weighted average effective tax rate, weighted by donations given by AGI brackets, combined with Write-down Value to estimate income tax deductions for the overvalued portion of art donations. Predicted Write-down Value uses the ratio of Audit Flag Rate and number of organizations in each sample to extrapolate Write-down Value for all art filing nonprofits. Predicted Estimated Tax Loss uses the weighted average effective tax rate, weighted by donations given by AGI brackets, combined with Predicted Art Write-down Value to estimate income tax deductions for the overvalued portion of art donations. Effective tax rates are estimated using the methodology described in the Appendix. All dollar amounts are in millions of 2022 dollars.

These tax losses are similar in magnitude to 0.2% of US individual income tax receipts in 2022 (U.S. Treasury, 2024). For these back-of-the-envelope calculations, I only focus on these effective income tax itemized charitable deductions from donations. I ignore capital gains appreciation of art, or many of the sales tax and other avoidance strategies known to be used with art (*e.g.*, the Masterworks on Loan sales tax avoidance strategy, documented by New York Times (2014)). I discuss these further tax loss possibilities in Appendix B. However, my extrapolated write-down value is similar, and often larger, than the total amount of art assets I observe. Yet tax losses from these numbers are likely low. In so doing, I believe these extrapolated estimates of tax avoidance present a rough starting point, and the true number may be significantly higher or lower than this estimate. Compared to other tax avoidance strategies, like the use of offshore tax havens, real estate, or tax avoidance by US multinationals, this amount is 0.36% of U.S. wealth held in offshore tax havens in 2022 (Al-

stadsæter, et al. (2023)). At a reasonable range of plausible returns to these assets, this is a significant fraction of the capital income from offshore accounts. Similarly, when comparing these numbers to missing U.S. multinational tax receipts from offshoring in 2020, this estimate is 17% of these tax losses as well (Alstadsæter, et al. (2023)). While the academic literature has primarily focused on international tax avoidance methods, a growth in audits that help differentiate nonprofits' role in aiding tax avoidance versus their ostensible provision of public goods may be quite impactful.

## 5. Conclusion

Nonprofits in the United States spend a substantial amount on the private provision of public goods. These organizations, which assist in many charitable causes, are an important aspect of the U.S. economy. However, the tax incentives of charitable donations also allow nonprofits to aid individuals in tax avoidance. Focusing on the role of art donations to nonprofits, I show that nonprofits accept art donations both following their charitable mission as well as aiding in tax avoidance. These donations are substantial, with the value of art held by nonprofits worth $6 billion in 2022. While these assets are substantial, only 17% of organizations that accept art donations or hold art value it. Much of this low percentage of disclosure is driven by the choice to not capitalize collections of art. This behavior is associated with specific organization types, like museums and libraries, as well as the stated use of art within organizations, e.g. for public exhibit or preservation. Accepting overvalued art donations is also related to donation values below the value threshold requiring qualified appraisal attached to a donor's income tax deduction form. This avoidance of required third party valuation leads to greater acceptance of overvalued donations, particularly in conjunction with donation valuation methods that are associated with a donor provided valuation.

Finally, even the observation of art is driven by a failure to completely file tax forms. An increase in audit probability, driven by nonprofit behavior, significantly impacts the likelihood of nonprofits to identify donations, value donations, and revalue these donations to their true value. This revaluation reveals the amount of overvaluation that occurs with art donations, revealing significant tax avoidance. Donations to nonprofits in 2022 were worth 9.88% of the revenue of the U.S. government. However, juxtaposed against this significant contribution to social welfare is my estimate of tax loses from the donation of art to nonprofits, worth 1.3% of IRS estimates of the 2014–2016 net tax gap. Nonprofits provide both important public goods and sizeable tax loopholes.

## References


Allingham, Michael G. and Agnar Sandmo. 1972, "Income tax evasion: a theoretical analysis," Journal of Public Economics, 1:3-4, 323–338.

Ang, Yuen Yuen. 2020, "China's Gilded Age: The Paradox of Economic Boom and Vast Corruption," Cambridge University Press.

Alstadsæter, Annette, Sarah Godar, Panayiotis Nicolaides, and Gabriel Zucman. 2023, "Global Tax Evasion Report 2024," EU Tax Observatory Report.

Alstadsæter, Annette, Niels Johannesen, and Gabriel Zucman. 2018, "Who Owns the Wealth in Tax Havens? Macro Evidence and Implications for Global Inequality," Journal of Public Economics, 162: 89–100.

Andreoni, James and A. Abigail Payne. 2003, "Do Government Grants to Private Charities Crowd Out Giving or Fundraising?" The American Economic Review, 93:3, 792–812.

Andreoni, James and A. Abigail Payne. 2011, "Is Crowding Out Due Entirely to Fundraising? Evidence From a Panel of Charities," Journal of Public Economics, 93:5-6, 334–343.

Besley, Timothy and Maitreesh Ghatak. 2005, "Competition and Incentives with Motivated Agents," The American Economic Review, 95:3, 616–636.

Boning, William C., Nathaniel Hendren, Ben Sprung-Keyser and Ellen Stuart. 2023, "A Welfare Analysis of Tax Audits Across the Income Distribution," NBER Working Paper No. 31376.

Boserup, Simon H., Wojciech Kopczuk, and Claus T. Kreiner. 2016, "The Role of Bequests in Shaping Wealth Inequality: Evidence from Danish Wealth Records." American Economic Review, 106 (5): 656–61.

Carr, Riggs, and Ingram. 2023, "Six Common Nonprofit IRS Audit Triggers," https://cricpa.com/insight/six-common-nonprofit-irs-audit-triggers/.

Caruana-Galizia, Paul, and Matthew Caruana-Galizia. 2016, "Offshore financial activity and tax policy: evidence from a leaked data set." Journal of Public Policy, 36: 457–488.

Cornes, Richard and Jun-Ichi Itaya. 2010, "On the private provision or two or more public goods," Journal of Public Economic Theory, 12:2, 363–385.

De Simone, Lisa, Rebecca Lester, Kevin Markle. 2020. "Transparency and Tax Evasion: Evidence from the Foreign Account Tax Compliance Act (FATCA)." Journal of Accounting Research, 58 (1): 105–153.

Diamond, Peter. 2006, "Optimal tax treatment of private contributions for public goods with and without warm glow preferences," Journal of Public Economics, 90: 897–919.

Duquette, Nicolas J. 2016, "Do Tax Incentives Affect Charitable Contributions? Evidence from Public Charities' Reported Revenues." Journal of Public Economics 137:51–69.

Duquette, Nicolas J. 2019, "Do Share-of-Income Limits on Tax-Deductibility of Charitable Contributions Affect Giving?" Economics Letters, 174:1–4.

Engel, Ellen, Rachel M Hayes, and Xue Wang. 2007, "The Sarbanes–Oxley Act and firms' going-private decisions," Journal of Accounting and Economics, 44:116–145.

Foundation Group. 2023, "Top 10 Form 990 Audit Triggers No One Told You About," https://www.501c3.org/top-10-form-990-audit-triggers-no-one-told-you-about/.

Gee, Laura, and Jonathan Meer. 2019, "The Altruism Budget: Measuring and Encouraging Charitable Giving," NBER Working Paper No. 25938.

Glaeser, Edward L. 2002, "The Governance of Not-For-Profit Firms," NBER Working Paper No. 8921.

Glazer, Amihai and Kai A. Konrad. 1996, "A Signaling Explanation for Charity," The American Economic Review, 86:4, 1019–1028.

Gravelle, Jane G. 2015, "Tax Havens: International Tax Avoidance and Evasion". Congressional Research Service 7-5700.

Green, Sarah. 2018, "Art + Taxes = The Dirty Truth." https://www.youtube.com/watch?v=QZz2PhTQJCA

Guyton, John, Patrick Langetieg, Daniel Reck, Max Risch, and Gabriel Zucman. 2021, "Tax Evasion at the Top of the Income Distribution: Theory and Evidence." National Bureau of Economic Research Working Paper Series No. 28542.

Hanlon, Michelle, Edward L. Maydew, and Jacob R. Thornock. 2015, "Taking the Long Way Home: U.S. Tax Evasion and Offshore Investments in U.S. Equity and Debt Markets." The Journal of Finance 70: 257–287.

Hemel, Daniel J. 2022, "The United States as the Ultimate Tax Haven: Testimony Before the House Ways and Means Subcommittee on Oversight." Univ. of Chicago Public Law & Legal Theory Working Paper No. 793.

Hines, James R. Jr. 2023, "The role of trusts in taxing the rich," Oxford Review of Economic Policy, 39:3, 460–477.

Huizinga, Harry, and Gäetan Nicodème. 2004, "Are international deposits tax-driven?" Journal of Public Economics, 88: 1093–1118.

ICIJ. 2022, "Magazine spread of 'most beautiful house in America' conceals allegedly stolen Cambodian relics," International Consortium of Investigative Journalists. https://www.icij. org/investigations/pandora-papers/lindemann-cambodia-relics-altered-photo-magazine/

Internal Revenue Service. 2022, "Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2014-2016," Publication 1415 (rev. 10-2022), Washington, D.C.

Johannesen, Niels. 2014, "Tax evasion and Swiss bank deposits." Journal of Public Economics 111: 46–62.

Johannesen, Niels, Jakob Miethe and Daniel Weishaar. 2022, "Homes Incorporated: Offshore Ownership of Real Estate in the U.K," CESifo Working Paper No. 10159.

Johannesen, Niels, and Gabriel Zucman. 2014, "The End of Bank Secrecy?" American Economic Journal: Economic Policy 6: 65–91.

Kleven, Henrik, Martin B Knudsen, Claus T Kreiner, Søren Pedersen, and Emmanuel Saez. 2011, "Unwilling or Unable to Cheat? Evidence from a Tax Audit Experiment in Denmark," Econometrica, 79:3, 651-692.

Kleven, Henrik, Claus Kreiner, and Emmanuel Saez. 2016, "Why Can Modern Governments Tax So Much? An Agency Model of Firms as Fiscal Intermediaries," Economica, 83:219–246.

Langenmayr, Dominika. 2017, "Voluntary disclosure of evaded taxes—Increasing revenue, or increasing incentives to evade?" Journal of Public Economics 151: 110–125.

Langenmayr, Dominika and Lennard Zyska. 2021, "Escaping the exchange of information: Tax evasion via citizenship-by-investment". CESifo Working Paper No. 8956.

Leenders, Wouter, Arjan Lejour, Simon Rabaté, and Maarten van't Riet. 2023, "Offshore tax evasion and wealth inequality: Evidence from a tax amnesty in the Netherlands". Journal of Public Economics 217: 104785.

List, John A. 2011, "The Market for Charitable Giving." Journal of Economic Perspectives, 25:157–80.

Londoño-Vélez, Juliana and Javier Ávila-Mahecha. 2023, "Behavioral Responses to Wealth Taxation: Evidence from Colombia," Working Paper.

Marx, Benjamin. 2018, "Optimizing Policy Notches: Theory and Evidence from a Reporting Requirement for Charities," Working Paper.

Mayo, Jennifer. 2023a, "Navigating the Notches: Charity Responses to Ratings," Journal of Political Economy: Microeconomics, forthcoming.

Mayo, Jennifer. 2023b, "The Impact of Sanctioning in the Nonprofit Sector," Working Paper.

McCandrew, Clare. 2023, "The Art Basel and UBS Art Market Report 2024," https://theartmarket.artbasel.com.

Meer, Jonathan and Benjamin Priday. 2020a, "Generosity across the Income and Wealth Distributions," NBER Working Paper No. 27076.

Meer, Jonathan and Benjamin A. Priday. 2020b, "Tax Prices and Charitable Giving: Projected Changes under the 2017 TCJA," Tax Policy and The Economy, 34:113–38.

New York Times. 2014, "Buyers Find Tax Break on Art: Let It Hang Awhile in Oregon," https://www.nytimes.com/2014/04/13/business/buyers-find-tax-break-on-art-let-it-hang-awhile-in-portland.html.

Omartian, James. 2016, "Do Banks Aid and Abet Asset Concealment: Evidence from the Panama Papers." SSRN Working Paper.

Oliver, John. 2022, "Last Week Tonight: Museums." Home Box Office, 9 (24).

Ottoni-Wilhelm, Mark, Lise Vesterlund, and Huan Xie. 2017, "Why Do People Give? Testing Pure and Impure Altruism," The American Economic Review, 107:11, 3617–33.

Prendergast, Canice. 2021, "A Fascinating, Sexy, Intellectually Compelling, Unregulated Global Market." Freakonomics Radio, 484.

Rose, Evan K., and Yotam Shem-Tov. 2024, "On Recoding Ordered Treatments as Binary Indicators," NBER Working Paper No. 32234.

Saez, Emmanuel. 2004, "The optimal treatment of tax expenditures," Journal of Public Economics, 88: 2657–2684.

Saez, Emmanuel and Gabriel Zucman. 2021, "The Rise of Income and Wealth Inequality in America: Evidence from Distributional Macroeconomic Accounts", Journal of Economic Perspectives, 34(4), 3–26.

Tørsløv, Thomas, Ludvig Wier, and Gabriel Zucman. 2022, "The Missing Profits of Nations," The Review of Economic Studies, 0, 1–36.

U.S. Senate, Permanent Subcomittee on Investigations. 2020, "The Art Industry and U.S. Policies That Undermine Sanctions". https://www.hsgac.senate.gov/wp-content/uploads/imo/media/doc/2020-07-29%20PSI%20Staff%20Report%20-%20The%20Art%20Industry% 20and%20U.S.%20Policies%20that%20Undermine%20Sanctions.pdf.

U.S. Treasury. 2024, "FiscalData.Treasury.gov," https://fiscaldata.treasury.gov/americas-finance-guide/

Villanacci, Antonio, and Zenginobuz, Ünal. 2007, "On the neutrality of redistribution in a general equilibrium model with public goods," Journal of Public Economy Theory, 9:2, 183–200.

Villanacci, Antonio, and Zenginobuz, Ünal. 2012, "Subscription equilibrium with production: Nonneutrality and constrained suboptimality," Journal of Economic Theory, 147, 407–425.

Zucman, Gabriel. 2013, "The Missing Wealth of Nations: Are Europe and the US net Debtors or net Creditors?" Quarterly Journal of Economics, 128(3), 1321–1364.

Zucman, Gabriel. 2014, "Taxing Across Borders: Tracking Personal Wealth and Corporate Profits," Journal of Economic Perspectives, 28(4), 121–148.

# Appendix

## *A. Additional Figures and Tables*

**FIGURE A1. Time Series of Art Donations and Average Donation Value by NTEE**

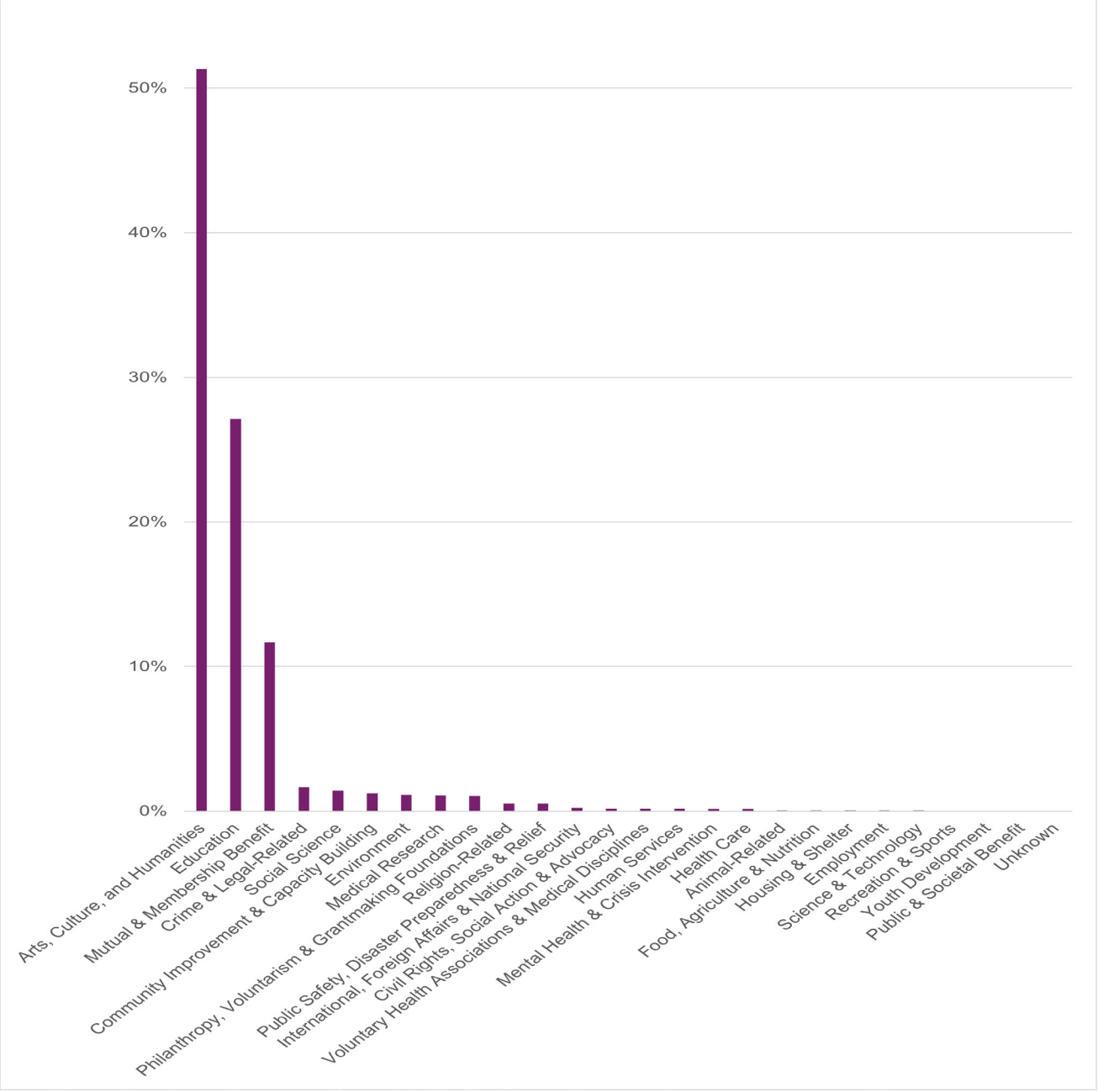


**FIGURE A2. Time Series of Art Donations and Average Donation Value by Organization Type**

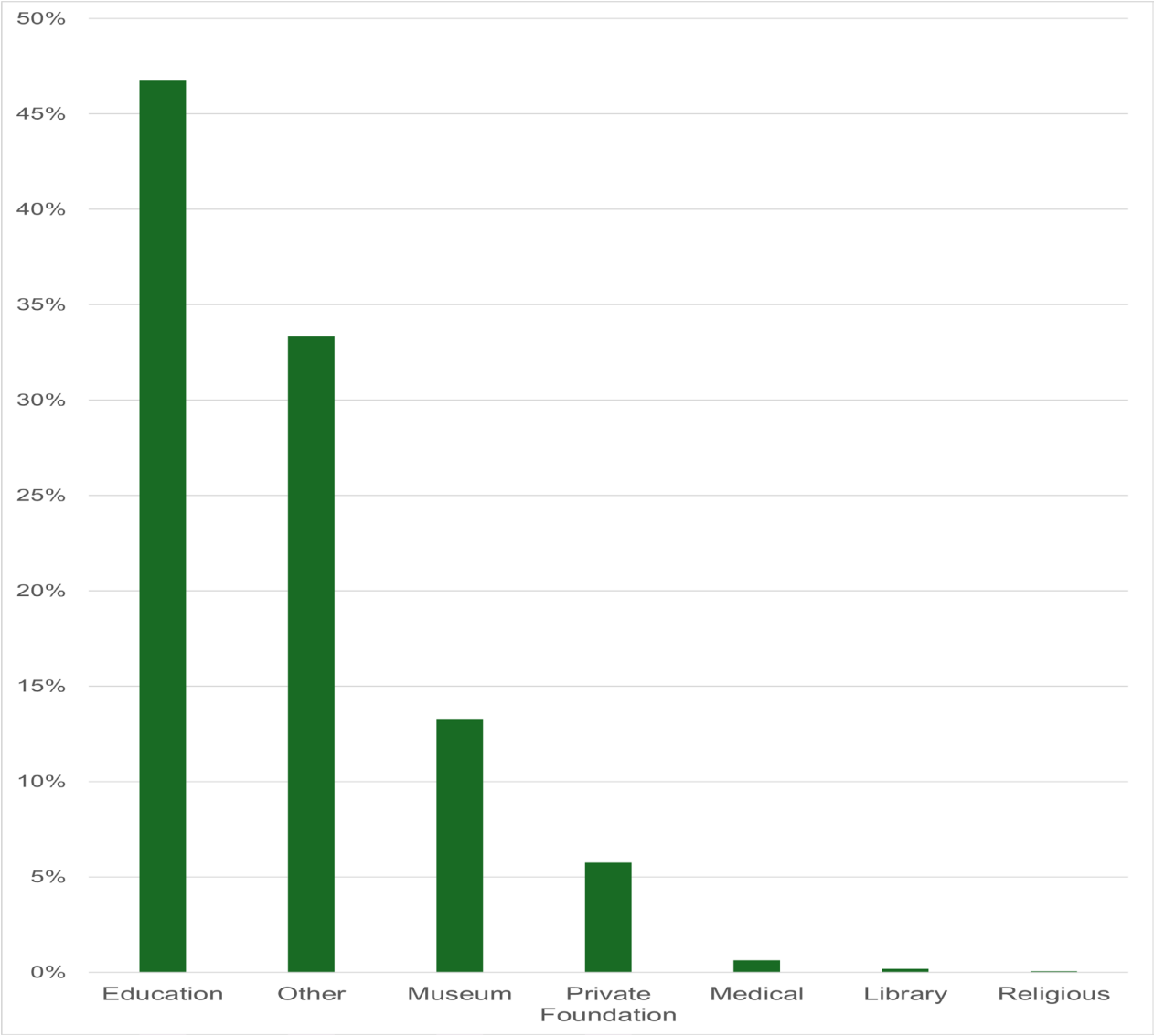


**TABLE A1. Determinants of Accepting, Valuing, and Overvaluing Art Donations: Organization Type**

A. **Art Filing (N=5,364,313)**

| | Family Foundation (1) | Private Foundation (2) | Medical (3) | Educational (4) | Religious (5) | Library (6) | Museum (7) | Foreign Operated (8) |
|---|---|---|---|---|---|---|---|---|
| Org. Type | -0.079*** | -1.090*** | -2.10*** | 2.236*** | -0.35*** | 3.424*** | 34.497*** | 5.017*** |
| | (0.028) | (0.019) | (0.154) | (0.098) | (0.038) | (0.375) | (0.812) | (0.460) |
| log(Total Assets) | 0.550*** | 0.546*** | 0.563*** | 0.535*** | 0.549*** | 0.549*** | 0.516*** | 0.543*** |
| | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) |
| Total Revenue | -0.460** | -0.464** | -0.486** | -0.435** | -0.460** | -0.458** | -0.439** | -0.511** |
| | (0.210) | (0.206) | (0.216) | (0.213) | (0.210) | (0.211) | (0.223) | (0.202) |
| Salary Expense | 0.105*** | 0.104*** | 0.111*** | 0.102*** | 0.105*** | 0.105*** | 0.107*** | 0.102*** |
| | (0.013) | (0.013) | (0.014) | (0.013) | (0.013) | (0.013) | (0.014) | (0.013) |
| Contributions / Total Revenue | 0.060 | 0.054 | 0.059 | 0.065 | 0.060 | 0.060 | 0.050 | 0.060 |
| | (0.042) | (0.037) | (0.041) | (0.045) | (0.041) | (0.041) | (0.034) | (0.041) |
| R-squared | 0.025 | 0.027 | 0.026 | 0.028 | 0.025 | 0.026 | 0.078 | 0.026 |

B. **Art Value (N=62,541)**

| | Family Foundation (1) | Private Foundation (2) | Medical (3) | Educational (4) | Religious (5) | Library (6) | Museum (7) | Foreign Operated (8) |
|---|---|---|---|---|---|---|---|---|
| Org. Type | 2.020** | -1.582 | -1.881 | 20.815*** | -3.558 | -11.14*** | -4.207*** | 6.957** |
| | (0.916) | (2.094) | (3.186) | (1.505) | (6.629) | (1.437) | (0.956) | (2.821) |
| log(Total Assets) | 4.292*** | 4.312*** | 4.328*** | 2.693*** | 4.318*** | 4.310*** | 4.282*** | 4.213*** |
| | (0.204) | (0.205) | (0.206) | (0.204) | (0.204) | (0.204) | (0.202) | (0.204) |
| Total Revenue | -1.879** | -1.895** | -1.874** | -1.104 | -1.900** | -1.931** | -1.952** | -2.130*** |
| | (0.847) | (0.859) | (0.870) | (0.725) | (0.857) | (0.848) | (0.839) | (0.772) |
| Salary Expense | 1.669 | 1.676 | 1.745 | 0.619 | 1.681 | 1.706 | 1.693 | 1.676 |
| | (2.193) | (2.210) | (2.237) | (1.941) | (2.207) | (2.192) | (2.176) | (2.050) |
| Contributions / Total Revenue | -297.29*** | -309.94*** | -312.65*** | 242.19*** | -307.88*** | -296.21*** | -254.89*** | -310.02*** |
| | (95.306) | (95.878) | (95.569) | (92.754) | (95.603) | (95.340) | (95.463) | (95.239) |
| R-squared | 0.087 | 0.086 | 0.086 | 0.120 | 0.086 | 0.088 | 0.088 | 0.087 |

C. **Overvalued Art (N=9,801)**

| | Family Foundation (1) | Private Foundation (2) | Medical (3) | Educational (4) | Religious (5) | Library (6) | Museum (7) | Foreign Operated (8) |
|---|---|---|---|---|---|---|---|---|
| Org. Type | -1.120 | 9.047 | -2.789 | 0.759 | -36.1*** | 29.851 | -0.099 | 9.267** |
| | (2.453) | (11.519) | (6.660) | (3.054) | (3.889) | (20.432) | (3.829) | (4.606) |
| log(Total Assets) | 0.534 | 0.560 | 0.517 | 0.437 | 0.492 | 0.554 | 0.515 | 0.295 |
| | (0.585) | (0.585) | (0.584) | (0.667) | (0.584) | (0.583) | (0.602) | (0.589) |
| Total Revenue | -1.251 | -1.459 | -1.021 | -1.175 | -1.220 | -1.240 | -1.235 | -1.898 |
| | (3.189) | (3.240) | (3.256) | (3.209) | (3.197) | (3.197) | (3.197) | (3.189) |
| Salary Expense | 1.574 | 1.913 | 1.313 | 1.528 | 1.560 | 1.553 | 1.573 | 2.378 |
| | (5.964) | (6.036) | (6.072) | (5.987) | (5.974) | (5.976) | (5.974) | (5.934) |
| Contributions / Total Revenue | -436.979 | -418.941 | -446.710 | -402.507 | -443.054 | -425.414 | -427.846 | -460.618 |
| | (293.770) | (294.524) | (298.150) | (299.186) | (294.349) | (292.154) | (294.578) | (293.709) |
| R-squared | 0.020 | 0.020 | 0.020 | 0.019 | 0.020 | 0.021 | 0.019 | 0.022 |

## TABLE A2. Determinants of Accepting, Valuing, and Overvaluing Art Donations: Collection Stated Use

A. **Art Value (N=62,541)**

| | Public Exhibit (1) | Preservation (2) | Research (3) | Loan (4) | Other (5) |
|---|---|---|---|---|---|
| Art Stated Use | 7.543*** | 3.991*** | 0.864 | 0.273 | 7.676*** |
| | (1.632) | (0.898) | (0.909) | (1.010) | (2.513) |
| R-squared | 0.166 | 0.167 | 0.165 | 0.165 | 0.165 |

B. **Overvalued Art (N=9,801)**

| | Public Exhibit (1) | Preservation (2) | Research (3) | Loan (4) | Other (5) |
|---|---|---|---|---|---|
| Art Stated Use | -4.957 | -5.599** | -2.182 | 8.633*** | 7.512 |
| | (3.989) | (2.656) | (2.757) | (2.947) | (4.785) |
| R-squared | 0.017 | 0.020 | 0.018 | 0.023 | 0.018 |

Notes: This table presents panel linear probability model (LPM) regression estimates of the relation between collection stated uses and art filing outcomes. In Panel A, conditional on disclosure, art filing outcomes are measured by the valuation of art donations and holdings. In Panel B, conditional on valuation, art filing outcomes are measured by the overvaluation of art donations and holdings. Controls include log(Total Assets), Total Revenue, Salary Expense, and Contributions/Total Revenue, all in the prior tax year, as well as Charity Nav. Stars, and dummies for Charity Nav. Rating, Family Foundation, Private Foundation, Medical Organization, Educational, Religious, Library, Museum, Foreign. Donation Valuation, and Year fixed effects are included. All variables are defined in the Appendix. Standard errors are reported in parentheses and are clustered by the organization. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively. Coefficients multiplied by 100 for readability.

## TABLE A3. Determinants of Accepting, Valuing, and Overvaluing Art Donations: Donation Valuation Method

| | Auction (1) | Comparable Sale (2) | Cost (3) | Donor Supplied (4) | Org. Estimate (5) | FMV (6) | Insurance (7) | Appraisal (8) | Artist (9) | Other / Unknown (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Art Value (N=62,541)** | | | | | | | | | |
| Method | -0.986 | 12.187*** | 4.776 | 11.943*** | 11.037** | 14.522*** | -1.429 | 27.867*** | 13.240** | -20.22*** |
| | (3.705) | (4.367) | (3.136) | (4.005) | (4.566) | (1.555) | (4.740) | (2.017) | (6.664) | (1.122) |
| $R^2$ | 0.126 | 0.127 | 0.127 | 0.127 | 0.127 | 0.135 | 0.126 | 0.153 | 0.126 | 0.163 |
| | **Overvalued Art (N=9,801)** | | | | | | | | | |
| Method | 34.267** | 0.883 | 18.349*** | 3.618 | 0.330 | 1.840 | -11.821 | -0.334 | -3.439 | -2.763 |
| | (15.004) | (6.670) | (6.532) | (6.374) | (7.451) | (2.647) | (11.878) | (2.544) | (11.518) | (1.882) |
| $R^2$ | 0.020 | 0.019 | 0.022 | 0.019 | 0.019 | 0.019 | 0.020 | 0.019 | 0.019 | 0.020 |

Notes: This table presents panel linear probability model (LPM) regression estimates of the relation between donation valuation methods and art filing outcomes. In Panel A, conditional on disclosure, art filing outcomes are measured by the valuation of art donations and holdings. In Panel B, conditional on valuation, art filing outcomes are measured by the overvaluation of art donations and holdings. Controls include log(Total Assets), Total Revenue, Salary Expense, and Contributions/Total Revenue, all in the prior tax year, as well as Charity Nav. Stars, and dummies for Charity Nav. Rating, Family Foundation, Private Foundation, Medical Organization, Educational, Religious, Library, Museum, Foreign. Donation Valuation, and Year fixed effects are included. All variables are defined in the Appendix. Standard errors are reported in parentheses and are clustered by the organization. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively. Coefficients multiplied by 100 for readability.

## TABLE A4. Organization Type with NTEE Classification

| | Art Filing (1) | Art Value (2) | Overvalued Art (3) |
|---|---|---|---|
| Audit Flag | 0.388*** | 3.536*** | 4.050* |
| | (0.038) | (0.771) | (2.111) |
| Observations | 5,295,137 | 72,147 | 10,413 |
| R-squared | 0.642 | 0.167 | 0.046 |

Notes: This table presents panel linear probability model (LPM) regression estimates of the relation between nonprofit audit flags and art filing outcomes. In model (1), art filing outcomes are measured by the disclosure of art donations or holdings. In model (2), conditional on disclosure, art filing outcomes are measured by the valuation of art donations and holdings. In model (3), conditional on valuation, art filing outcomes are measured by the overvaluation of art donations and holdings. Controls include log(Total Assets), Total Revenue, Salary Expense, and Contributions/Total Revenue, all in the prior tax year, as well as Charity Nav. Stars, and dummies for Charity Nav. Rating, Family Foundation, Private Foundation, Medical Organization, Educational, Religious, Library, Museum, Foreign. National Taxonomy of Exempt Entity Code, Art Use, Donation Valuation, and Year fixed effects are included. All variables are defined in the Appendix. Standard errors are reported in parentheses and are clustered by the organization. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively. Coefficients multiplied by 100 for readability.

## TABLE A5. Audit Flags and Logit Model

| | Art Filing (1) | Art Value (2) | Overvalued Art (3) |
|---|---|---|---|
| Audit Flag(t-1) | 0.488*** | 0.193*** | 0.231*** |
| | (0.022) | (0.032) | (0.059) |
| Don Val F.E. | - | Yes | Yes |
| Art Use F.E. | - | Yes | Yes |
| Observations | 5,295,137 | 62,541 | 9,794 |

Note: This table presents panel logistic binary choice regression (logit) estimates of the relation between nonprofit audit flags and art filing outcomes. In model (1), art filing outcomes are measured by the disclosure of art donations or holdings. In model (2), conditional on disclosure, art filing outcomes are measured by the valuation of art donations and holdings. In model (3), conditional on valuation, art filing outcomes are measured by the overvaluation of art donations and holdings. Controls include log(Total Assets), Total Revenue, Salary Expense, and Contributions/Total Revenue, all in the prior tax year, as well as Charity Nav. Stars, and dummies for Charity Nav. Rating, Family Foundation, Private Foundation, Medical Organization, Educational, Religious, Library, Museum, Foreign. Art Use, Donation Valuation, and Year fixed effects are included. All variables are defined in the Appendix. Standard errors are reported in parentheses and are clustered by the organization. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively. Coefficients multiplied by 100 for readability.

## TABLE A6. Audit Flag Ordinal Specification

| | Art Filing (1) | Art Value (2) | Overvalued Art (3) |
|---|---|---|---|
| Audit Flags(t-1)=1 | 0.464*** | 2.963*** | 4.445** |
| | (0.037) | (0.804) | (2.206) |
| # Audit Flags(t-1)=2 | 0.615*** | 4.030* | 7.196 |
| | (0.164) | (2.264) | (4.673) |
| # Audit Flags(t-1)=3 | -0.480 | -7.478 | -12.681 |
| | (0.585) | (6.360) | (17.779) |
| # Audit Flags(t-1)=4 | -7.732** | -21.826*** | |
| | (3.604) | (1.960) | |
| # Audit Flags(t-1)=5 | 10.251*** | -9.744*** | |
| | (0.392) | (1.656) | |
| Observations | 5,295,137 | 72,147 | 10,413 |
| R-squared | 0.652 | 0.172 | 0.028 |

Notes: This table presents panel linear probability model (LPM) regression estimates of the relation between nonprofit audit flags and art filing outcomes. In model (1), art filing outcomes are measured by the disclosure of art donations or holdings. In model (2), conditional on disclosure, art filing outcomes are measured by the valuation of art donations and holdings. In model (3), conditional on valuation, art filing outcomes are measured by the overvaluation of art donations and holdings. # Audit Flags(t-1)=n is a binary indicator for the total number of audit flags in each nonprofit-year. Controls include log(Total Assets), Total Revenue, Salary Expense, and Contributions/Total Revenue, all in the prior tax year, as well as Charity Nav. Stars, and dummies for Charity Nav. Rating, Family Foundation, Private Foundation, Medical Organization, Educational, Religious, Library, Museum, Foreign. National Taxonomy of Exempt Entity Code, Art Use, Donation Valuation, and Year fixed effects are included. All variables are defined in the Appendix. Standard errors are reported in parentheses and are clustered by the organization. ***, **, * indicate statistical significance at the 1%, 5%, and 10% level, respectively. Coefficients multiplied by 100 for readability.

## *B. Variable Definitions*

### *B.1 Organization Variables*

| Variable | Definition |
|---|---|
| Art Filing | An indicator variable, equal to one (Yes) when an organization indicates they have accepted art donations or have art assets in a tax year, and equal to zero (No) otherwise. Art donations are indicated by checking the box on Form 990, Part IV, Question 30, while art assets are indicated by checking the box for Form 990, Part IV, Question 8. |
| Art Value | Art Value is an indicator variable equal to one (Yes) after an organization first indicates they have accepted art donations or assets, and lists those assets and donations by filing both a Schedule D and a Schedule M in the same tax year, with donations recognized in revenue on Schedule M, and assets recognized as part of total assets in Schedule D, with valuations provided for both. Otherwise, this variable is equal to zero (No). |
| Overvalued Art | This variable is an indicator variable equal to one (Yes) when an organization's Art Value is equal to one, and when the total, cumulative net art donations over the history of an organization, subtracted by end of year total art assets is greater than zero. See Equation (1) for more details on construction. |
| log(Total Assets) (t-1) | The natural logarithm of organization total assets in the prior tax year. |
| Total Revenue (t-1) | Total revenue of the organization in the prior tax year, in millions of nominal USD. |
| Salary Expense (t-1) | Total salary expenses of the organization in the prior tax year, in millions of nominal USD. |
| Contributions/Revenue (t-1) | The fraction of organization's total revenue coming from direct donor contributions to the organization, rather than investment, asset sales, or other sources in the prior tax year. |
| Audit Flag(t-1) | An indicator for when a nonprofit identifies that it has engaged in one of the following practices on its Form 990 in the prior tax year:<br>a **diversion of assets**, which is identified by whether the box is checked form Form 990, Part VI, Question 5 or on attached Expenditure Responsibility Statements for diversion of any grants,<br>**acknowledging prohibited political activity**, identified by an organization marking yes to Part IV, Question 3, or identifying any political expenditures or volunteer work on Schedule C,<br>**unrelated business income**, filing of a Form 990-T, which is indicated on a Form 990 by Part V, Questions 3a and 3b,<br>**excess benefit transactions** or **loans to disqualified persons**, which is identified by Part X, Line 6, Part IV, Questions 25a and b, as well as Schedule L,<br>**excess compensation**, executive compensation in the top 1% of executive compensation for nonprofits in that tax year,<br>**foreign grant activity**, identified by Schedule A, Part IV, Question 4 a, b, or c, and<br>**income and expense discrepancies from fundraising events**, which is identified by an organization being in the highest or lowest%ile of Net Fundraising Revenue, or when Net Fundraising Revenue is negative.<br>This variable takes the value of one (Yes) when any of these practices is identified in a given tax year for an organization, and zero (No) otherwise. |

## *B.2 Organization Types*

| Variable | Definition |
|---|---|
| Family Foundation | Is an indicator variable equal to one (Yes) for an organization if any of the following rules are satisfied:<br>the last name of an individual on its Board, Executive Employees, Substantial Contributors, or Contributing Managers is part of the name of the organization,<br>the nonprofit's name includes the words "family foundation",<br>the number of identical last names within the previously mentioned individuals is equal to 2 or greater,<br>the last name of a substantial contributor is among the last names of all other individuals identified above that are associated with an organization,<br>the organization has been identified as a family foundation by the anonymous nonprofit rating agency from Brounstein (2023), or<br>the Family contributions to the organization, identified on Schedule A, Part IV, Line 11a-c. |
| Private Foundation | Identifies any organization that has filed a 990-PF over the course of our sample or identifies itself as a private foundation on its Form 990 at any point. |
| Medicine | Indicates the organization is either a hospital or other medical practice or healthcare providing organization. |
| Education | Indicates the organization lists schooling as its primary purpose, and includes K-12 schools as well as universities, colleges, trade schools, and other such organizations. |
| Religious | Includes all organizations identifying as houses or organizations of religious worship. |
| Library | Indicates libraries or archival organizations. |
| Museum | Indicates organizations with works for public exhibit. |
| Other | Classifies all organizations that do not fit into these, including community assistance, conservancies, and organizations that do not identify their primary mission on tax filings. |
| Foreign Operated | An indicator equal to one (Yes) if the organization's mailing or business address is listed as outside the United States, and equal to zero (No) otherwise. |
| Charity Nav. Rating | An indicator equal to one (Yes) if the organization is rated by Charity Navigator, including zero-star ratings, and equal to zero (No) otherwise. |
| Charity Nav. Stars | The number of stars assigned to the nonprofit by Charity Navigator, including 0 stars, with missing ratings also assigned 0 stars. |

### *B.3 Donation Valuation Types*

I scan all valuation type free text fields for keywords and classify valuation types based on the most common responses.

| Variable | Definition |
|---|---|
| Auction | The piece was valued by a recent auction purchase. |
| Comparable Sales | Donation was valued by recent sales of comparable items. |
| Cost | The item was either valued at the cost to acquire, cost to transport and donate, or a de minimis or other minimal value. |
| Donor | The donation value was supplied directly by the donor. |
| Organization Estimate | The organization used its own methodology or expertise to value the donation. |
| Insurance | The donation was valued by a third-party insurer. |
| Appraisal | The donation was valued by an appraiser, either provided by the organization, the donor, or a third party. |
| Artist | The artist who created the donated work provided the valuation. |
| Other | All donations that do not have valuation types or have valuation types that do not fit into any of the above categories. |

### *B.4 Art Stated Use Types*

If multiple stated uses are identified by an organization, the priority of use is given in the order of Public Exhibit, Preservation, Research, Loan and Other.

| Variable | Definition |
|---|---|
| Public Exhibit | Indicates the work was used for public exhibit. |
| Preservation | Indicates the work was used for “preservation for future generations”. |
| Research | Indicates the work was used for “scholarly research”. |
| Loan | Indicates the work was used for a “loan or exchange program” |
| Other | Indicates the work was used for “other”, which is free response category, or whether the stated use for the work was unknown or left blank on the filed Schedule M |

### *C. Tax Advantages of Art Donations*

Donating a piece of art to charity may avoid several sets of taxes within the United States. First, the donation may be income tax deductible, reducing the donors' individual tax burden. I use this measure, and only this measure, for our back-of-the-envelope calculations. As I cannot observe the individual tax choices of art donors, I cannot observe the tax impact these donations, and their valuations, have on the total amount of taxes avoided.

Second, if the asset has appreciated in value, a nonprofit donation avoids luxury capital gains taxes as the donation does not trigger a step up in basis. Art is taxed as a "collectible", and therefore has a long-term capital gains rate of 28% plus an additional Net Investment Income tax on income of 3.8% for a total of 31.8%. Therefore, an additional minimum of 3.8% of income as well as 28% of capital gains are avoided by donating the art itself rather than selling it and donating the proceeds.

Donating art to charity can act as a substantial tax avoidance mechanism while still preserving some of the key benefits that fine art derives its subjective value from. That is, while a private investor may receive substantial tax benefits for donating a piece of art to a museum, they may still derive some benefit by seeing the piece in the museum's collection—"use" of the asset, so to speak, is not necessarily lost when ownership changes.

There is another way to gain this subjective viewing benefit while still reaping tax rewards. Even if the piece is not put up for public observation in a museum or other public space, donations can be given in terms of fractional interest, where some fraction of the donation's value is tax deductible and a time-sharing agreement is achieved where the donor still has use of the piece for a fraction of the year, and the nonprofit organization uses it for the remainder. While I observe fractional interest donation categories in our data, I do not observe any usage of this method of donation in our sample. This allows a wealthy donor, for example, to enjoy the art in a vacation home while visiting for the season and reap the tax benefits and absence of storage costs for the rest of the year.

There is an additional tax loophole of this variety—the display of a newly purchased collection in a museum located in a specific state, which avoids use taxes common among most U.S. states for out of state purchases of luxury goods (See https://www.nytimes.com/2014/04/13/business/buyers-find-tax-break-on-art-let-it-hang-awhile-in-portland.html). This use of tax avoidance is unobserved in our sample, as I cannot observe the home state of the donor.

Finally, the use of freeports avoids import/export fees and duties, and any sales tax made from transactions within the freeport. Therefore, a donor may purchase a piece already in a freeport, avoiding sales tax, and rather than importing a work of art may donate it from the freeport, causing the nonprofit organization to take on these fees in addition to the other tax benefits they receive for this donation.

# Indirect Deterrence Effects From Filing and Payment Compliance Programs

*Brett Collins, Corbin Miller, Mark Payne, Sean Roh, Yan Sun, Alex Turk, and Chris Wilson (IRS, RAAS)*[1]

## 1. Introduction

The Internal Revenue Service (IRS) plays a central role in maintaining the integrity of the U.S. tax system, which relies on voluntary compliance. In Fiscal Year 2023, the IRS processed over 271 million federal tax returns and supplemental documents (Internal Revenue Service 2024a). Most taxpayers meet their filing and payment obligations without direct intervention, but ensuring sustained compliance remains a challenge. In Tax Year 2022, approximately 85% of total tax liabilities were paid voluntarily and on time, yet a substantial portion remained unpaid, contributing to a tax gap of $696 billion (Internal Revenue Service 2024b).[2] Even after enforcement efforts, a significant share of these unpaid liabilities—$606 billion—remains uncollected, underscoring the persistent difficulty of achieving full compliance and highlighting the critical role of enforcement strategies in influencing taxpayer behavior (Internal Revenue Service 2024b).

The ability of the IRS to close this tax gap through enforcement, however, has been increasingly constrained by limited resources. Between 2010 and 2019, the agency's enforcement budget declined by more than 28% in real terms, while its responsibilities expanded due to legislative changes and increased administrative burdens. Over the same period, the number of full-time equivalent (FTE) employees dedicated to enforcement fell by 34%, from 50,400 to 33,484, reducing the IRS's capacity to conduct audits and pursue delinquent taxpayers (Internal Revenue Service 2011, 2020). These constraints necessitate a more strategic allocation of enforcement resources to not only recover unpaid taxes but also maximize compliance through deterrence effects. A key question, therefore, is how enforcement efforts—particularly those related to filing and payment compliance—affect taxpayer behavior, both directly and indirectly.

Prior research distinguishes between direct enforcement effects, which apply to taxpayers who receive an enforcement action, and indirect effects, where enforcement influences individuals who were not directly contacted but adjust their behavior based on perceived risk or awareness of IRS activities. A growing body of literature highlights the importance of these spillover effects, suggesting that enforcement actions can shape compliance norms within communities and networks. For instance, Boning et al. (2019) demonstrate that IRS field visits not only increase compliance among targeted firms but also among businesses connected through the same tax preparer network. While these studies underscore the role of social networks in amplifying enforcement effects, much of the literature focuses on noncompliant taxpayers. The extent to which IRS enforcement actions reinforce compliance among previously compliant taxpayers remains an open and underexplored question.

This study seeks to fill that gap by investigating the indirect deterrence effects of IRS enforcement on taxpayers who were compliant in the previous year but may become delinquent in the current year. Specifically, we examine how IRS enforcement actions function as a preventative mechanism, sustaining voluntary compliance among historically compliant taxpayers. Our analysis focuses on three key enforcement actions related to filing and payment compliance: Automated Collection System (ACS) notices, which are mailed reminders sent to taxpayers with outstanding balances; CP59 notices, which target nonfilers, requesting submission of overdue returns; and field collection visits, where IRS revenue officers conduct in-person interventions to address persistent delinquencies. These enforcement strategies vary in their intensity and reach. ACS and CP59 notices are correspondence-based enforcement tools, allowing the IRS to contact a broad population of taxpayers at a relatively low cost. Field collection visits, in contrast, are resource-intensive and geographically localized, with revenue officers directly engaging delinquent taxpayers. While these actions are primarily designed to address existing noncompliance, their visibility within communities may influence taxpayers who have not yet fallen into delinquency, reinforcing the perceived risk of noncompliance and encouraging continued compliance.

---

[1] The view and opinions presented in this paper reflect those of the authors. They do not necessarily reflect the views or the official position of the Internal Revenue Service.

[2] The tax gap is the difference between the total true tax liability owed by taxpayers for a given tax year and the amount that is paid voluntarily and on time. It consists of three components: (1) Nonfiling—tax not paid on time by those who do not file required returns, (2) Underreporting—tax that is understated on timely filed returns, and (3) Underpayment—tax that is reported but not paid on time.

To assess these indirect effects, we leverage a two-stage least squares (2SLS) regression model with instrumental variables, using fluctuations in IRS enforcement resources from 2011 to 2019 as a natural experiment. These fluctuations provide a natural experiment to help isolate the causal impact of enforcement actions from potential confounding factors. Additionally, we employ the Facebook Social Connectedness Index (SCI) to capture how enforcement awareness spreads through social networks, rather than relying solely on geographic proximity for a proxy of enforcement exposure. This approach allows us to quantify how enforcement actions propagate compliance effects beyond directly treated individuals and across socially connected communities.

Our findings reveal that IRS enforcement actions have significant indirect deterrence effects on previously compliant taxpayers. ACS and CP59 notices, in particular, generate measurable spillover effects due to their broad reach and visibility. We estimate that a 10% increase in ACS notices reduces newly accrued delinquent balances among previously compliant taxpayers by 16%, highlighting the substantial compliance benefits of scalable, correspondence-based enforcement. Moreover, these effects are amplified in regions with higher social connectedness, suggesting that enforcement actions influence taxpayer behavior through both direct treatment and social spillovers.

This study makes several key contributions to the tax compliance literature. First, it broadens the scope of enforcement research by demonstrating how compliance interventions can sustain voluntary compliance rather than merely rectifying noncompliance. Second, by incorporating network-based measures of enforcement exposure, our analysis offers a more comprehensive understanding of how taxpayers perceive enforcement risk. Third, the utilization of instrumental variables and natural fluctuations in enforcement resources allows us to identify causal relationships between enforcement actions and taxpayer behavior. Furthermore, our results imply that conventional estimates of the direct effects of low-cost, frequent enforcement actions may substantially understate their total impact by neglecting spillovers to compliant taxpayers via social networks. Recognizing the influence of enforcement salience among social contacts could enhance the design and effectiveness of compliance programs.

The remainder of the paper is organized as follows. Section 2 reviews the institutional background of IRS enforcement mechanisms and the literature on tax compliance spillovers. Section 3 details the data sources and empirical methodology, including our instrumental variables and network-based enforcement measures. Section 4 presents the primary empirical findings, quantifying indirect enforcement effects. Finally, Section 5 discusses policy implications and concludes with suggestions for future research.

## 2. Background

### *2.1 Program Trends*

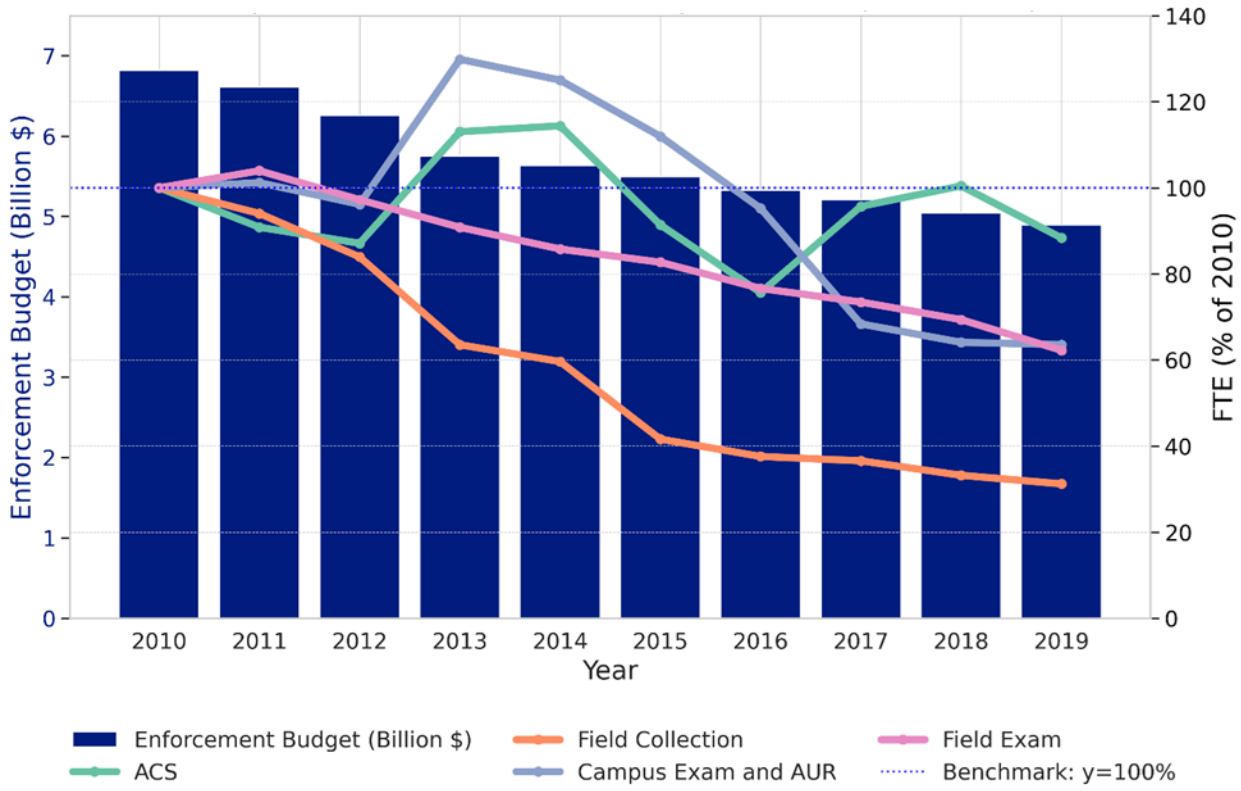
The IRS enforcement budget declined by more than 28% in real (inflation-adjusted) terms from 2010 to 2019, as illustrated in Figure 1. This budget contraction occurred alongside an increasing number of tax returns to process and growing administrative responsibilities related to new legislation and compliance issues. The smooth downward trend in budget figures masks the substantial challenges faced during this period, including rising instances of identity theft, multiple federal government shutdowns—most notably the longest in U.S. history during 2018-2019—and the need to adapt to significant legislative changes such as the Tax Cuts and Jobs Act of 2017.

To manage its expanding responsibilities amid constrained resources, the IRS had to redistribute its workforce across various programs, leading to cutbacks in several filing and payment compliance initiatives. Between 2010 and 2019, the number of FTE employees at the IRS decreased by 34%, from 50,400 to 33,484 (Internal Revenue Service (2011, 2020)). Figure 1 illustrates the relationship between enforcement budget trends and FTE allocations across different enforcement activities. While the overall trend shows a decline in staffing, there are notable shifts in how FTEs were allocated to different enforcement types. These shifts reflect administrative decisions that were shaped by resource constraints and broader enforcement priorities rather than direct responses to individual taxpayer compliance behavior.

These heterogeneous changes in the allocation of FTE positions subsequently affected the number of enforcement actions—such as ACS letters, CP59 notices, and field visits—conducted under each type of program. This variation in enforcement intensity, driven by exogenous shifts in FTEs, creates a natural experiment that allows us to tease out the causal

effect of enforcement activities on taxpayer compliance. Such insights are particularly valuable for optimizing resource allocation, especially given the increased funding for enhanced enforcement efforts following the Inflation Reduction Act of 2022.

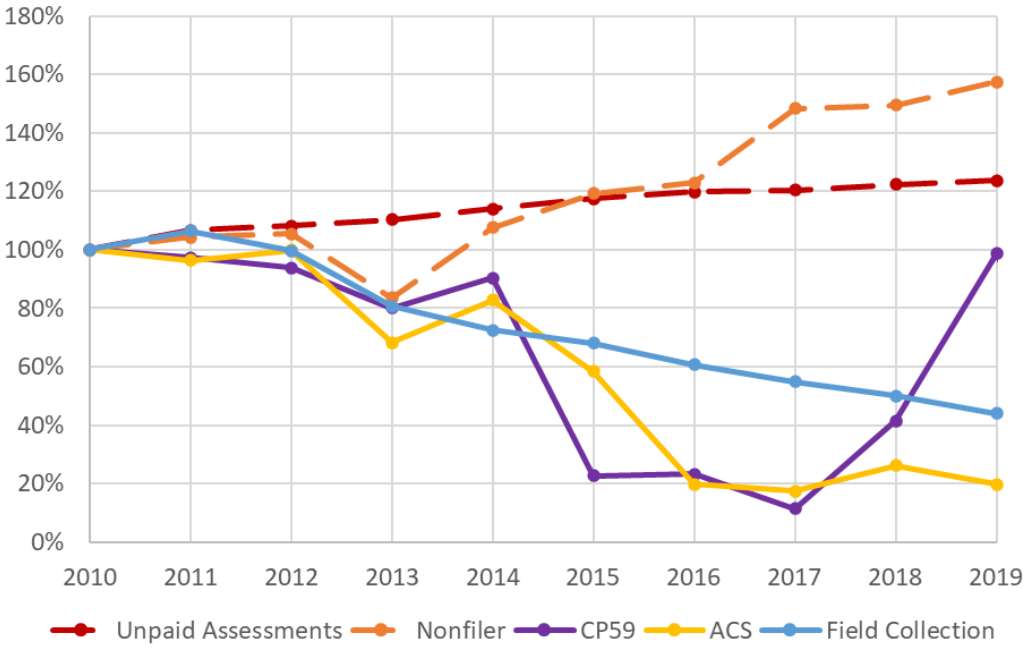
**FIGURE 1. IRS Enforcement Budget and FTEs, 2010 to 2019**




Note: The bar chart represents the annual enforcement budget from 2010 to 2019. The line graphs show the annual FTE positions for various enforcement programs, including the ACS, Campus Examinations, Automatic Underreporter (AUR), Field Collection visits, and Field Examinations. All line graphs are normalized to 2010 as the base year to allow for consistent trend comparison across different FTE scales. Source: IRS Data Book, Table 30, inflation adjustment calculated with Bureau of Labor Statistics CPI-U consumer price index.

Reductions in appropriations and FTEs coincided with increases in the numbers of taxpayers in a delinquent status. As shown in Figure 2, which tracks key trends for individual Form 1040 taxpayers over our study period by using 2010 as a base year, the number of taxpayers with unpaid assessments (UA) and those identified as nonfilers through the Case Creation Nonfiler Identification Process increased. This rise in noncompliant taxpayers contrasts with a decline in enforcement actions, as reflected by the reduced issuance of delinquent return notices (CP59), selected ACS letters, and field collection assignments. Within these overall declines, there is considerable variation in both the timing and intensity of these enforcement programs, which enhances the ability of our models to isolate each program's impact on compliance.

**FIGURE 2. Compliance Trends**

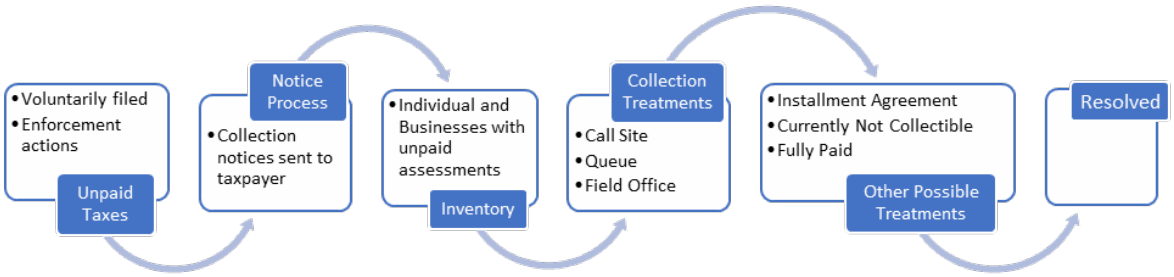



Notes: Based on authors' calculation from administrative tax data. This figure shows counts of noncompliant taxpayers (dotted lines) and IRS enforcement actions (solid lines) indexed to 2010.

## *2.2 Program Operation*

The IRS collection process begins with a return matching system that cross-references taxpayer-reported income against third-party sources, such as employer-reported wages and financial institution filings. If a taxpayer fails to file a return, then this matching process cannot proceed, and the case is placed into the delinquent return inventory. Regardless of filing status, taxpayers with unpaid assessments enter the collection process (Internal Revenue Service (2024b)). Upon entering the collection process, taxpayers typically receive a balance due notice, informing them of their outstanding liability and providing instructions for resolution (Figure 3). Nonfilers may receive a CP59 notice, which notifies taxpayers that the IRS has no record of a filed tax return for a given year. Recipients are advised to submit their return immediately or provide justification for not filing. If the taxpayer does not respond or resolve their liability after these initial notices, the IRS may escalate enforcement efforts by opening a Taxpayer Delinquent Account (TDA), which may trigger further collection actions.

**FIGURE 3. IRS Collection Process**



Cases are assigned to different IRS collection programs based on factors such as the size of the unpaid balance, the complexity of the taxpayer's financial situation, and available enforcement resources. The IRS employs both automated and in-person enforcement mechanisms. The Automated Collection System (ACS), which operates from IRS campus facilities, issues a series of notices to prompt taxpayer compliance. Among these, LT11 (Final Notice of Intent to Levy and Your Notice of a Right to a Hearing) serves as the final warning before the IRS proceeds with asset seizure, informing taxpayers of their right to contest the levy. The LT16 notice is a reminder urging immediate resolution of unpaid balances to prevent potential enforcement actions, such as levies or liens. The LT26 notice is directed at nonfilers who have ignored prior IRS communications, demanding that they file their outstanding tax returns (Internal Revenue Service (n.d.)).

For cases requiring in-person enforcement, Revenue Officers (ROs) conduct field collection visits, typically reserved for high-priority cases involving significant unpaid balances, uncooperative taxpayers, or complex financial circumstances. Field visits allow the IRS to obtain financial disclosures, issue levies, or negotiate installment agreements directly. However, in a recent policy shift, the IRS has largely ended unannounced visits by revenue officers to improve taxpayer safety and reduce confusion. Instead, ROs now initiate contact through an appointment letter (IRS Form 725-B) to schedule meetings. Depending on the taxpayer's response, their case may be reassigned to different enforcement treatments, placed in the IRS "queue" pending further review, or resolved through full payment, installment agreements, or designation as Currently Not Collectible (CNC) status, the equivalent of a write off.

In this study, we focus on ACS notices (LT11, LT16, and LT26) and CP59 notices, as well as field collection visits by ROs. These interventions differ in their implementation: ACS and CP59 notices are remote enforcement mechanisms issued from centralized IRS facilities, whereas field visits involve direct engagement by local IRS offices. Given their higher resource intensity, field visits are much less common than ACS notices. A critical distinction in IRS enforcement is between discretionary enforcement actions, such as ACS notices and field visits, and automatic collection procedures, such as balance due notices. While balance due notices represent a mandatory early-stage enforcement step, they do not constitute discretionary enforcement actions. Unlike ACS notices, which can be intensified or strategically deployed based on IRS priorities, balance due notices are systematically issued to all taxpayers with outstanding balances (Internal Revenue Service 2024a).

Because this study seeks to estimate the causal effects of discretionary enforcement actions, we exclude balance due notices from our treatment variables. These notices lack exogenous variation, making it difficult to disentangle their compliance effects from broader systemic enforcement trends. While the issuance of balance due notices was temporarily disrupted during the COVID-19 pandemic, isolating the compliance effects of this pause from other pandemic-related factors—such as economic stimulus payments and temporary IRS enforcement suspensions—falls beyond the scope of this analysis.

### *2.3 Literature Review*

The impact of IRS compliance programs can be broadly categorized into direct effects—changes in current and future behavior for taxpayers subject to enforcement, and indirect effects—where nontreated taxpayers adjust their behavior based on perceived IRS enforcement activity. Indirect effects suggest that nontreated taxpayers acquire information about enforcement likelihood through various channels, such as preparer networks, public data on IRS activities, social circles, or news outlets. These effects are particularly relevant for tax administration, given the critical role of voluntary compliance in the U.S. tax system (Bloomquist (2012); Datta et al. (2015); Boning et al. (2019)). Typically, in the literature, the success of IRS compliance programs is evaluated based on observed changes in taxpayer behavior, such as timely filing and payment or the resolution of prior delinquencies.

Most previous studies that estimate indirect effects have focused on specific programs but demonstrate notable indirect effects. For instance, Datta et al. (2019) analyzed the Automated Substitute for Return (ASFR) program, finding that indirect effects increased the likelihood of filing for nontreated taxpayers by up to 27%, surpassing direct effects and persisting over time. Similarly, Turk and Ashley (2002) examined the Notice of Federal Tax Lien (NFTL) program and leveraged a policy change to assess both the direct and indirect effects on delinquent taxpayers' likelihood of resolving their tax debts. Although indirect effects have gained increasing attention, the research remains constrained due to technical

complexities and data limitations (Boning et al. (2019)). Studies in this area typically rely on three methodological approaches: field experiments, laboratory experiments, and natural experiments.

### 2.3.1 Field Experiments

Lopez-Luzuriaga and Scartascini (2023) conducted a field experiment in Argentina in 2011 to examine how various interventions affected nonpayers' compliance with unpaid property taxes. They compared three types of treatment messages: deterrence, reciprocity, and peer-effect. The study found that a deterrence letter—emphasizing penalties and the likelihood of detection—was the most effective in increasing compliance.

Their model predicted that taxpayers with higher tax morale or risk aversion are more likely to comply, while liquidity constraints pose challenges to compliance. Although the focus was on direct effects, the study reinforced that the perception of penalties and detection probability are key factors in tax compliance, echoing findings from Boning et al. (2019). Furthermore, the dissemination of information on penalties and detection is not restricted to the treated taxpayers but spreads through social networks, suggesting that compliance behaviors may change through indirect channels such as group effects (Bloomquist (2012)) or network effects (Boning et al. (2019)).

Boning et al. (2019) conducted a randomized experiment in 2015 to study both direct and indirect effects of IRS enforcement on employer Federal Tax Deposit collections. The study tested the effects of sending letters and conducting in-person visits to at-risk firms. They found that in-person visits had significant and persistent direct effects on tax payments, while letters had smaller effects. The study specifically examined network effects, where information about enforcement activities spread through shared tax preparers. Firms whose tax preparers had other clients receiving in-person visits from IRS Revenue Officers were more likely to remit taxes. The aggregate network effect was larger than the direct effect, producing 1.2 times more revenue. No similar network effect was observed for letters.

### 2.3.2 Agent-Based Models

Bloomquist (2012) emphasizes the importance of indirect effects in tax administration and identified three types of these effects: induced, subsequent period, and group effects. Although Bloomquist considers changes in taxpayer behavior due to prior audits as indirect, we treat these as direct effects. More relevant are group effects, where individuals alter their behavior based on others' experiences, such as those within the same community or workplace. Bloomquist estimates that every $1 detected through audit selection generates $6 to $11.60 in indirect effects. Bloomquist developed an Agent-Based Model (ABM) to quantify these indirect effects, using artificial taxpayer data from Tax Year 2001 to simulate income tax reporting behavior in a small region. The model shows that audit selection strategies incorporating indirect effects—particularly group effects—yield a greater impact on voluntary compliance, as taxpayers adjust their reporting behavior based on the audit experiences of their neighbors and coworkers.

### 2.3.3 Natural Experiments

Using a natural experiment stemming from declining IRS budgets and reduced enforcement activity, Datta et al. (2015) analyzed the direct and indirect impacts of the IRS ASFR program on delinquent tax collections and subsequent compliance. The study also estimated the program's indirect effects on nonfilers more broadly. The dataset comprised a random 10% sample of delinquent tax returns from Tax Years 2007 to 2009, and subsequent returns. The study first calculated the predicted probability of taxpayers being selected for ASFR treatment to capture indirect effects, followed by estimates of revenue collected in subsequent years. Results showed significant direct and indirect impacts on compliance. Treated cases generated $672 to $1,640 in revenue, depending on the model, while untreated cases exhibited indirect effects ranging from $194 to $1,187. The study also identified stronger, longer-lasting indirect effects on filing compliance.

Turk et al. (2016) examined the direct effects of the IRS NFTL program on delinquent tax collections for individuals and businesses. The study used a policy change in NFTL filing thresholds in 2011 as a natural experiment, tracking taxpayer outcomes for two years after cases were transferred from the ACS to the Field Collection Queue. The research found that NFTLs significantly increased the likelihood of reducing outstanding balances, with individual taxpayers' balances falling by 22 to 23% over 1 year and 33 to 35% over 2 years. Business taxpayers experienced larger reductions, ranging from 38 to 40% over 1 year to 60 to 65% over 2 years.

# 3. Data and Methodology

## *3.1 Data*

The study population consists of individual filers who fully paid and timely filed their return in the prior year and had no unpaid tax assessments from any prior tax period. Our analysis covers the period from 2011 to 2019, a time marked by significant reductions in IRS compliance program resources but preceding the COVID-19 pandemic, which significantly altered taxpayer interactions with the IRS. To assess compliance, we consider only the most recently filed return and current balance due, drawing a 1-percent random sample of compliant taxpayers each year. This results in a repeated cross-sectional data structure, with approximately 1.2 to 1.3 million observations per year. No taxpayer appears in more than 1 year, and the total sample size for all years combined is 11.6 million.

Since most compliant taxpayers remain compliant across time (approximately 95%), a substantial portion of the sample is lost when analyzing models that focus on those who fall out of compliance. To address this issue, we generate an alternative 10% sample using a different seed for the random sampling. This approach significantly increases the number of noncompliant taxpayers in the dataset, expanding the sample size from around 600,000 to nearly 6,000,000 observations.

Tax information for the study population during the pilot year was compiled from individual return filings, data on unpaid tax assessments, and information return filings. These datasets provide comprehensive details on income types and amounts, changes in outstanding balances, compliance risk scores, exam classification groups, and other characteristics. We use prior year ($t$-1) data as controls and predictors in our models for current year ($t$) outcomes. Our primary dependent variables are current year filing and payment compliance, and for taxpayers who fail to file and pay on time, we examine the magnitude of their outstanding balance due over the course of the current year. Since all sampled taxpayers had a zero-balance due at the start of year (they were compliant), the balance can only remain at zero or increase if they fail to fully pay during the year.

In addition to analyzing the sample of compliant taxpayers, we construct treatment variables at the ZIP Code level to capture broader changes in IRS compliance programs. Our analysis primarily focuses on IRS campus programs that correspond with taxpayers—specifically, delinquent return notices (CP59) aimed at nonfilers and ACS communications (LT11, LT16, and LT26 letters) directed to those who have unmet filing or payment obligations. We also include field collection programs, which involve in-person visits to taxpayers with unpaid assessments.

The focus on ACS letters, CP59 notices, and field visits aligns closely with our research objective, as these enforcement actions are directly relevant to taxpayer behavior related to timely filing and full payment. To ensure a comprehensive analysis of IRS enforcement's indirect effects, we also include measures of IRS underreporting compliance: campus and field examinations. Campus exams, conducted remotely, are designed for straightforward issues and simpler cases, whereas field exams are in-person audits for more complex situations, often involving businesses or high-income individuals, with IRS agents reviewing records directly at the taxpayer's location. Field exams, therefore, tend to be more thorough and resource intensive.

To quantify these enforcements, we aggregate the volume of letters, field visits, and campus and field exams conducted within each ZIP Code. We include ZIP Code fixed effects to account for unobserved heterogeneity across geography. This approach enhances the accuracy of our results by controlling for time-invariant characteristics that might otherwise confound our estimates.

Table 1 presents the summary statistics for the dependent variable and key treatment variables used in this study, highlighting several important trends. The dependent variable captures taxpayer compliance behavior, categorized into three levels based on filing and payment status:

- **Fully compliant** taxpayers are those who file their tax returns and pay their liabilities by the original due date.
- **Late filers** are taxpayers who miss the filing deadline but still manage to pay the full tax amount by the original return due date.
- **Delinquent** taxpayers include those who fail to pay the full tax amount by the original due date, which includes nonfilers and those who still have outstanding liabilities despite filing on time or late.

The summary statistics reveal a noticeable decline in the proportion of compliant taxpayers from 2011 to 2019, accompanied by a corresponding increase in the noncompliant population over the same period. This shift suggests a growing challenge of maintaining compliance levels during the study period.

**TABLE 1. Summary Statistics – Means of Key Variables**

| Type of Variables | All (2011 2019) (1) | 2011 (2) | 2019 (3) |
|---|---|---|---|
| **Outcome Variables** | | | |
| Fully Compliant (%) | 94.9 | 95.4 | 94.1 |
| Late Filer (%) | 2.1 | 1.7 | 2.6 |
| Delinquent (%) | 3.0 | 2.9 | 3.3 |
| $\Delta$ Balance Due (All Prev. Compliant, $) | 157 | 115 | 186 |
| $\Delta$ Balance Due (Delinquent Only, $) | 5213 | 3873 | 5666 |
| **Treatment Variables** | | | |
| ACS | 172.5 | 286.3 | 72.8 |
| CP59 | 113.0 | 183.1 | 87.4 |
| Field collection | 11.0 | 13.6 | 7.7 |
| Campus exam | 9.9 | 12.4 | 6.5 |
| Field exam | 43.5 | 56.4 | 33.4 |
| **Control Variables** | | | |
| Married filing jointly | 0.37 | 0.39 | 0.36 |
| Log total positive income | 10.33 | 10.24 | 10.42 |
| Timely filed in past four years | 0.73 | 0.75 | 0.72 |
| Balance due (before remittance) | 0.13 | 0.11 | 0.14 |
| % of income under-withheld | -0.07 | -0.08 | -0.06 |
| $\geq$50% of income not subject to withholding | 0.10 | 0.10 | 0.10 |
| Observations | 10,246,313 | 1,086,418 | 1,181,211 |

Notes: Means of each variable are presented for each category over the entire sample period (2011-2019) in All and separately for the years 2011 and 2019. The treatment variables (ACS, CP59, Field collection, Campus exam, and Field exam) are ZIP Code-level counts. The control variables reflect taxpayer-level measures from the prior tax year. Units are specific to each variable, where applicable.

In addition to changes in taxpayer compliance, our main treatment variables, which represent different enforcement actions—ACS notices, CP59 notices, and field visits—show significant decreases over the study period. Specifically, the average number of ACS notices per ZIP Code fell from 286 in 2011 to 73 in 2019, a reduction of approximately 74%. CP59 notices also declined, dropping from 183 to 87 per ZIP Code, a decrease of about 52%. Field visits saw a similar downward trend, decreasing from 14 to 8 per ZIP Code, representing a 43% reduction.

The data also show considerable disparities in the frequency of enforcement actions. On average, 173 ACS notices were sent per ZIP Code annually, which is 53% higher than the average number of CP59 notices. Field visits were even less frequent, with ACS notices being issued 16 times more often than field visits, which averaged only 11 per ZIP Code. These

disparities in the frequency and scale of enforcement actions suggest that the marginal impact of each treatment variable on taxpayer compliance may vary substantially.

Overall, the summary statistics underscore the critical role of enforcement actions in shaping tax compliance behavior and highlight the significant variation in the intensity of different enforcement strategies. The sharp decline in enforcement activities over the years raises concerns about the IRS's ability to sustain compliance rates, especially as resource constraints continue to limit its operational capacity.

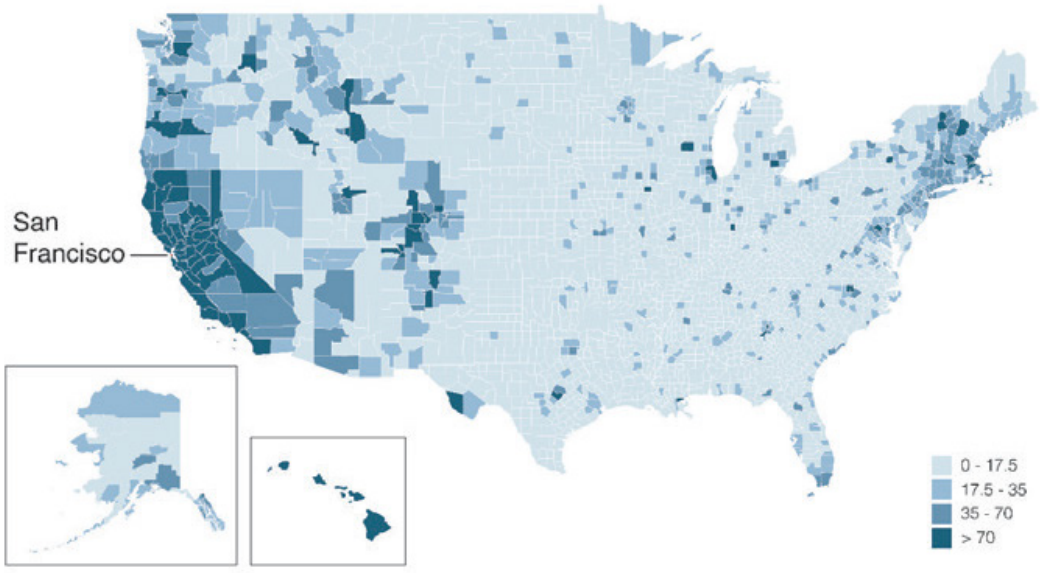
### *3.2 Social Connectedness Index*

To accurately assess the indirect effects of IRS enforcement, it is crucial to understand how information about enforcement actions circulates through social networks, which may span geographic areas, preparer networks, or supply chains. For this purpose, we use the Social Connectedness Index (SCI), developed by Bailey et al. (2018). The SCI measures the intensity of connections between ZIP Code pairs using anonymized Facebook friendship data from 2016, a time when approximately two-thirds of all U.S. adults used Facebook (Greenwood et al. (2016)), to reflect the density of social connections across the U.S. Given Facebook's extensive user base and a demographic profile that mirrors the general population, the SCI provides a reliable indicator of social networks, offering valuable insights into how social ties influence perceptions of enforcement actions.

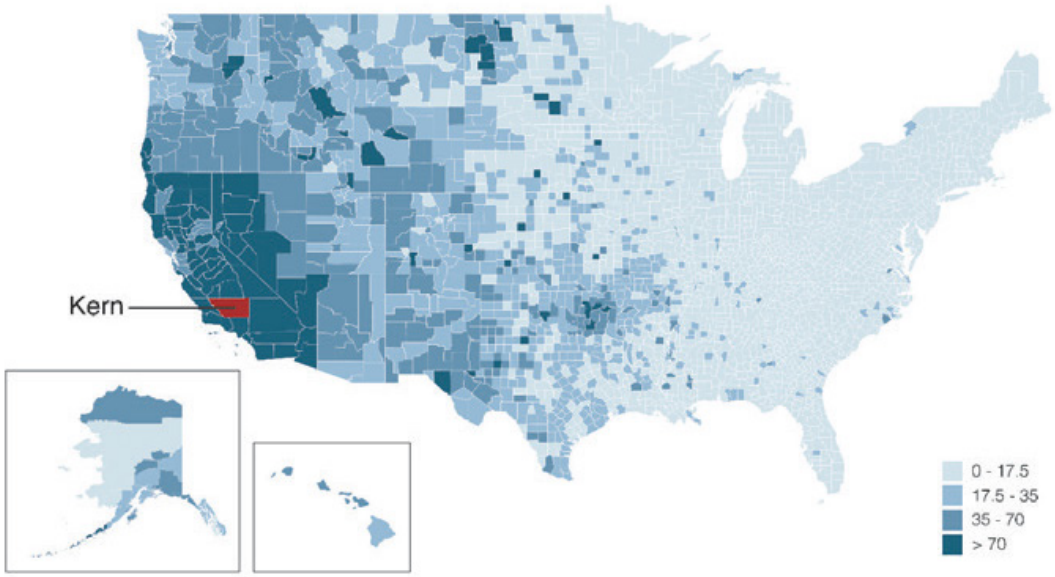
Unlike traditional measures of social proximity that rely on geographic location, the SCI captures actual social connections, offering a more nuanced understanding of how individuals are linked across regions. For example, as depicted in Figure 4, while San Francisco County and Kern County in California have similar population sizes, their social networks are markedly different. San Francisco's connections are dispersed nationally, particularly into the Northeast, while Kern County's network is concentrated on the West Coast, with strong ties to regions such as Oklahoma and Arkansas due to historical migration patterns. Specifically, 57% of Kern County's friendships are within 50 miles, closely matching the U.S. average of 55.4%, whereas only 27% of San Francisco's friendships are within the same range, highlighting its broader social dispersion. By calculating the "relative probability of friendship"—adjusted for the number of Facebook users—the SCI provides a more precise measure of social connectedness that goes beyond simple geographic proximity. This measure is crucial for understanding how perceptions of IRS enforcement actions spread within and across communities, as geographic closeness alone does not fully capture the strength and influence of social ties.

**FIGURE 4. County-Level Friendship Maps**

(a) Relative Probability of Friendship Link to San Francisco County, CA




(b) Relative Probability of Friendship Link to Kern County, CA




Note: The heat maps illustrate the relative likelihood of a Facebook user in each county having a friendship connection with San Francisco County, CA (Panel a) and Kern County, CA (Panel b). Darker shades indicate counties where there is a greater likelihood of a friendship connection from a person in the home county (San Francisco or Kern) to county. The "relative probability of friendship" is derived by dividing the Social Connectedness Index between counties and by the product of the total number of Facebook users in both counties. Source: Bailey et al. (2018).

### 3.2.1 Data Coverage and the Social Connectedness Index

During the study period from 2011 to 2019, an average of 148 million individual tax returns were filed annually across 58,960 ZIP Codes, covering all 50 states and the District of Columbia. This figure includes not only standard geographic

ZIP Codes but also P.O. Box-only ZIP Codes, unique codes for large organizations, and military ZIP Codes. By contrast, according to U.S. Postal Service data from 2024, there are approximately 41,704 standard geographic ZIP Codes in the United States. Our 1% random sampling of individual tax returns reduces the number of ZIP Codes in our dataset to 39,794 out of the 58,960 total ZIP Codes. Crucially, only 0.01% of tax returns are filed in ZIP Codes outside of this sample, ensuring that our data remains highly representative of taxpayer behavior across the U.S.

The SCI, used to capture social connections between ZIP Codes, further limits coverage due to privacy concerns, excluding ZIP Codes with very few users. As a result, the SCI encompasses 22,718 ZIP Codes, representing the ZIP Codes for which our weighted average treatment variables are available. While the exclusion of some ZIP Codes might seem significant, it is important to note that the 22,718 ZIP Codes covered by the SCI account for 97% of all tax returns filed during the 2011 to 2019 period. The remaining 3% of tax returns come from ZIP Codes in remote areas with sparse populations and minimal tax return activity, meaning their exclusion has little impact on the representativeness of our analysis. Therefore, our dataset captures the majority of taxpayer interactions and remains robust for the purposes of our analysis.

### 3.2.2 Treatment Variables Transformation

To capture the indirect effects of IRS enforcement actions, we transform key treatment variables, ACS notices, CP59 notices, and field visits using weighted averages based on the SCI. This transformation accounts for how enforcement actions in one ZIP Code may influence taxpayer behavior in socially connected ZIP Codes, reflecting the spread of enforcement perceptions through social networks.

For example, the transformation of ACS notices is calculated as follows:

$$ACS_{jt} = \sum_{k} w_{jk} ACS_{kt}^{raw} \tag{1}$$

where $ACS_{jt}$ represents the weighted average of ACS letters sent to ZIP Code $j$ in year $t$, is the social connection measure between ZIP Code $j$ and $k$, and is the number of ACS letters sent to ZIP Code $k$ in year $t$. This method enables our models to capture how social connections, rather than geographic proximity alone, shape the dissemination of enforcement perceptions and influence taxpayer behavior. By incorporating SCI-weighted averages, we reflect the indirect effects of enforcement actions as they propagate through connected communities.

Table 2 summarizes the treatment variables after applying the SCI transformation. The overall trends remain similar to the raw data, showing noticeable declines in enforcement actions over time. However, the SCI-weighted variables are larger on average, reflecting the amplifying effect of social connections. Importantly, the standard deviations of the treatment variables decrease significantly after the transformation, indicating that the SCI smooths out extreme variations in the raw data, explained in more detail below. This reduction is because some ZIP Codes that received fewer direct enforcement actions in the raw data are socially connected to others that received more intensive enforcement, allowing for a more accurate measure of the indirect effects through social spillovers.

**TABLE 2. Treatment Variables after Transformation**

| Variable | All | | 2011 | | 2019 | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| Unweighted ACS | 173 | 280 | 286 | 383 | 73 | 141 |
| Unweighted CP59 | 113 | 182 | 183 | 248 | 87 | 168 |
| Unweighted Field collection | 11 | 14 | 14 | 18 | 7.7 | 9.0 |
| Unweighted Campus exam | 9.9 | 15 | 12 | 18 | 6.5 | 9.4 |
| Unweighted Field exam | 44 | 75 | 56 | 97 | 33 | 51 |
| SCI-Weighted ACS | 191 | 144.7 | 319 | 152 | 81.6 | 38.9 |

| Variable | All | | 2011 | | 2019 | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| SCI-Weighted CP59 | 124 | 92.9 | 206 | 99.4 | 100 | 51.1 |
| SCI-Weighted Field collection | 11.9 | 5.9 | 15.2 | 7.1 | 8.4 | 3.3 |
| SCI-Weighted Campus exam | 49.2 | 33.4 | 64.0 | 42.7 | 37.2 | 21.3 |
| SCI-Weighted Field exam | 10.8 | 7.1 | 13.8 | 8.3 | 7.0 | 3.6 |

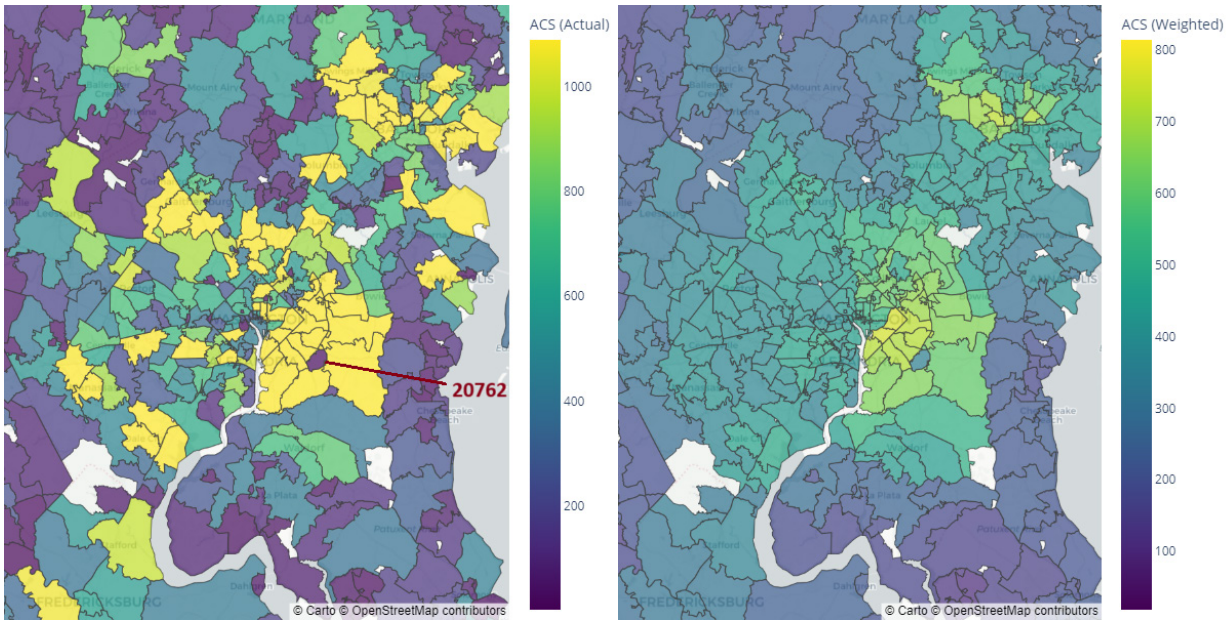
Notes: "Mean" and "SD" denote the mean and standard deviation for the entire sample period (2011-2019) in All and separately for the years 2011 and 2019.

### 3.2.3 Smoothing Effect of the SCI Transformation

Figure 5 compares the distribution of actual ACS letters sent to various ZIP Codes in the Washington D.C. area with the distribution after the SCI transformation. This comparison highlights two key points. First, the SCI transformation smooths out the varying values of ACS notices across different ZIP Codes. In the left panel, some ZIP Codes received over 1,000 notices, while adjacent ZIP Codes received only a few. This stark variation can be misleading for analyzing indirect effects, as these effects propagate through social connections, which often align with—but do not strictly adhere to—geographic proximity. The right panel, which uses SCI-weighted data, shows a more gradual variation in ACS notices, offering a clearer understanding of how enforcement messages spread through social networks.

Second, the contrasting examples of ZIP Codes 20762 (Joint Base Andrews) and 20742 (University of Maryland) illustrate that geographic proximity alone does not fully explain how enforcement effects propagate. In 2011, ZIP Code 20762 received only about 20 notices, despite surrounding areas receiving over 1,000. Even after the SCI transformation, the low social connectivity of this military base results in a relatively low number of notices. In contrast, ZIP Code 20742, which initially received fewer than 1% of the notices compared to its neighbors, shows almost no difference after the SCI transformation due to its higher social connectivity. These examples highlight the critical role of social networks—rather than geographic distance alone—in determining how enforcement messages disseminate across regions.

**FIGURE 5. Comparison of Raw and SCI-Weighted ACS Letters in the Washington D.C. Area**




Notes: These panels illustrate the distribution of ACS letters across different ZIP Codes in the Washington D.C. area. The left panel displays the raw counts of notices sent in 2011, winsorized at the top 5% level to enhance visual clarity. The right panel presents the ACS notices weighted by the Social Connectedness Index. The color bars indicate the respective ranges for each panel.

### *3.3 Regression Modeling Framework*

To evaluate the causal impact of IRS enforcement actions on taxpayer behavior, we employ a two-stage econometric approach that accounts for both the extensive and intensive margins of compliance. The first stage estimates the likelihood of behavioral transitions among previously compliant taxpayers, categorizing them into three distinct compliance states: continued full compliance, late filing, and delinquency with an outstanding balance. We use a multinomial logistic model to analyze how exposure to enforcement actions indirectly affects these transitions.

In the second stage, we investigate the financial consequences for taxpayers who enter delinquency, estimating the effect of enforcement actions on the magnitude of outstanding balances. This is accomplished using an ordinary least squares (OLS) regression, where the dependent variable captures changes in the balance due. By integrating these two stages, our framework enables a comprehensive assessment of enforcement effectiveness, distinguishing its role in preventing noncompliance (deterrence effect) and mitigating the financial severity of delinquency (recovery effect).

#### 3.3.1 Two-Stage Multinomial Logistic Model for Filing and Payment Compliance

To address potential endogeneity in the ACS, CP59, Field, Campus, and Field Exam variables—where regions with higher noncompliance may experience greater enforcement efforts—we employ a two-stage least squares (2SLS) approach. This method allows us to disentangle the effects of enforcement actions from the reverse causality driven by underlying noncompliance rates. By using the IRS's annual FTE allocations for specific types of enforcement as instrumental variables (IVs), we isolate exogenous variations in enforcement, producing unbiased estimates of enforcement effects on compliance. These FTE allocations are determined administratively and are, therefore, exogenous to taxpayer compliance behaviors, making them ideal instruments for this setting.

**First Stage**. In the first stage, we model enforcement variables (ACS, CP59, Field, Campus, and Field Exam) for each year and ZIP Code, using FTE positions allocated annually to each type of enforcement as instrumental variables (IVs). Unlike ACS and field collection programs, CP59 notices do not have dedicated FTEs. Instead, FTEs allocated to collection enforcement units—such as ACS and field collection—are interchangeably used for CP59 cases as well. To reflect the shared and overlapping nature of IRS collection efforts, we utilize both ACS and field collection FTEs, along with interaction terms, to predict the number of ACS, CP59, and Field interventions.

In contrast, the exam units operate more distinctly from the collection units. Campus and field exams have their own specific FTE allocations, and these are used directly in our models. To capture potential nonlinear relationships, such as diminishing or increasing returns from increased staffing, we include quadratic terms for each type of exam-related FTEs. This nuanced modeling approach enables us to better understand how variations in IRS staffing—whether shared among collection units or specific to exams—impact enforcement activities.

The model formulations for the first stage are as follows:

$$ACS_{jt} = \alpha_1 + \beta_1 FTE_t^{ACS} + \beta_2 FTE_t^{Field} + \beta_3 FTE_t^{ACS} * FTE_t^{Field} + \gamma_{zip} + \nu_{jt} \tag{2}$$

$$CP59_{jt} = \alpha_2 + \beta_4 FTE_t^{ACS} + \beta_5 FTE_t^{Field} + \beta_6 FTE_t^{ACS} * FTE_t^{Field} + \gamma_{zip} + \nu_{jt} \tag{3}$$

$$Field_{jt} = \alpha_3 + \beta_7 FTE_t^{ACS} + \beta_8 FTE_t^{Field} + \beta_9 FTE_t^{ACS} * FTE_t^{Field} + \gamma_{zip} + \nu_{jt} \tag{4}$$

$$Campus_{jt} = \alpha_4 + \beta_{10} FTE_t^{Campus} + \beta_{11} FTE_t^{Campus} * FTE_t^{Campus} + \gamma_{zip} + \nu_{jt} \tag{5}$$

$$FieldExam_{jt} = \alpha_5 + \beta_{12} FTE_t^{Field} + \beta_{13} FTE_t^{Field} * FTE_t^{Field} + \gamma_{zip} + \nu_{jt} \tag{6}$$

**Second Stage**. The second stage involves regressing the probability of taxpayer compliance outcomes ($P_{ijt}$) on the predicted values from the first stage. $P_{ijt}$ is categorized as follows:

- $P_{ijt} = 0$: Fully compliant (filed and paid on time).
- $P_{ijt} = 1$: Filed late but no outstanding balance (paid in full).
- $P_{ijt} = 2$: Has an outstanding balance due at the end of time *t*.

The second stage model is specified as follows:

$$P_{ijt} = F\left(\alpha + \beta_1 \widehat{ACS}_{jt-1} + \beta_2 \widehat{CP59}_{jt-1} + \beta_3 \widehat{Field}_{jt-1} + \beta_4 \widehat{Campus}_{jt-1} + \beta_5 \widehat{FieldExam}_{jt-1} + \sum_k \theta_k X_{ijt-1} + \gamma_{zip} + \eta_{year}\right) + e_{ijt} \quad (7)$$

Here, $F(\cdot)$ represents the multinomial logit link function. The predicted values from the first stage ($\widehat{ACS}_{jt-1}$, $\widehat{CP59}_{jt-1}$ $\widehat{Field}_{jt-1}$, $\widehat{Campus}_{jt-1}$, and $\widehat{FieldExam}_{jt-1}$) are used as independent variables along with control variables ($X_{(ijt-1)}$), ZIP Code ($\gamma_{zip}$) and year ($\eta_{year}$) fixed effects. This setup leverages administrative FTE allocations as instruments, allowing us to derive causal insights on the effects of enforcement actions while effectively controlling for potential biases from time-invariant regional and temporal factors.

**Additional Control Variables**. We include a comprehensive set of taxpayer characteristics based on the most recent return filed in previous year (*t*-1), which is represented in $X_{(ijt-1)}$. These control variables are fully listed in Appendix Table A1 and account for differences in compliance behavior, income, and risk characteristics, and include:

- Indicator for married filing jointly status.=
- Log transform of total positive income.=
- Indicator for filing on-time consecutively for the last 4 years.=
- Indicator for having a balance due (from line 37 of Form 1040, before remittance).=
- Under-withholding as a percent of total positive income (balance due/ total positive income), restricted to = between -100 and 100%.
- Indicator for 50% or more of income derived from sources that cannot withhold taxes, such as self-employment income.
- Indicators for activity code/audit class indicators and their interactions with the Discriminant Function (DIF) = score.

These measures help capture the taxpayer's risk profile, with the DIF score serving as a proxy for reporting compliance and the likelihood of an audit, allowing us to control for potential selection biases. The DIF score is uniquely defined by the activity class of the return, so we also include indicator variables for these classes and interaction terms between them and the DIF score. Because our taxpayers were selected based on being compliant in the prior year, they aren't directly treated by a filing and payment compliance program, and there is no direct measure of the impact of compliance programs in our model. For this population we aim to measure only an indirect impact of these programs.

### 3.3.2 Linear Model for Change in Balance Due

Our sample of taxpayers begins each year *t* with no outstanding balance due. While most taxpayers will maintain this status throughout the year, some will fall out of compliance and receive a balance due notice. For these individuals, we model the indirect effects of compliance programs on the change in their outstanding balance. This approach allows us to evaluate whether these programs can positively influence compliance by reducing the size of the taxpayer's debt, even if

they do not entirely prevent noncompliance. For filers who receive a balance due notice after underpaying, we use the total balance shown on the initial notice. For nonfilers, we calculate the balance due based on the information returns provided to the IRS. This framework enables us to estimate the intensive margin for taxpayers who do not remain fully compliant.

We define our outcome variable for the change in balance due, $U_{ijt}$ as follows:

- For taxpayers with $P_{ijt}$=2, is the amount of tax not timely filed and paid. For filers, this is the total balance due on the first notice sent to the taxpayer. For nonfilers, it is the balance due on a potential substitute for return (SFR) generated through the Case Creation Nonfiler Identification Process=
- Otherwise,$U_{ijt}$ =0 (late filers with $P_{ijt}$ =1 or compliant taxpayers with $P_{ijt}$=0)

Our 1% sample of compliant taxpayers includes about 11.6 million observations, but only around 350,000 (3%) of them ended up with an outstanding tax debt ($U_{ijt}>0$). Because this number is insufficient for robust analysis across the comprehensive set of ZIP Codes and multiple years used in our two-stage approach, we employ an alternative 10% sample for the balance due model. This adjustment increases the sample size to approximately 3.5 million previously compliant taxpayers who later accrued outstanding tax debts.

We use this 10% sample to run a linear model for the change in balance due, focusing on the 3.5 million taxpayers with an outstanding balance. The dependent variable, $U_{ijt}$, is log-transformed to mitigate bias caused by skewed unpaid tax amounts with extreme outliers. We calculate $U_{ijt}$ using tax year $t$-1, which is filed in year $t$. Following the two-stage approach outlined in models (2)-(7), we run the following ordinary least squares (OLS) regression for taxpayers with $P_{ijt}$=2:

$$\log(U_{ijt}) = \alpha + \beta_1 \widehat{ACS}_{jt-1} + \beta_2 \widehat{CP59}_{jt-1} + \beta_3 \widehat{Field}_{jt-1} + \beta_4 \widehat{Campus}_{jt-1} + \beta_5 \widehat{FieldExam}_{jt-1} + \sum_k \theta_k X_{ijt-1} + \gamma_{zip} + \eta_{year} + e_{ijt} \quad (8)$$

Similar to model (7), model (8) predicts the unpaid assessment amount $U_{ijt}$ for non-compliant taxpayer $i$ in ZIP Code $j$ in year $t$. It is regressed on the predicted values of endogenous variables, $\widehat{ACS}_{jt-1}$,$\widehat{CP59}_{jt-1}$, $\widehat{Field}_{jt-1}$ , $\widehat{Campus}_{jt-1}$ , and $\widehat{FieldExam}_{jt-1}$ with the same control variables $X_{(ijt-1)}$ as in model (5), along with ZIP Code ($\gamma_{zip}$) and year ($\eta_{year}$) fixed effects to account for omitted variables that may influence $U_{ijt}$. Using the two-stage process in model (8) also has the same advantages as with model (7), better controlling for unobserved factors through the incorporation of fixed effects and addressing potential endogeneity between the ZIP Code-level indirect treatments and $U_{ijt}$ using the IRS enforcement budget as an instrumental variable.

Further tests confirm that the SCI-based model outperforms alternative models that rely on simple geographical distances or unweighted counts of enforcement actions to replace the SCI-based treatment variables. Comprehensive results and comparisons from these additional model tests, presented in the Appendix Table A2, substantiate the effectiveness of using the SCI to capture the indirect effects of IRS enforcement strategies.

# 4. Results

## *4.1 Two-Stage Multinomial Logistic Model for Filing and Payment Compliance*

### 4.1.1 Model Results and Interpretation

Our two-stage model utilizes FTE allocations as instrumental variables in the first stage to predict enforcement variables, followed by a second stage that models compliance outcomes based on these predicted values. Table 3 presents the first stage regression results, which show that FTE allocations positively affect the number of enforcements, with a strong model fit indicated by the R-squared and F-statistics. The negative interaction term between collection FTEs reflects their interchangeable allocation, while the quadratic terms for exam FTEs suggest diminishing returns, consistent with typical labor input-output relationships.

**TABLE 3. First-Stage Regression Results for Two-State Least Squares Model**

| Variable | ACS (1) | CP 59 (2) | Field Collection (3) | Campus Exam (4) | Field Exam (5) |
|---|---|---|---|---|---|
| Intercept | -312.2** (8.138) | -213.4** (7.325) | 2.879** (0.378) | -47.49** (3.397) | -0.989 (0.942) |
| ACS FTE | 0.129** (0.002) | 0.081** (0.001) | 0.0007** (0.00006) | - | - |
| Field FTE | 0.076** (0.001) | 0.029** (0.000) | 0.0015** (0.00003) | - | - |
| ACS FTE X Field FTE | -0.019** (0.0004) | -0.005** (0.0002) | -0.0003** (0.00001) | - | - |
| Campus FTE | - | - | - | 0.043** (0.000) | - |
| Campus $FTE^2$ | - | - | - | -0.053** (0.0003) | - |
| Field Exam FTE | - | - | - | - | 0.001** (0.00004) |
| Field Exam $FTE^2$ | - | - | | - | -0.0001** (0.00001) |
| R-squared | 0.821 | 0.816 | 0.911 | 0.900 | 0.912 |
| F-statistic | 211.0 | 49.67 | 152.1 | 124.0 | 76.26 |

Note: N=185,593. Standard errors in parentheses. ** indicate significance levels of $p < 0.05$, and $p < 0.01$, respectively. All models include ZIP Code fixed effects.

Table 4 presents the results from the multinomial compliance model. The findings align with intuitive expectations—positive coefficients in the compliant category (=0) suggest that increased enforcement efforts improve compliance, while negative coefficients for the non-compliant categories (=1 and =2) indicate a reduction in the likelihood of noncompliance. Among the three enforcement programs, ACS letters demonstrate the strongest influence, followed by CP59 notices. Although the sample consists of generally compliant taxpayers, the model reveals that increased enforcement—especially through ACS letters—has a significant preventative effect, enhancing voluntary compliance rates. CP59 notices similarly contribute to compliance improvements, though to a lesser extent than ACS letters. Field collection interventions, while impactful, exhibit a more modest effect in comparison to the other two programs.

**TABLE 4. Selected Parameter Estimates for Two-Stage Multinomial Compliance Model**

($P_{ijt}$=0: compliant, 1: noncompliant no balance due, 2: noncompliant with balance due)

| Variable | $P_{ijt}$=0 (1) | $P_{ijt}$ 1 (2) | $P_{ijt}$=2 (3) |
|---|---|---|---|
| Intercept | 0.211** (0.016) | -0.074** (0.026) | -0.137** (0.022) |
| ACS weighted average | 5.539** (0.005) | -2.108** (0.009) | -3.431** (0.008) |
| CP59 weighted average | 3.155** (0.003) | -1.202** (0.005) | -1.954** (0.004) |
| Field collection weighted average | 0.278** (0.000) | -0.102** (0.000) | -0.176** (0.000) |
| Campus exam weighted average | 1.036** (0.001) | -0.371** (0.001) | -0.665** (0.001) |
| Field exam weighted average | 0.243** (0.000) | -0.089** (0.000) | -0.154** (0.000) |

**TABLE 4. Selected Parameter Estimates for Two-Stage Multinomial Compliance Model (Continued)**

($P_{ijt}$=0: compliant, 1: non-compliant no balance due, 2: non-compliant with balance due)

| Variable | $P_{ijt}$ =0 (1) | $P_{ijt}$ 1 (2) | $P_{ijt}$ =2 (3) |
|---|---|---|---|
| Married filing jointly | 0.186** (0.002) | -0.179** (0.004) | -0.007** (0.003) |
| Log total positive income | -0.128** (0.001) | 0.011** (0.002) | 0.117** (0.001) |
| Timely filed in past four years | 0.591** (0.002) | -0.322** (0.003) | -0.269** (0.003) |
| Balance due (before remittance) | -0.136** (0.003) | -0.049** (0.005) | 0.185** (0.004) |
| Percent of income underwithheld | -1.478** (0.010) | -0.142** (0.016) | 1.620** (0.015) |
| 50% income not subject to withholding | -0.101** (0.003) | -0.008 (0.005) | 0.110** (0.005) |

Note: N=10,246,313. Standard errors in parentheses. ** indicate significance levels of p < 0.05, and p < 0.01, respectively. All models include year and ZIP Code fixed effects.

### 4.1.2 Impact Analysis

The ACS intervention demonstrates the most substantial influence on compliance among the programs studied. Over the study period from 2011 to 2019, ACS notices were sent to approximately 45,000 ZIP Codes. In comparison, CP59 notices and field collections were administered to around 40,000 and 33,000 ZIP Codes, respectively. Additionally, the frequency of ACS treatments per ZIP Code significantly outpaces that of CP59 and field visits. On average, 173 ACS letters were sent per ZIP Code annually, compared to 113 CP59 notices and just 11 field visits per ZIP Code each year.

The disparity in both the breadth and intensity of enforcement efforts leads to differing impacts across programs. Our findings emphasize that the wide reach and frequent interactions of the ACS program are particularly effective in enhancing voluntary compliance. These indirect effects, which spread through social networks, extend the impact of enforcement actions beyond directly treated individuals. By ensuring compliance programs have sufficient resources to contact taxpayers, the IRS can amplify the spread of compliant behavior across a wider population. In contrast, direct effects are limited to those directly treated and follow a different dynamic. For instance, while field visits are more limited in scope, they may exert a stronger direct effect due to the intensity of in-person contact, prompting immediate compliance.

### 4.1.3 Average Marginal Effects and Impact of Increased Enforcement Levels

The average marginal effects from the multinomial model, shown in Table 5, convert log odds into probabilities, offering a clearer interpretation of the enforcement programs' impact on compliance. An increase of 1,000 ACS letters leads to significant reductions in both late filings and delinquencies, indicating substantial improvements in compliance. Similarly, increases in CP59 notices and field visits also lower noncompliance rates, though to a lesser degree. The results highlight the varying effectiveness of these enforcement tools, with ACS letters proving to be particularly powerful in fostering taxpayer compliance.

Table 6 expands on these findings by showing the marginal effects of a 10% increase in each program's enforcement levels. A 10% increase in ACS letters is associated with a 0.3 percentage point decrease in late filings and a 0.5 percentage point decrease in delinquencies, corresponding to 15% and 17% reductions, respectively. CP59 notices also yield positive effects, with a 10% increase reducing late filings by 0.1 percentage points (5% improvement) and delinquencies by 0.2 percentage points (6% decrease). Field collection visits have a more modest effect, highlighting that while effective, their reach is more limited compared to the broader, more frequent ACS letters and CP59 notices.

**TABLE 5. Average Marginal Effects for Two-Stage Multinomial Compliance Model**

($P_{ijt}$=0: compliant, 1: noncompliant no balance due, 2: noncompliant with balance due)

| Variable | $P_{ijt}$=0 (1) | $P_{ijt}$=1 (2) | $P_{ijt}$=2 (3) |
|---|---|---|---|
| ACS weighted average | 0.397** (0.003) | -0.149** (0.001) | -0.249** (0.002) |
| CP59 weighted average | 0.226** (0.001) | -0.085** (0.000) | -0.142** (0.001) |
| Field collection weighted average | 0.020** (0.000) | -0.007** (0.000) | -0.013** (0.000) |
| Campus exam weighted average | 0.074** (0.000) | -0.027** (0.000) | -0.047** (0.000) |
| Field exam weighted average | 0.017** (0.000) | -0.006** (0.000) | -0.011** (0.000) |
| Married filing jointly | 0.012** (0.000) | -0.007** (0.000) | -0.005* (0.000) |
| Log total positive income | -0.009** (0.000) | 0.003** (0.000) | 0.007** (0.000) |
| Timely filed in past four years | 0.049** (0.000) | -0.021** (0.000) | -0.028** (0.000) |
| Balance due (before remittance) | -0.011** (0.000) | 0.002** (0.000) | 0.010** (0.000) |
| Percent of income under-withheld | -0.112** (0.001) | 0.025** (0.000) | 0.087** (0.002) |
| 50% income not subject to withholding | -0.008** (0.000) | 0.002** (0.000) | 0.006** (0.000) |

Note: N=10,246,313. Standard errors in parentheses. * and ** indicate significance levels of $p < 0.05$ and $p < 0.01$, respectively. All models include year and ZIP Code fixed effects.

**TABLE 6. Marginal Effect Estimates for 10% Increase in Program Levels**

| Compliance Program | Δ Probability for Late Filers | Δ Probability for Delinquent Cases |
|---|---|---|
| ACS Letters | -0.3 | -0.5 |
| CP59 Notices | -0.1 | -0.2 |
| Field Collection | -0.001 | -0.002 |
| Campus Exam | -0.02 | -0.03 |
| Field Exam | -0.0008 | -0.001 |

Our results confirm the differential effectiveness of IRS compliance programs. ACS letters, due to their broad distribution and frequency, are especially potent in encouraging taxpayer compliance. In contrast, CP59 notices and field visits, while effective, have a more limited reach. These findings suggest that strategic resource allocation focusing on extensive and frequent outreach, particularly through ACS, is critical for enhancing voluntary compliance. Policymakers can use these insights to optimize enforcement efforts and refine program designs for greater efficiency.

## *4.2 Linear Model for Change in Balance Due*

### 4.2.1 OLS Results

Table 7 shows the OLS results for the change in outstanding balance due, shown in Equation (8). Table 7 reveals patterns consistent with our findings from the filing and payment compliance models. Specifically, the parameter estimates for

ACS letters and CP59 notices are negative and statistically significant, indicating that enforcement actions contribute to reducing unpaid tax balances, even when they do not prevent taxpayers from becoming delinquent altogether.

The results indicate that ACS letters exert the greatest impact in reducing the outstanding balance due, followed by CP59 notices. Field collection interventions, while statistically significant, show only a marginal effect, with significance at the 10% level. This suggests that although field visits are a more intensive enforcement action and may generate substantial direct effects, their overall indirect impact on reducing balances is minimal compared to the broader influence of ACS letters and CP59 notices. These findings underscore the effectiveness of widespread, less resource-intensive interventions in mitigating delinquent balances through indirect channels.

**TABLE 7. Selected Parameter Estimates for Linear Model of Change in Balance Due**

| Variable | $Log(U_{ijt})$ |
|---|---|
| Intercept | 5.525** (0.017) |
| ACS weighted average | -0.029** (0.006) |
| CP59 weighted average | -0.015** (0.003) |
| Field collection weighted average | -0.001* (0.000) |
| Campus exam weighted average | -0.000 (0.001) |
| Field exam weighted average | -0.001** (0.000) |
| Married filing jointly | 0.068** (0.002) |
| Log total positive income | 0.202** (0.001) |
| Timely filed in past four years | -0.177** (0.002) |
| Balance due (before remittance) | -0.184** (0.003) |
| Percent of income under-withheld | 0.334** (0.011) |
| 50% income not subject to withholding | 0.031** (0.003) |

Notes: N=3,286,146. Standard errors in parentheses. * and ** indicate significance levels of $p < 0.05$ and $p < 0.01$, respectively. All models include year and ZIP Code fixed effects.

## *4.3 Nationwide Delinquent Balances and Enforcement Impacts*

### 4.3.1 Nationwide Estimates of Delinquent Balances

To estimate the nationwide balance due for previously compliant taxpayers who became delinquent, we employ two complementary approaches, both of which yield consistent estimates of approximately $19.6 billion in yearly delinquent balances for the period 2011-2019. The first approach uses a 10% sample of taxpayers who were compliant at the start of the year but ended the year with a delinquent balance. The aggregated balance from this sample is scaled up to represent the entire population:

$$One\ Year\ Delinquent\ Balance = \left(\sum_i balance_i\right) * \frac{10}{9} \tag{9}$$

The second approach leverages a 1% sample of taxpayers to estimate the national delinquent balance by combining three components: the average number of compliant taxpayers ($TCP \approx 129$ million), the probability of transitioning to delinquency ($P_{ijt}=2$, ≈ 3%), and the expected balance among delinquents ($E[Balance\ P_{ijt}=2] \approx \$5{,}000$):

$$One\ Year\ Delinquent\ Balance\ = TCP * P_2 * E\left[Balance \middle| P_{ijt} = 2\right] \tag{10}$$

Both approaches provide independent but consistent estimates of the annual nationwide delinquent balance for taxpayers who were compliant at the start of the year.

### 4.3.2 Total National Impact of Enforcement Actions

To quantify the effect of enforcement actions, we estimate the Total National Impact (TNI) of interventions, including ACS letters, CP59 notices, and field collections. The TNI is calculated by combining the changes in extensive and intensive margins:

$$TNI = TCP * ( \Delta P_2 * E[Balance\ P_{ijt}=2] + P_2 * \Delta E[Balance\ P_{ijt}=2]), \tag{11}$$

where:

- $\Delta P_2$ represents the change in the probability of delinquency, derived from the multinomial logit regression.
- $\Delta E[Balance\ P_{ijt}=2]$ represents the change in the expected balance among delinquents, derived from our OLS regression.

The change in the intensive margin, $\Delta E[Balance\ P_{ijt}=2]$, is calculated using the coefficient from the OLS regression of $\log(U_{ijt})$ on enforcement actions:

$$\Delta E[Balance\ P_{ijt}=2]) = E[Balance\ P_{ijt}=2] * (e^{\beta}-1) \tag{12}$$

Applying this framework, a 10% increase in ACS interventions leads to an estimated $3.2 billion reduction in newly created delinquent balance, representing a 16% decrease. This estimate applies specifically to taxpayers who were fully compliant in the prior year but became delinquent in the current year. The impact reflects both a reduced probability of transitioning into delinquency, and a reduction in the amount of unpaid balance accrued by those who do become delinquent. A similar 10% increase in CP59 notices results in a $1.3 billion decrease (7%), while a 10% increase in field visits yields a much smaller reduction of $11 million (0.06%).

These results demonstrate the efficacy of broad and frequent interventions, such as ACS and CP59 notices, in reducing outstanding balances. In contrast, field collections—despite their direct and intensive nature—have limited indirect impact on individual taxpayer balances. It is noteworthy that field collections are likely more impactful for business taxpayers, consistent with Boning et al. (2019).

### 4.3.3 Confidence Intervals and Delta Method

We compute the confidence intervals (CIs) for TNI using the delta method, which provides a first-order approximation of variance for nonlinear functions of estimated parameters. Specifically, the variance of TNI is expressed as:

$$Var(TNI) = (TCP * E[Balance\ P_{ijt}=2] * SE_{(\Delta P\_2)})^2 + (TCP * P_2 * SE_{\Delta E[Balance|\ P_{ijt}=2]})^2$$

where:

- $SE_{(\Delta P_2)}$ is the standard error of $\Delta P_2$
- $SE_{(\Delta E[Balance|\bar{P}_{ijt=2}])}$ is the standard error of $\Delta E[Balance \mid P_{ijt}=2]$

Using the variance, the 95-percent confidence interval for TNI is calculated as:

$$CI_{TNI} = TNI \pm 1.96\sqrt{Var(TNI)}$$

**TABLE 8. Reductions in Delinquent Balances from a 10% Increase in Enforcement**

| Compliance Program | Dollar reduction ($B) | Percentage Reduction (%) |
|---|---|---|
| ACS Letters | -3.18 [-3.24, -3.12] | -16.2 [-16.5, -15.9] |
| CP59 Notices | -1.34 [-1.36, -1.33] | -6.85 [-6.93, -6.77] |
| Field Collection | -0.01 [-0.01, -0.01] | -0.06 [-0.06, -0.06] |
| Campus Exam | -0.172 [-0.173, -0.172] | -0.88 [-0.88, -0.88] |
| Field Exam | -0.009 [-0.009, -0.009] | -0.04 [-0.04, -0.04] |

## 5. Conclusion

This study offers a rigorous approach to estimate voluntary compliance effects of IRS enforcement strategies, focusing on ACS letters, CP59 notices, and field collection interventions and their influence on filing and payment compliance for individual taxpayers. By employing a two-stage multinomial logistic model in combination with the SCI, our analysis underscores the significant role these programs play in maintaining and enhancing voluntary compliance, particularly through their indirect effects across various taxpayer segments.

Our results indicate that ACS letters have the most pronounced impact on promoting voluntary compliance filing and payment obligations. This is reflected by their extensive coverage, reaching approximately 45,000 ZIP Codes and averaging 173 letters per ZIP Code annually. In contrast, CP59 notices and field collections, though impactful, show less influence, indicating their relatively narrower reach and lower frequency of interaction. Our findings emphasize the importance of strategic outreach, where programs with broad reach and consistent interaction are notably effective in fostering voluntary compliance through indirect channels.

Furthermore, the analysis of average marginal effects emphasizes the substantial benefits of even modest increases in enforcement. A 10% increase in ACS letters is associated with significant reductions in both late filings and delinquencies, demonstrating the effectiveness of widespread, targeted enforcement actions. This suggests that a well-distributed approach can yield meaningful improvements in taxpayer behavior, enhancing overall compliance rates.

A key insight from this study is the heterogeneity in enforcement effectiveness across different regions, shaped in part by social dynamics. Our findings suggest that compliance responses to enforcement actions tend to be stronger in areas with higher levels of social connectedness, implying that community networks may facilitate the transmission of compliance-related information and behavioral norms. Understanding these dynamics can help inform broader discussions on optimizing enforcement strategies without altering fundamental allocation principles.

From an economic perspective, the fiscal impact of enhanced enforcement is substantial. Our analysis, utilizing a combination of multinomial logit and linear regression models, reveals that a 10% increase in ACS interventions is linked to a $3.2 billion reduction in newly accrued delinquent balances among previously compliant taxpayers, an approximately 16% decrease. Similarly, a 10% increase in CP59 notices yields a $1.3 billion reduction (7%), while a comparable increase in field visits produces a $11 million reduction (0.06%). These figures highlight the significant fiscal returns that can be realized through strategic improvements in resources for IRS enforcement activities. These indirect effects are in addition to the substantial direct treatment effects of filing and payment compliance programs.

This study underscores the critical role of indirect effects in IRS enforcement strategies and provides actionable insights for policymakers to refine program designs. We plan to extend our approach to estimate indirect effects for business taxpayers. Future research and policy efforts should continue to explore these dynamics to deepen our understanding of enforcement spillover effects and inform the development of evidence-based compliance strategies.

## References


Bailey, Michael, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. "Social Connectedness: Measurement, Determinants, and Effects." *Journal of Economic Perspectives* vol. 32, no. 3, pp. 259–280 (2018).

Bloomquist, Kim. "Incorporating Indirect Effects in Audit Case Selection: An Agent-Based Approach." *Internal Revenue Service* (2012).

Boning, William C., John Guyton, Ronald Hodge, and Joel Slemrod. "Heard it through the grapevine: The direct and network effects of a tax enforcement field experiment on firms." *Journal of Public Economics* vol. 190, issue C (2019).

Datta, Saurabh, Stacy Orlett, and Alex Turk. "Individual Nonfilers and IRS-Generated Tax Assessments: Revenue and Compliance Impacts of IRS Substitute Assessments When Taxpayers Don't File." *Internal Revenue Service* (2015).

Greenwood, Shannon, Andrew Perrin, and Maeve Duggan. "Social Media Update 2016." Retrieved from https://www.pewresearch.org/internet/2016/11/11/social-media-update-2016/. *Pew Research Center*, (2016).

López-Luzuriaga, Andrea, and Carlos G. Scartascini. "Willing but Unable to Pay?: The Role of Gender in Tax Compliance." IDB Publications (Working Paper) 12983, *Inter-American Development Bank* (2023).

Internal Revenue Service. "Data Book, 2010." *Publication 55-B*, Washington, DC (March 2011).

Internal Revenue Service. "Data Book, 2019." *Publication 55-B*, Washington, DC (June 2020).

Internal Revenue Service. "Data Book, 2023." *Publication 55-B*, Washington, DC (April 2024).

Internal Revenue Service. "Tax Gap Projections for Tax Year 2022." *Publication 5869 (Rev. 10-2024)*, Washington, DC (October 2024).

Internal Revenue Service. "Understanding Your CP59 Notice." Retrieved from https://www.irs.gov/individuals/understanding-your-cp59-notice. Accessed January 2025.

Internal Revenue Service. "Understanding Your LT11 Notice." Retrieved from https://www.irs.gov/individuals/understanding-your-lt11-notice-or-letter-1058. Accessed January 2025.

Internal Revenue Service. "Understanding Your LT16 Notice." Retrieved from https://www.irs.gov/individuals/understanding-your-lt16-notice. Accessed January 2025.

Internal Revenue Service. "Understanding Your LT26 Notice." Retrieved from https://www.irs.gov/individuals/understanding-your-lt26-notice. Accessed January 2025.

Internal Revenue Service. "IRS Ends Unannounced Revenue Officer Visits to Taxpayers." Retrieved from https://www.irs.gov/newsroom/irs-ends-unannounced-revenue-officer-visits-to-taxpayers. Accessed January 2025.

Treasury Inspector General for Tax Administration (TIGTA). "Trends in Compliance Activities Through Fiscal Year 2022." *Report No*. 2024-300-011 (2023).

Turk, Alex, and Terry Ashley. "Accounts Receivable Resolution and the Impact of Lien Filing Policy on Sole Proprietor Businesses." *2002 Federal Forecasters Conference Proceedings*, pp. 323–332 (2002).

Turk, Alex, John Iuranich, Stacy Orlett, and Saurabh Datta. "Resolving Unpaid Taxes and the Notice of Federal Tax Lien: Evidence from the Fresh Start Initiative." *Internal Revenue Service* (2016).

# Appendix

This appendix presents additional information on the datasets constructed for the analysis and full regression results.

## TABLE A1. Variable Descriptions

| Name | Description |
|---|---|
| Time Trend | Linear trend line, increases by one each year |
| CP59 coverage rate | Total number of taxpayers receiving CP59 notices for year *t*-1 divided by the total number of taxpayers with delinquent accounts (balance due in collections data) in year *t*-1 |
| ACS letter coverage rate | Total number of taxpayers receiving ACS letters LT11, LT16, or LT26 for year *t*-1 divided by the total number of taxpayers with delinquent accounts (balance due in collections data) in year *t*-1 |
| Field coverage rate | Total number of taxpayers in field collection status at any point in year *t*-1 divided by the total number of taxpayers with delinquent accounts (balance due in collections data) in year *t*-1 |
| Married filing jointly | Indicator for married filing jointly filing status on most recent return, filed in year *t*-1 |
| Log total positive income | Natural log transformation of total positive income (amount of income excluding losses) from most recent return, filed in year *t*-1 |
| Timely filed in past four years | Indicator for taxpayers who fully paid and filed timely in the four most recent years, including years *t*-1, *t*-2, *t*-3, and *t*-4 |
| Balance due (before remittance) | Indicator for taxpayers who had an amount greater than or equal to $100 on the "Amount you owe" line from the most recent return, filed in year *t*-1 |
| % of income underwithheld | Ratio of balance due amount ("Amount you owe" line) to total positive income from most recent return, filed in year *t*-1, capped at -1 (cases with refunds equal or greater than total positive income) and 1 (cases with balance due on filing greater than or equal to total positive income) |
| 50% or more of income not subject to withholding | Indicator for taxpayers with a ratio of income not subject to withholding (e.g., farm income from Schedule F, business income from Schedule C, etc.) to total income greater than 0.5 for the most recent return, filed in year *t*-1 |
| Activity code 266 | Indicator for taxpayers in activity code (examination class) 266 (Forms 1040PR/1040SS) on the most recent return, filed in *t*-1 |
| Activity code 270 | Indicator for taxpayers in activity code (examination class) 270 (returns with earned income tax credit, total positive income below $200,000 and Schedule C/F gross receipts below $25,000 or not present) on the most recent return, filed in *t*-1 |
| Activity code 271 | Indicator for taxpayers in activity code (examination class) 271 (returns with earned income tax credit, total positive income below $200,000 and Schedule C/F gross receipts $25,000 or more) on the most recent return, filed in *t*-1 |
| Activity code 272 | Indicator for taxpayers in activity code (examination class) 272 (returns with no earned income credit, total positive income below $200,000 and no Schedule C/E/F or Form 2106) on the most recent return, filed in *t*-1. |
| Activity code 273 | Indicator for taxpayers in activity code (examination class) 273 (returns with no earned income credit, total positive income below $200,000 and with Schedule E or Form 2106 but no Schedule C/F) on the most recent return, filed in *t*-1 |
| Activity code 274 | Indicator for taxpayers in activity code (examination class) 274 (returns with no earned income credit, total positive income below $200,000 and nonfarm business with Schedule C/F receipts below $25,000) on the most recent return, filed in *t*-1 |
| Activity code 275 | Indicator for taxpayers in activity code (examination class) 275 (returns with no earned income credit, total positive income below $200,000 and nonfarm business with Schedule C/F receipts $25,000-$99,999) on the most recent return, filed in *t*-1 |

| Name | Description |
|---|---|
| Activity code 276 | Indicator for taxpayers in activity code (examination class) 276 (returns with no earned income credit, total positive income below $200,000 and nonfarm business with Schedule C/F receipts $100,000-$199,999) on the most recent return, filed in *t*-1 |
| Activity code 277 | Indicator for taxpayers in activity code (examination class) 277 (returns with no earned income credit, total positive income below $200,000 and nonfarm business with Schedule C/F receipts $200,000 or more) on the most recent return, filed in *t*-1 |
| Activity code 278 | Indicator for taxpayers in activity code (examination class) 278 (returns with no earned income credit, total positive income below $200,000 and farm business not classified elsewhere) on the most recent return, filed in *t*-1 |
| Activity code 279 | Indicator for taxpayers in activity code (examination class) 279 (returns with no earned income credit, with Schedule C/F and total positive income $200,000-$999,999) on the most recent return, filed in *t*-1 |
| Activity code 280 | Indicator for taxpayers in activity code (examination class) 280 (returns with no earned income credit, no Schedule C/F and total positive income $200,000-$999,999) on the most recent return, filed in *t*-1 |
| Activity code 281 | Indicator for taxpayers in activity code (examination class) 281 (returns with no earned income credit and total positive income $1,00,000 or more) on the most recent return, filed in *t*-1. Note that activity code 281 is dropped from the models and serves as the reference category for the series of activity code indicator variables |
| Activity code*DIF | Interaction term for each activity code indicator and the Discriminant Index Function (DIF) score, which ranks the likelihood of tax changes for taxpayers in the event of an audit and is modeled separately for each activity code. The DIF score can take on positive and negative values, and may be thought of as a risk indicator, but only has meaning in context |
| Year X | Dummy variable for year X |
| CP59 weighted average | Used in two-stage models as an alternative for CP59 coverage rate, number of CP59 notices in a specific ZIP Code, weighted by SCI index, distance, or unweighted, as described in equation (3). For the unweighted models, a log transformation is applied to address skewness |
| ACS weighted average | Used in two-stage models as an alternative for ACS letter coverage rate, number of ACS letters in a specific ZIP Code, weighted by SCI index, distance, or unweighted, as described in equation (2). For the unweighted models, a log transformation is applied to address skewness |
| Field collection weighted average | Used in two-stage models as an alternative for field coverage rate, number of taxpayers in field collection in a specific ZIP Code, weighted by SCI index, distance, or unweighted, as described in equation (4). For the unweighted models, a log transformation is applied to address skewness |
| ZIP Code X | Dummy variable for ZIP Code X (parameter estimates not shown, as ZIP Codes number in the tens of thousands) |

**TABLE A2. Full Parameter Estimates for Two-Stage Logistic Compliance Model**

Response Variable: $P_{ijt}$ (0: compliant, 1: non-compliant)

| Variable | SCI Weighted | Distance Weighted | Unweighted |
|---|---|---|---|
| Intercept | -4.762** (0.027) | -5.236** (0.029) | -4.744** (0.008) |
| ACS weighted average | -1.367** (0.009) | -0.081** (0.007) | -0.037** (0.002) |
| CP59 weighted average | -0.753** (0.005) | -0.039** (0.004) | -0.033** (0.002) |
| Field collection weighted average | -0.066** (0.000) | -0.002** (0.000) | -0.015** (0.002) |
| Married filing jointly | -0.260** (0.003) | -0.243** (0.004) | -0.237** (0.003) |
| Log total positive income | 0.236** (0.001) | 0.235** (0.002) | 0.233** (0.002) |
| Timely filed in past four years | -0.884** (0.003) | -0.873** (0.003) | -0.876** (0.003) |
| Balance due (before remittance) | 0.232** (0.004) | 0.232** (0.004) | 0.235** (0.004) |
| % of income under-withheld | 2.576** (0.016) | 2.603** (0.016) | 2.582** (0.016) |
| 50% income not subject to withholding | 0.210** (0.005) | 0.202** (0.005) | 0.208** (0.005) |
| Activity code 266 | 0.291** (0.094) | 0.588 (0.425) | 0.105 (0.088) |
| Activity code 270 | 0.830** (0.025) | 0.520** (0.024) | 0.509** (0.023) |
| Activity code 271 | 1.029** (0.039) | 0.840** (0.039) | 0.829** (0.038) |
| Activity code 272 | 0.543** (0.024) | 0.249** (0.023) | 0.242** (0.022) |
| Activity code 273 | 0.632** (0.025) | 0.339** (0.023) | 0.340** (0.022) |
| Activity code 274 | 1.002** (0.024) | 0.700** (0.023) | 0.694** (0.022) |
| Activity code 275 | 1.068** (0.027) | 0.747** (0.027) | 0.739** (0.026) |
| Activity code 276 | 1.106** (0.055) | 0.458** (0.059) | 0.437** (0.056) |
| Activity code 277 | 1.419** (0.055) | 0.804** (0.059) | 0.747** (0.057) |
| Activity code 278 | 0.655** (0.029) | 0.358** (0.029) | 0.358** (0.029) |
| Activity code 279 | 0.416** (0.026) | 0.083** (0.025) | 0.079** (0.024) |
| Activity code 280 | 0.779** (0.026) | 0.457** (0.025) | 0.446** (0.024) |
| Activity code 266*DIF | 0.005** (0.000) | -0.001 (0.002) | 0.003** (0.000) |
| Activity code 270*DIF | 0.001** (0.000) | 0.001** (0.000) | 0.001** (0.000) |
| Activity code 271*DIF | 0.002** (0.000) | 0.002** (0.000) | 0.002** (0.000) |
| Activity code 272*DIF | 0.001** (0.000) | 0.001** (0.000) | 0.001** (0.000) |
| Activity code 273*DIF | 0.001** (0.000) | 0.001** (0.000) | 0.001** (0.000) |
| Activity code 274*DIF | 0.002** (0.000) | 0.001** (0.000) | 0.001** (0.000) |
| Activity code 275*DIF | 0.001** (0.000) | 0.001** (0.000) | 0.001** (0.000) |
| Activity code 276*DIF | 0.001** (0.000) | 0.002** (0.000) | 0.002** (0.000) |
| Activity code 277*DIF | 0.000** (0.000) | 0.001** (0.000) | 0.001** (0.000) |
| Activity code 278*DIF | 0.000 (0.000) | 0.000 (0.000) | 0.000 (0.000) |
| Activity code 279*DIF | 0.001** (0.000) | 0.001** (0.000) | 0.001** (0.000) |
| Activity code 280*DIF | 0.001** (0.000) | 0.001** (0.000) | 0.001** (0.000) |
| Year 2012 | -0.134** (0.006) | -0.069** (0.007) | -0.087** (0.006) |
| Year 2013 | -0.242** (0.006) | -0.075** (0.006) | -0.112** (0.006) |
| Year 2014 | -0.371** (0.006) | -0.051** (0.006) | -0.120** (0.006) |
| Year 2015 | -0.342** (0.006) | 0.012* (0.006) | -0.064** (0.007) |

**TABLE A2. Full Parameter Estimates for Two-Stage Logistic Compliance Model (Continued)**

Response Variable: $P_{ijt}$ (0: compliant, 1: non-compliant)

| Variable | SCI Weighted | Distance Weighted | Unweighted |
|---|---|---|---|
| Year 2016 | -0.278** (0.005) | 0.121** (0.006) | 0.036** (0.007) |
| Year 2017 | -0.265** (0.005) | 0.188** (0.006) | 0.089** (0.007) |
| Year 2018 | -0.339** (0.005) | 0.142** (0.006) | 0.043** (0.007) |
| Year 2019 | -0.383** (0.005) | 0.146** (0.006) | 0.041** (0.007) |

Notes: N=11,616,809. Standard errors in parentheses. * and ** indicate significance levels of $p < 0.05$ and $p < 0.01$, respectively. This model simplifies the multinomial framework into a logistic model with only taking values of 0 for compliance and 1 for noncompliance, to better highlight the comparison between alternative approaches to weighting connections between ZIP Codes.

## TABLE A3. Full Parameter Estimates for Two-Stage Multinomial Compliance Model

($P_{ijt}$=0: compliant, 1: non-compliant no balance due, 2: non-compliant with balance due)

| Variable | $P_{ijt}$=0 | $P_{ijt}$=1 | $P_{ijt}$=2 |
|---|---|---|---|
| Intercept | 0.208**** (0.016) | -0.072**** (0.026) | -0.135**** (0.022) |
| ACS weighted average | 5.547**** (0.006) | -2.095**** (0.009) | -3.452**** (0.008) |
| CP59 weighted average | 3.078**** (0.003) | -1.152**** (0.005) | -1.926**** (0.004) |
| Field collection weighted average | 0.275 **** (0.000) | -0.100**** (0.000) | -0.175**** (0.000) |
| Married filing jointly | 0.187**** (0.002) | -0.179**** (0.004) | -0.008*** (0.003) |
| Log total positive income | -0.128**** (0.001) | 0.010**** (0.002) | 0.118**** (0.001) |
| Timely filed in past four years | 0.592**** (0.002) | -0.321**** (0.003) | -0.270**** (0.003) |
| Balance due (before remittance) | -0.146**** (0.003) | -0.044**** (0.005) | 0.190**** (0.004) |
| % of income under-withheld | -1.497 **** (0.010) | -0.153**** (0.016) | 1.649**** (0.015) |
| 50% income not subject to withholding | -0.091**** (0.003) | -0.008 (0.005) | 0.099**** (0.005) |
| Activity code 266 | 0.730**** (0.065) | -0.187** (0.105) | -0.543**** (0.093) |
| Activity code 270 | 0.246 **** (0.014) | -0.386**** (0.022) | 0.140**** (0.018) |
| Activity code 271 | 0.030 (0.025) | -0.189**** (0.041) | 0.158**** (0.033) |
| Activity code 272 | 0.387**** (0.013) | -0.143**** (0.021) | -0.244**** (0.016) |
| Activity code 273 | 0.318 **** (0.013) | -0.218 **** (0.022) | -0.100**** (0.017) |
| Activity code 274 | 0.080**** (0.013) | -0.167**** (0.021) | 0.087**** (0.017) |
| Activity code 275 | 0.066**** (0.016) | -0.229**** (0.027) | 0.163**** (0.021) |
| Activity code 276 | 0.336**** (0.042) | -0.238**** (0.071) | -0.098** (0.051) |
| Activity code 277 | 0.210**** (0.042) | -0.240**** (0.072) | 0.030 (0.051) |
| Activity code 278 | 0.397**** (0.018) | -0.271**** (0.030) | -0.126**** (0.025) |
| Activity code 279 | 0.430**** (0.014) | -0.145**** (0.024) | -0.285**** (0.019) |
| Activity code 280 | 0.231 **** (0.014) | -0.242**** (0.024) | 0.011 (0.019) |
| Activity code 266*DIF | -0.003**** (0.000) | 0.003**** (0.000) | 0.001** (0.000) |
| Activity code 270*DIF | -0.001 **** (0.000) | 0.000**** (0.000) | 0.000**** (0.000) |
| Activity code 271*DIF | -0.001**** (0.000) | 0.000**** (0.000) | 0.001**** (0.000) |
| Activity code 272*DIF | -0.001**** (0.000) | -0.000**** (0.000) | 0.001**** (0.000) |
| Activity code 273*DIF | -0.001**** (0.000) | 0.000**** (0.000) | 0.001**** (0.000) |
| Activity code 274*DIF | -0.001**** (0.000) | 0.000**** (0.000) | 0.001**** (0.000) |
| Activity code 275*DIF | -0.001**** (0.000) | 0.000**** (0.000) | 0.001**** (0.000) |
| Activity code 276*DIF | -0.001**** (0.000) | 0.000 (0.000) | 0.001**** (0.000) |
| Activity code 277*DIF | -0.001**** (0.000) | 0.000 (0.000) | 0.001**** (0.000) |
| Activity code 278*DIF | -0.000 (0.000) | -0.000*** (0.000) | 0.000**** (0.000) |
| Activity code 279*DIF | -0.001**** (0.000) | -0.000**** (0.000) | 0.001**** (0.000) |
| Activity code 280*DIF | -0.000**** (0.000) | -0.000 (0.000) | 0.000**** (0.000) |
| Year 2012 | 0.300**** (0.004) | -0.077**** (0.007) | -0.223**** (0.006) |
| Year 2013 | 0.727**** (0.004) | -0.133**** (0.006) | -0.594**** (0.006) |
| Year 2014 | 1.342**** (0.004) | -0.460**** (0.006) | -0.882**** (0.005) |
| Year 2015 | 1.446**** (0.004) | -0.536**** (0.006) | -0.910**** (0.005) |
| Year 2016 | 1.534**** (0.004) | -0.431**** (0.006) | -1.102**** (0.005) |
| Year 2017 | 1.695**** (0.003) | -0.433**** (0.005) | -1.263**** (0.005) |
| Year 2018 | 1.861**** (0.003) | -0.511**** (0.006) | -1.350**** (0.005) |
| Year 2019 | 2.063**** (0.003) | -0.588**** (0.005) | -1.475**** (0.005) |

Notes: N=11,616,809. Standard errors in parentheses. ** and *** indicate significance levels of $p < 0.05$ and $p < 0.01$, respectively.

**TABLE A4. Full Parameter Estimates for Linear Model of Change in Balance Due**

Response Variable:

| Variable | For 2 |
|---|---|
| Intercept | 5.503** (0.017) |
| ACS weighted average | -0.019** (0.006) |
| CP59 weighted average | -0.009** (0.003) |
| Field collection weighted average | -0.000** (0.000) |
| Married filing jointly | 0.068** (0.002) |
| Log total positive income | 0.202** (0.001) |
| Timely filed in past four years | -0.177** (0.002) |
| Balance due (before remittance) | -0.185** (0.003) |
| % of income under-withheld | 0.337** (0.011) |
| 50% or more of income not subject to withholding | 0.030** (0.003) |
| Activity code 266 | 1.111** (0.066) |
| Activity code 270 | -0.587** (0.014) |
| Activity code 271 | -0.481** (0.023) |
| Activity code 272 | -0.827** (0.013) |
| Activity code 273 | -1.024** (0.014) |
| Activity code 274 | -0.590** (0.014) |
| Activity code 275 | -0.654** (0.016) |
| Activity code 276 | -0.542** (0.033) |
| Activity code 277 | -1.008** (0.032) |
| Activity code 278 | -0.944** (0.019) |
| Activity code 279 | -0.601** (0.015) |
| Activity code 280 | -0.415** (0.014) |
| Activity code 266*DIF | -0.001** (0.000) |
| Activity code 270*DIF | 0.001** (0.000) |
| Activity code 271*DIF | 0.002** (0.000) |
| Activity code 272*DIF | 0.002** (0.000) |
| Activity code 273*DIF | 0.002** (0.000) |
| Activity code 274*DIF | 0.002** (0.000) |
| Activity code 275*DIF | 0.001** (0.000) |
| Activity code 276*DIF | 0.001** (0.000) |
| Activity code 277*DIF | 0.002** (0.000) |
| Activity code 278*DIF | 0.002** (0.000) |
| Activity code 279*DIF | 0.001 ** (0.000) |
| Activity code 280*DIF | 0.001** (0.000) |
| Year 2012 | -0.023** (0.004) |
| Year 2013 | -0.188** (0.004) |
| Year 2014 | -0.091** (0.004) |
| Year 2015 | 0.011** (0.004) |
| Year 2016 | 0.062** (0.004) |
| Year 2017 | 0.085** (0.004) |
| Year 2018 | 0.064** (0.004) |
| Year 2019 | -0.010** (0.004) |

Notes: N=3,487,662. Standard errors in parentheses. ** and *** indicate significance levels of $p < 0.05$ and $p < 0.01$, respectively.

# 3

# Trusting the Tax Man: Metrics, AI, and Audits

**Holtzblatt**

**Szulczewski ♦ Feldman ♦ Silva**
**Anderson ♦ Graff**

# Measuring Success: New Performance Metrics for A New Internal Revenue Service

*Janet Holtzblatt (Urban-Brookings Tax Policy Center)*[1]

## 1. Introduction

In April 2023, the Internal Revenue Service (IRS) took a major step toward modernization by releasing the IRS Inflation Reduction Act Strategic Operating Plan detailing how the agency will invest the 10-year $80 billion boost to its budget provided under the 2022 Inflation Reduction Act (IRS 2023b). Although that funding was cut by over 25% a month later in the Fiscal Responsibility Act, the IRS maintained its commitment to the plan with the understanding that funding for some initiatives would run out much sooner than initially anticipated.[2]

The plan is a serious and comprehensive effort to bring the agency into the 21st century, but lingering questions must be addressed to ensure its success. First, what is the long-term plan? The remaining $58.6 billion budget-boost is a 10-year investment, but most of the Strategic Operating Plan, as well as a 2024 update, provides details for just the next few years (IRS 2023b, 2024d). Many features of the plan involve recruitment, research, evaluation, and pilot programs—the first steps toward development and implementation of effective long-term strategies.

Second, how will success be measured? The strategic plan contains objectives and a summary of what success would look like for each new initiative. Although the update lists "outcomes," it does not provide specific metrics or targets for evaluating the agency's performance in achieving many of the plan's goals—either for the specific initiatives or for the entire plan once fully implemented.

Given the early stages of the plan's implementation, holding the IRS to tough metrics now would be premature. Establishing targets too soon could further discourage efforts to test different approaches to determine which is the most efficient and fair to taxpayers.

But identifying and designing serious performance metrics should begin before the IRS proceeds too far in implementing the various initiatives. This would allow development of serious metrics reflecting thoughtful and careful analysis, in coordination across IRS divisions, along with input from outside experts.

In this paper, I establish several principles for designing metrics:

- The goals of the performance measures should be transparent. A goal of a metric may be to support the IRS's budget request for overall funding, another set of metrics could inform internal decisions as to how to best allocate appropriations across programs, and a third set could help the IRS refine and improve a program.
- The metrics should be consistent with the outcomes emphasized in the IRS mission statement: taxpayer services, enforcement, and equitable treatment of taxpayers.
- Within those three broad outcome categories, the IRS should measure output metrics, which measure the IRS's achievement of specific actions—such as the number of phone calls answered or the successful targeting of audits.
- In some instances, it may be helpful to establish input metrics—for example, the costs of implementation of certain activities—or efficiency metrics, such as the return on investment.
- A methodology should be developed to distinguish between the IRS's role in administering the tax role from factors that are beyond the control of the IRS—such as economic conditions, the tax code, and the agency's funding.

1 This report was funded by an anonymous foundation. We are grateful to them and to all our funders, who make it possible for the Urban-Brookings Tax Policy Center to advance its mission. The author thanks Tracy Gordon, Arnstein Ovrum, and attendees at the 14th Annual IRS/TPC Joint Research Conference on Tax Administration for their helpful comments on an earlier draft and Lillian Hunter for assistance in compiling the report. The views expressed are those of the author and should not be attributed to the Urban-Brookings Tax Policy Center, the Urban Institute, the Brookings Institution, their trustees, or their funders. Funders do not determine research findings or the insights and recommendations of our experts. Further information on Urban's funding principles is available at https://www.urban.org/about/organizational-principles; further information on Brookings' donor guidelines is available at https://www.brookings.edu/support-brookings/donor-guidelines/.

2 Shortly before publication of this bulletin, Congress passed the "Full Year Continuing Appropriations and Extensions Act of 2025," which cut the IRA funding by an additional $20.2 billion.

Importantly, the metrics should not be viewed in isolation. Some metrics should be bundled together so that policymakers and administrators can assess the trade-offs between goals and weigh the choices between activities—whether it is a choice between services and enforcement, answering phones and opening the mail, and correspondence and in-person audits.

Finally, the effectiveness of all IRS actions cannot be reduced to a single quantitative metric or even a bundle of metrics. Performance measures are not a replacement for a thorough evaluation of the IRS's actions.

In this paper, I review the objectives of the IRS Strategic Operating Plan as well as prior legislation that has required the agency and other government agencies to set performance measures for at least some of its activities. I identify the shortfalls in the current patchwork of metrics and present a more holistic perspective on measuring the performance of the IRS. I then evaluate several examples of outcome, output, and efficiency metrics, pointing to ways those measures could be refined or expanded to provide more insight into the performance of the agency in achieving its mission of providing taxpayer services, enforcing the tax code, and treating taxpayers equitably.

The metrics include the following elements:

- Taxpayer services
  1. Taxpayer satisfaction (outcome)
  2. Compliance burdens (output) and
  3. Telephone service (output)
- Enforcement
  1. Tax gap (outcome)
  2. Audit rates (output) and
  3. Return on investment (efficiency)
- Fairness
  1. Compliance burden by income (outcome)
  2. Underreported income and taxes by income group (outcome) and
  3. Audit rates by race (output)

I do not discuss the establishment of targets for performance measures. The IRS is typically required to set targets for future performance, but the methodology for deriving those goals is rarely (if ever) described in IRS documents and studies. Lifting the veil on the methods used to set targets would provide more insight into the choice and design of the performance measures.

## 2. The Inflation Reduction Act

In May 2021, the Treasury Department released a multifaceted plan to reduce tax noncompliance. As detailed in "The American Families Plan Compliance Agenda" (US Treasury Department 2021), the impetus was the 19% reduction in the IRS's budget, after adjusting for inflation, from Fiscal Years 2010 through 2020. Those funding reductions contributed to a 20% reduction in the IRS workforce and the depreciation of the agency's technological infrastructure.

A key component of Treasury's agenda was a proposal to increase the IRS budget by $80 billion over the next decade. The agenda provided a broad overview of how the IRS would use those funds. Most of the proposed funding would be dedicated to increasing audits of large corporations, partnerships, and global high-wealth and high-income taxpayers. In addition, the massive budget infusion would finance investments in modern technology and new data analytical tools to help select returns for enforcement actions. Finally, the IRS would take steps to improve taxpayer services and facilitate

claims of refundable tax credits.

Beyond the unprecedented magnitude of the proposed budget boost, the agenda was unique for two other reasons: First, the funding would cover an entire decade in contrast to the one-year funding typically provided through annual appropriations. Providing a 10-year stream of funds was intended to facilitate long-term planning and investment in technology and staff. Second, the $80 billion boost was conceived as a supplement on top of the annual appropriation—that is, the IRS would still receive funds through the annual appropriations legislation for its normal operating expenses.

In August 2022, Congress passed the Inflation Reduction Act (IRA), which authorized the $80 billion boost for tax administration from 2022 to 2031—with $79 billion to the IRS and the remaining funds divided between other Treasury offices and the US Tax Court (Table 1). The IRA contained even fewer details than the agenda about how the funding would be used, mandating only the division of the funds among the four broad IRS budget accounts: Taxpayer Services, Enforcement, Operations Support, and Modernization. Over half the funds were allocated to tax enforcement, with just 4% set aside for taxpayer services.

Since the IRA's enactment, Congress has chipped away at the funding—both directly and indirectly. As part of the agreement between Congress and the President to lift the debt ceiling in 2023, the IRA funding was immediately cut by $1.4 billion in FY 2023, with the reductions to be allocated between the enforcement and operations support accounts. (Ultimately, the IRS reduced the enforcement account by $1.4 billion.) Another $20.2 billion was rescinded in the Further Consolidated Appropriations Act of 2024—all coming from the enforcement allocation.

Moreover, Congress did not erect guardrails between the IRA 10-year funds and the annual appropriations. For Fiscal Years 2022 through 2024, the annual IRS appropriations have been frozen at $12.3 billion for taxpayer services, enforcement, and operations support. Appropriations for business systems modernization have been eliminated, with the justification that the IRA funds will be used for technological advancements.

In total, the IRS had spent $5.7 billion—or 10%—of the IRA funding as of March 2024. However, $2 billion was used to pay for normal operating expenses because of the shortfalls in the annual appropriations (TIGTA 2024).

**TABLE 1. Internal Revenue Service's Mandatory Funding, Fiscal Years 2022–2031**

| | IRA 2022 | | Rescissions | | | Mandatory Funding After Rescissions | |
|---|---|---|---|---|---|---|---|
| | | | FRA 2023 | FCAA 2024 | Total | | |
| | ($) | (%) | ($) | ($) | ($) | ($) | (%) |
| Taxpayer services | 3,181 | 4.0 | - | - | - | 3,181 | 5.6 |
| Enforcement | 45,637 | 57.8 | 1,400 | 20,200 | 21,600 | 24,037 | 41.9 |
| Operations support | 25,326 | 32.1 | - | - | - | 25,326 | 44.2 |
| Modernization | 4,751 | 6.0 | - | - | - | 4,751 | 8.3 |
| Direct File study | 15 | * | - | - | - | 15 | * |
| Total IRS funding | 78,911 | 100 | 1,400 | 20,200 | 21,600 | 57,311 | 100 |
| Addendum: Funding for organizations other than IRS | | | | | | | |
| TIGTA | 403 | | - | - | - | 403 | |
| Treasury Department Office of Tax Policy | 105 | | - | - | - | 105 | |
| Treasury departmental offices | 50 | | - | - | - | 50 | |
| US Tax Court | 153 | | - | - | - | 153 | |
| Total | 79,621 | | 1,400 | 20,200 | 21,600 | 58,021 | |

Notes: *=less than 0.5. Dollar values are in millions. IRA 2022 = Inflation Reduction Act of 2022, Pub. L. No. 117-169; FRA 2023 = Fiscal Responsibility Act of 2023; Pub. L. No. 118-5; FCAA 2024 = Further Consolidations Appropriations Act of 2024, Pub. L. No. 118-47. TIGTA = Treasury Inspector General for Tax Administration

## 3. Strategic Operating Plan

A week after the IRA's passage, Treasury Secretary Janet Yellen instructed the IRS to release an operating plan within six months.[3] Along with more details on how the funds would be used, Secretary Yellen also requested that the plan include metrics for measuring performance.

### *3.1 2023 Strategic Operating Plan*

In April 2023, the IRS released a 150-page strategic plan (IRS 2023b). Center to the IRS's Strategic Operating Plan were five objectives:

1. Dramatically improve services to help taxpayers meet their obligations and receive the tax incentives to which they are eligible
2. Quickly resolve taxpayer issues when they arise
3. Focus expanded enforcement on taxpayers with complex tax returns and high-dollar noncompliance to address the tax gap
4. Deliver cutting-edge technology, data, and analytics to operate more efficiently
5. Attract, retain, and empower a highly skilled, diverse workforce and develop a culture that is better equipped to deliver results for taxpayers

For each objective, the IRS listed "indicators of success." However, those were not always fully specified. In some cases, they were simply restatements of the objectives. For example, the first indicator of success in dramatically improving services was "increasing service levels."

Other indicators revealed features of the plan but still without specific metrics—for instance, a "wider array of digital options to help taxpayers and tax professionals interact with the IRS and have a more seamless customer experience." And still others described a metric for evaluation but did not set a quantitative target. An example is "decreased percentage of returns filed with math errors or errors related to third-party information reported to the IRS."

The Strategic Operating Plan also listed 42 initiatives aimed at helping the IRS achieve each objective. The explanation of each initiative included a description of what success would look like. As with the overall indicators of success, those descriptions varied in degree of specificity.

### *3.2 2024 Update to Strategic Operating Plan*

With the 2024 release of an update to the Strategic Operating Plan (IRS 2024d), the IRS made strides toward defining metrics and setting targets for many of its objectives. The update matched objectives to outcomes and specified priority efforts for 2024 and 2025.

Consider again the first objective identified in the Strategic Operating Plan of dramatically improving services. In the 2024 update, the IRS cited nine desired outcomes that would indicate achievement of improved services (e.g., "When taxpayers call the IRS, they are able to reach an agent in a timely manner and have high levels of satisfaction with the interaction"). And to achieve that outcome, the IRS listed three priority efforts (e.g., an 85% rate of answered phone calls on the IRS helpline during the filing season with an average wait time of fewer than five minutes).

As in the example above, the priority efforts were sometimes defined in quantitative measures. In other cases, the effort was described more generally, especially when quantifying the success of the action is not feasible: for instance, other priority efforts intended to meet the objective of better taxpayer services were described simply as "improve Where's My Refund? tool" or "assess impact of Direct File."

3 US Secretary of the Treasury Janet Yellen, memorandum for IRS Commissioner Charles Rettig, August 17, 2022, https://www.taxnotes.com/tax-notes-today-federal/tax-system-administration/yellen-requests-irs-plan-resource-implementation/2022/08/18/.

## 4. Legislative Requirements for Measuring the IRS's Performance

Setting performance metrics is not a new task for the IRS. Other legislation—dating back at least 20 years—has required that the IRS and other government agencies evaluate their performance in certain areas. Those provisions include:

- the Government Performance and Results Act (GPRA) of 1993,[4] as amended by the GRPA Modernization Act of 2010[5]
- the Paperwork Reduction Act (PRA) of 1980,[6] as amended by the Paperwork Reduction Act of 1995[7]
- the Improper Payments Information Act (IPIA) of 2002,[8] as amended by the Payment Integrity Information Act of 2019[9]

Of the three, the GPRA is the most extensive. Under the act, government agencies must produce annual performance metrics for both services and enforcement. The PRA and IPIA establish metrics for evaluating certain aspects of services and enforcement, respectively.[10]

### *4.1 Government Performance and Results Act*

GPRA requires government agencies to set goals, periodically prepare strategic plans, measure programs' recent effective-ness each year, and set targets for the future. If an agency does not meet those goals, then it must produce a performance improvement plan.

Under GPRA, the performance metrics must measure or assess a program's outputs, service levels, and outcomes. An output measure is defined as the tabulation, calculation, or recording of an activity or effort, whereas an outcome measure is an assessment of how well the program achieved its goals. In its instructions to agencies, the Office of Management and Budget (OMB) encourages agencies to use outcome measures when feasible and appropriate but also lists a broader range of performance indicators than referred to in GPRA.[11] Those include measures for inputs (time or monetary costs) and efficiency (the ratio of the outputs or outcomes to the inputs).

In its Fiscal Year 2025 budget released in February 2024, the IRS identified 25 performance measures (IRS 2024a) categorized by the relevant IRS budget account—taxpayer services, enforcement, operation support, and business mod-ernization. Most metrics focused on the outputs—for example, the number of answered phone calls and other services provided by the IRS to the public or the exam rates and other counts of enforcement activities. A few measured the inputs (such as rentable space feet per person) or the efficiency of the activity (the costs of collecting $100).

But only two performance measures came close to gauging the outcomes: the percentage of surveyed taxpayers sat-isfied with the IRS (an indicator of the effectiveness of taxpayer services) and the share of individual taxpayers who are noncompliant two years after a prior bad act (an indicator of the effectiveness of IRS enforcement actions).

The IRS's list of performance metrics has evolved. For example, in response to growing interest in equitable treatment of taxpayers, the IRS added three new output metrics showing the number of newly undertaken audits of high-income taxpayers, partnerships, and large corporations beginning in the Fiscal Year 2022 budget.[12]

4 Government Performance and Results Act of 1993, Pub. L. No. 103-62, 107 Stat. 285 (1993).

5 GRPA Modernization Act of 2010, Pub. L. 111-352, 124 Stat. 3866 (2011). -

6 Payment Reduction Act of 1980, Pub. L. 96-115, 94 Stat. 2812 (1980). -

7 Paperwork Reduction Act of 1995, Pub. L. 104-13, 109 Stat. 163 (1995). -

8 Improper Payments Information Act of 2002, Pub. L. 107-300, 116 Stat. 2350 (2002). -

9 Payment Integrity Information Act of 2019, Pub. L. 116-117, 134 Stat. 113 (2020). -

10 Other legislative acts have included provisions concerning performance metrics. The Internal Revenue Service Restructuring and Reform Act of 1998 (Pub. L. 105-206, 112 Stat. 685) required the IRS to set performance goals for organizational units and the establishment of a balanced performance measurement system. The IRS subsequently established a system (Establishment of a Balanced Measurement Program, 64 Fed. Reg. 42835 [August 6, 1999]) composed of three elements: customer satisfaction measures, employee satisfaction measures, and business results. Although the metrics were developed for monitoring individual units within the IRS, they are used—when appropriate—in establishing metrics for the entire agency under the Government Performance and Results Act. The Taxpayer First Act (Pub, L. 116-25, 133 Stat. 981 [2019]) mandates that the IRS identify metrics and benchmarks for quantitatively measuring the progress of the IRS in implementing a comprehensive customer service strategy.

11 Executive Office of the President, Office of Management and Budget, Preparation, Submission, and Execution of the Budget, Circular No. A-11 (2017).

12 High-income individuals are those with a total positive income of $10 million and above. Total positive amounts shown for the various sources of income are reported on the individual income tax return and exclude losses. Large corporate returns are those reporting assets of $250 million and more.

### *4.2 Paperwork Reduction Act*

Even before GRPA, government agencies were required to report on at least one performance element. Under the Paperwork Reduction Act of 1980 (as amended in 1995), agencies—including the IRS—must annually release estimates of the compliance burdens imposed on individuals and businesses by filling out forms. For the IRS, those forms include (but are not limited to) tax returns, W-2s, and 1099s. The burden estimates are partially based on taxpayer surveys, which ask respondents about the amount of time and money spent to complete income tax returns.

While the IRS has devoted substantial resources to developing compliance burden measures, the PRA's mandate is limited and focuses solely on one aspect of taxpayers' interactions with the IRS—the costs of completing each IRS form. Thus, the measures understate the total compliance costs incurred by taxpayers—when, for example, waiting for an IRS operator to answer the phone, responding to a request for documentation to support a claim of a child dependent, or being audited.

### *4.3 Improper Payments Information Act*

Although the IRS GPRA performance measures do not include estimates of tax noncompliance, the IRS is required to report the amount of erroneous payments of several tax credits each year. Under the Improper Payments Information Act of 2002, as amended by the Payment Integrity Information Act of 2019, each agency must identify programs and activities that "may be susceptible to significant improper payments." Improper payments are defined as any payment that should not have been made or was made incorrectly (either too much or too little) under the law.[13]

Since the launch of IPIA, the OMB and the Treasury Department have included the earned income tax credit (EITC) in the list of programs subject to improper payment reporting. The list has expanded to include three other tax credits: the additional child tax credit (the refundable portion of the child tax credit), the American Opportunity Tax Credit, and the premium assistance tax credit.[14]

The common feature that differentiates those four credits from other tax provisions is that they are partially or fully refundable, meaning that credit claimants can receive payments even if they do not have any income tax liabilities. To the extent that the credits exceed income tax liabilities, the payments are counted as outlays in the federal budget, which is likely considered a justification for the inclusion of those credits with more conventional spending programs in the improper payment analysis.

But the refundable nature of the four credits, combined with income caps on eligibility, also means that lower- and middle-income families are the segment of the population most likely to receive those benefits. Hence, only noncompliance among those groups is required to be reported annually, even though the estimates of improper payments are derived from some of the data used to measure the tax gap—the National Research Program, a nationally representative sample of all individual income tax returns, selected randomly for audits.

## 5. Performance Measures and the IRS Mission

The current IRS metrics are a patchwork of measures mandated by various legislative or administrative requirements. Considered together, they are neither comprehensive nor cohesive.

The IRS's mission statement might be an appropriate starting point for developing a more holistic set of metrics. That statement says that the IRS's mission is to "provide America's taxpayers top quality service by helping them understand and meet their tax responsibilities and enforce the law with integrity and fairness to all." Outcome measures would focus on taxpayer services, enforcement, and equitable treatment of taxpayers to evaluate the IRS's achievement of its mission

13 In this context, the term "significant" means that, in the preceding fiscal year, the sum of a program or activity's improper payments and payments whose propriety cannot be determined by the executive agency due to lacking or insufficient documentation may have exceeded (1) $10,000,000 of all reported program or activity payments of the executive agency made during that fiscal year and 1.5% of program outlays or (2) $100,000,000.

14 "Payment Accuracy," US Federal Government, accessed July 17, 2024, https://www.paymentaccuracy.gov/.

statement. Output, input, and efficiency metrics could provide additional context as to how the IRS uses the tools at its disposal to achieve those outcomes.

Regardless of the type of performance measure, two issues must be resolved in its choice, design, and implementation—the purpose of the metric and the baseline against which the activity's performance should be measured.

### *5.1 Purpose*

From outside the agency, the IRS performance measures may resemble report card grades to evaluate the agency's funding. The most recent example was the response to estimates of returns on investment during the deliberations over the IRA funding. Though Treasury and Congressional Budget Office (CBO) analysts disagreed on the amount of revenue that could result from an increase in the IRS's enforcement budget, they concur that the net yield would be positive. That finding played a significant role in the ultimate passage of the IRA funding.

Less visible is the use of performance metrics to highlight trade-offs between programs. Consider statistics of the number of phone calls answered. The IRS issues press releases about the percentage of phone calls to the agency that are responded to during the filing season, reporters write about those results, and lawmakers query IRS and Treasury officials about those numbers during hearings about the filing season and appropriations.

Of equal importance, however, is that the same people who answer the calls also open the mail, and one task can divert resources from the other. In the aftermath of the COVID-19 pandemic, as the IRS staff tried to work through a backlog of unopened paper tax returns, the trade-off between the two tasks became much more visible—in part due to the comments of National Taxpayer Advocate Erin Collins.[15] Bundling together performance measures enables the IRS and others to recognize the trade-offs between goals and make informed judgments as to which activities to prioritize.

Even less transparent is how the IRS uses performance measures to determine how to fix a program with a less-than-satisfactory performance measure. Some exceptions exist, most notably in the IRS's recent focus on audit rates for Black and White taxpayers. In presentations (such as at the IRS–Tax Policy Center research conferences in 2023 and 2024), IRS researchers (Anderson, et al. 2024, Hertz, et al. 2023) have not only presented data on racial disparities but have also discussed their findings from a more in-depth analysis of the reasons for those differences and how their research has led to changes in the ways the IRS administers audits.

### *5.2 Baseline*

Another issue common to all metrics is the baseline for observing changes in performance. Often, changes in performance are measured from one year to another or in some cases, back to a year that supports an argument in favor of increases or cuts in funding for the IRS or other legislative actions. For example, supporters of increases to the IRS budget compared current audit rates to higher levels in 2010 (when funding was relatively high) or the number of answered telephone calls in 2023 to much lower levels during the heights of the pandemic when there was a substantial increase in callers asking questions about temporary assistance programs and the IRS's staff were working remotely or out caring for themselves and others.

However, a simple comparison of two measures from different years does not indicate either improvements or deterioration of the IRS's performance alone. Achieving the IRS's three missions is not solely the agency's responsibility. The IRS does not write the tax code, but complex laws increase the burden of filing a return, open vulnerabilities for avoidance and noncompliance, and may lead to inequitable treatment because of the ways complicated laws affect groups differently. Nor does the IRS control its funding. Changes in the economy also affect the IRS's performance. In the 21st century, lawmakers have turned to the IRS multiple times to rapidly deliver lump-sum payments to individuals—including people who typically are not required to file tax returns because their income is very low—to alleviate the economic burdens caused by recessions and the pandemic.

15 Michelle Singletary. "This Tax Season, the IRS Answered Just 10% of Taxpayer Calls," The Washington Post, June 24, 2022, https://www.washingtonpost.com/business/2022/06/24/irs-taxpayer-calls/.

Ideally, the IRS would develop metrics that could distinguish between outcomes attributable to its actions and those outside its power to influence. One comment in the IRS's 2023 Strategic Operating Plan suggests that the agency recognizes that need, stating that an indicator of success would be if the tax gap fell "relative to tax gap without the resources provided by the IRA" (IRS 2023b).

But that is only a partial step toward isolating the effectiveness of the agency in enforcing the tax code because it ignores the role of changes in tax laws and the economy. For a model of how to disaggregate the sources of changes to the tax code, the IRS could turn to the budget forecasts of the OMB and CBO. The agencies break down the differences between actual budget data and their prior projections into three categories—changes due to (1) revised economic assumptions, (2) technical adjustments, and (3) newly enacted legislation.

The IRS is partway there. In its reports on the tax gap, the IRS compares its results to those from the prior study and then decomposes the differences into two categories: (1) updated methods and (2) other factors. For example, the IRS estimated that the annual voluntary compliance rate over the 2014–16 period was 85%—up by 1.4 percentage points from the annual rate over the prior three years, of which only 0.1 percentage point was due to revisions in methodology (IRS 2022b).

## 6. Taxpayer Services

Currently, the IRS has two outcome metrics for taxpayer services: the GPRA metric for taxpayer satisfaction and the PRA's measure of compliance burden. In addition, there are eight GPRA output measures—most prominently, the number of telephone calls answered. Those measures could be expanded to provide more insight into the quality of the IRS's services for taxpayers.

### *6.1 Outcome: Taxpayer Satisfaction*

To measure individual taxpayers' satisfaction, the IRS relies on information collected by the American Customer Satisfaction Index (ACSI 2023), a private company founded by researchers at the University of Michigan. Each year, ACSI releases a report on citizen satisfaction with the federal government with breakouts for cabinet departments. The information in the report is based on interviews with a random sample of individuals, with their responses serving as input into an econometric model that derives scores of citizens' satisfaction ranging from 0 to 100.

The public report typically does not include those metrics for all sub-cabinet agencies. However, the IRS obtains the results for filers and includes the value in the agency's annual performance measures. In Fiscal Year 2022, the IRS's customer satisfaction score was 69. That aggregate score, however, illustrates one of the challenges of a broad performance metric: in this instance, the metric reveals overall satisfaction, but it is insufficient to identify the IRS's strengths and weaknesses. Consequently, the taxpayer-satisfaction metric informs policymakers of the IRS's overall performance but does not provide any insight into the agencies' weaknesses and areas for improvement.

ACSI collects more in-depth data, which would be useful for program evaluation. For example, the ACSI identifies four drivers of citizen satisfaction with the federal government:

- Efficiency and ease of government processes
- Ease of access and clarity of information
- Courtesy and professionalism of customer service
- Perceptions of government websites

The ACSI reports often include a score for each attribute, but only at the government-wide level. Yet, data for departments and agencies might be more useful in at least pointing in the general direction of the source of dissatisfaction.

While further details would be desirable, methodological concerns may constrain the use of the ACSI data. Relative to government household surveys, the description of ACSI's methodology is sparse on its website. But one anomaly stands out. The size of the survey fluctuates significantly from year to year: 1,291 in 2020 to 2,126 in 2022. In 2023, the sample

size shrunk to 847. Perhaps related to that substantial decline in sample size, the IRS did not report a score for taxpayer satisfaction in 2023 in its Fiscal Year 2025 budget, citing ongoing updates to the methodology.

Another potential source of information is the Comprehensive Taxpayer Attitude Survey or CTAS (IRS, 2022a). That survey is rooted in the Executive Order 12862, issued in 1993, which required agencies to survey customers—people or entities directly dealing with the organization—regarding their satisfaction with its current services.[16] The survey found that 75% of taxpayers reported being very or somewhat satisfied with their personal interactions with the IRS in 2021—about 5 percentage points higher than the ACSI found.

One option would be to replace the ACSI with the CTAS as the source of information on taxpayer satisfaction. The large sample size—2,099 adults—may facilitate reliable analysis of subgroups. And because the survey is solely about the IRS, questions can be tailored to address the concerns of taxpayers. However, most of the questions in the 2021 survey focused more on attitudes about the IRS and tax system and did not provide much insight into the specific administrative challenges faced by taxpayers.

A promising sign is President Biden's 2021 initiative to evaluate certain government services, drawing on techniques used to study user experiences in other sectors. As a consequence of the "Executive Order on Transforming Federal Customer Experience and Service Delivery to Rebuild Trust in Government" (Executive Order 14058),[17] the OMB designated the IRS as one of 38 "high-impact service providers" (HISP) in the federal government.[18] HISPs were selected based on the size of their client base or critical impact on those served—both criteria that apply equally to the IRS. Currently, OMB is working with each HISP to develop and implement users' feedback surveys that will be used to derive scores for seven categories: trust, satisfaction, effectiveness, ease, efficiency, transparency, and the quality of interactions with employees.[19] Implementation of this executive order is still in the nascent stage and, at least for the time being, is limited to prioritized services. For the IRS, the current priorities are taxpayers' experiences with return filing and online accounts.

### *6.2 Outcome: Compliance Burdens*

Since the 1980s, the IRS has produced estimates of compliance burdens, relying in part on random surveys of taxpayers. The first surveys asked respondents to report the amount of time they spent on recordkeeping, learning about the law and form, completing the tax form, and transmitting it to the IRS. The survey data were then matched to respondents' tax returns, and the matched data formed the basis of a mathematical model (the ADL model, so called because the survey and modeling were conducted by the Arthur D. Little company).

The ADL model had several shortcomings, which became more problematic over time as people became increasingly reliant on alternative methods of filing. First, the survey did not ask respondents about their monetary costs, including payments to preparers. Nor did the model anticipate the shift from paper returns to preparation software and electronic filing.

While the ADL model is still used for many forms, the IRS began shifting in 2003 to a new approach for estimating compliance burdens for individual and business income tax returns (IRS 2023a). As with the ADL model, the new approach begins with surveys of random samples of individuals and businesses. But the new surveys ask about out-of-pocket expenses as well as hours. Moreover, the random sample of surveyed individuals is stratified by preparation method and complexity category (ranging, for example, from low complexity for wage and salary income to high complexity for partnership income).[20] For the business survey, companies are divided into groups based on their organizational structure and

16 Exec. Order No. 12862, 58 FR 48257. https://www.google.com/books/edition/United_States_Code_Congressional_and_Adm/H9QkAAAAMAAJ?hl=en&gbpv=1&dq=Exec.+Order+No.+12862,+58+FR+48257&pg=SL2-PA73&printsec=frontcover

17 "Executive Order on Transforming Federal Customer Experience and Service Delivery to Rebuild Trust in Government." The White House, December 13, 2021, https://www.federalregister.gov/documents/2021/12/16/2021-27380/transforming-federal-customer-experience-and-service-delivery-to-rebuild-trust-in-government .

18 "Federal Customer Experience," US Federal Government, accessed on July 17, 2024, https://www.performance.gov/cx/.

19 Executive Office of the President, Office of Management and Budget, Preparation, Submission, and Execution of the Budget, Circular No. A-11 (2020). https://www.performance.gov/cx/assets/files/a11-280.pdf

20 Tax provisions were categorized by level of complexity based on recordkeeping intensity, tax planning activities, and overall difficulty of extracting information from the taxpayer's financial books (IRS 2023b).

size. The data are then used to build the individual and business burden models, in which the logarithm of the burden—both the monetarized hours and the out-of-pocket expenses—is linearly related to a set of explanatory variables, including income for individuals and assets and receipts for businesses.

The estimates for individual and corporate income tax returns are broken into two components:

- Hours spent on each of the following categories—recordkeeping, tax planning, form completing and submission, and other
- Total out-of-pocket expenditures, ranging from payments to preparers and purchases of tax return preparation software to much smaller items such as copying costs and postage

For example, the 2023 instructions for the individual income tax return (1040)[21] show that taxpayers, on average, take 13 hours to complete their tax return, with nearly half that time devoted to recordkeeping. In addition, they spend an average of $270. Average costs were higher for filers with business income (24 hours and $560) and lower for other filers (9 hours and $150).[22]

Despite the upgrades, the current measures of compliance burden fall short of measuring the IRS's performance in providing taxpayer services—as well as the burdens of interacting with the IRS during an enforcement action.

First, the reported measures of paperwork burden are not broken down by taxpayers' characteristics—other than those that might be inferred by their completion of a form (e.g., we can infer that the filer who attaches a Schedule EIC to his or her tax return is also reporting labor income, total income below a certain threshold, and probably children). Less can be inferred from knowing the compliance cost of completing a 1040 because the form is used by all types of filers (especially since the elimination of the simpler 1040A and 1040EZ in 2018).

A few IRS studies have provided additional information about the association between the paperwork burden and observable characteristics. Because the sample is stratified by the complexity of the return, IRS and Treasury Department researchers could provide more detailed information from the 2010 survey about the incidence of taxpayer burdens by the complexity of tax items on individuals' returns (Marcuss, et al. 2013). Over half of the compliance costs incurred in the individual income tax were associated with reporting and substantiating income, even for relatively simple returns. As discussed in the section on equity measures, other studies have constructed distributions of paperwork burdens by income group. More analysis of the burden distribution by the presence of children, age of taxpayer, and race and ethnicity would also be informative.

Second, the compliance burden measures meet the requirements of the Paperwork Reduction Act, but it is far from a comprehensive metric of the burden of interacting with the IRS. For example, the burden models do not include prefiling correspondence and discussions between the IRS and the taxpayers (or their advisers) to obtain guidance, such as a letter ruling, nor do the current measures include the burdens attributable to post-filing interactions between the taxpayer and the IRS, though questions about the costs of those interactions were included in at least one of those surveys. An analysis of that data from the 2012 survey indicated that the compliance costs associated with examinations could be as much as $900 more than the costs associated with filing a return for affected taxpayers; however, because only a small share of taxpayers deal with the IRS after filing a return, nearly 60% of aggregate compliance costs were incurred before tax returns were filed (Guyton and Hodge 2013). Regularly updating this type of analysis would provide a fuller picture of the costs entailed with dealings with the IRS.

Both preceding challenges have been long recognized by the IRS. In 1998, an IRS study group concluded that the ideal burden estimation model would provide compliance costs by type of tax, taxpayer, and activity (GAO 2000). The group also recommended that burden measures account for all prefiling, filing, and post-filing activities (including enforcement activities).

21 https://www.irs.gov/pub/irs-pdf/i1040gi.pdf

22 https://www.irs.gov/pub/irs-pdf/i1040gi.pdf

Third, the compliance burden surveys are restricted to people who *actually* file tax returns. Yet, the complexity of the tax system potentially burdens people who do not file a tax return because they are not required to do so. Many would have received a refund of over-withheld taxes or a refundable tax credit if they had filed. Those nonfilers may have incurred compliance costs, especially if they had tried and failed to navigate the tax filing process.

Finally, the term "compliance burden" is a misnomer. As measured, the costs include those borne by people not complying with the tax code and by taxpayers following the law. The current measures of compliance costs may be pushed up by taxpayers who search for strategies to avoid or evade their tax liabilities. The average costs could also be deflated by taxpayers who do not read or understand the instructions and thus make inadvertent errors.

Distinguishing between the costs incurred by compliant and noncompliant taxpayers would be especially useful if the scope of the compliance burden studies was permanently expanded to include post-filing activities. The costs involved in an audit can be viewed as part of the penalty when the affected taxpayer is, in fact, noncompliant, but they are unambiguously a burden when the compliant taxpayer must undergo the pain of an unnecessary examination. Linking the measure of compliance burdens to noncompliance research, if possible, would enable the IRS to distinguish between costs incurred by compliant and noncompliant taxpayers.

### *6.3 Output: Telephone Service*

Among the most cited IRS performance measures is the share of telephone calls to the IRS that are answered—the level of service (LOS). It may also be one of the most misunderstood measures.

Consider, for example, reporting of telephone service during 2023—the first filing season after the enactment of IRA. At the close of the 2023 filing season, IRS Commissioner Werfel heralded the historically large increase in the LOS from the prior year—from 16% in 2022 to 85% in 2023 (Werfel 2023)—the target set by Treasury Secretary Janet Yellen shortly after the passage of IRA.[23] An IRS announcement in April 2024, however, stated that the LOS during the 2023 filing season had been just 84%—sparking a reporter's investigation as to whether the IRS had actually missed by a percentage point the 85% target set by Yellen for that year (Rifaat 2024).

Other estimates of answered phones in 2023 differed by much more than a percentage point. The Treasury Inspector General for Tax Administration reported that only 52% of calls had been answered through May 2023—up from 29% in 2022 over the same period (TIGTA 2023). And even estimates by the IRS can differ by much more than a percentage point. In its congressional justification for the proposed Fiscal Year 2025 budget, the IRS showed that phone service increased from 17% in 2022 to just 52% in 2023 (IRS 2024b).

One reason for those differences is timing. In response to the reporter's query about the one-percentage-point difference for 2023, a spokesperson for the Treasury Department speculated that there might have been a slight data lag between "Tax Day itself and the end of the filing season." (Rifaat 2024).[24]

The much greater gap between the numbers in the April press release and the congressional justification the following year is due to the former covering just the filing season (from January to April) and the latter representing the entire fiscal year (from October 2022 through September2023). The lower estimates for the entire year reflect shifts in priorities for customer service representatives during the year—from being responsive to taxpayers' questions as they prepare their tax returns to later inputting data from paper returns and responding to taxpayers' correspondence.

The specification of the telephone response rate also contributes to different estimates. For many years, the IRS focused solely on the number of attempted toll-free calls routed to Accounts Management—the line for callers seeking general tax information and updates on tax returns, refunds, and balances due. The level of service (LOS) is the percentage of callers who speak to a customer service representative or receive prerecorded informational messages.[25] TIGTA's measure,

23 Yellen, Janet, "Memorandum for Commissioner Rettig: IRS Operational Plan," September 15 2022, https://www.taxnotes.com/tax-notes-today-federal/tax-system-administration/yellen-requests-irs-plan-resource-implementation/2022/08/18/7dym5?highlight=yellen%20rettig.

24 An IRS spokesperson told the reporter that they could not explain the discrepancy.

25 The denominator includes abandoned calls, disconnects, and busy signals.

the level of access, includes calls diverted to targeted automated lines based on the callers' responses to prompts. However, TIGTA also limits its telephone metric to attempted calls during the IRS open hours.

Beginning with its Fiscal Year 2024 congressional justification, the IRS introduced a new metric for measuring the performance of customer service representatives—the LOS(A). This measure includes callers who received answers to their questions through an automated tool, though the IRS did not also adopt the TIGTA restriction of only including calls made during working hours. Relative to the original LOS, the LOS(A) was 22 percentage points higher in 2022 and 15 percentage points higher in 2023 (IRS 2024b). Currently, both the LOS and LOS(A) are presented in the IRS's congressional justifications, though it is not obvious which measure is now being used in the April releases.

A third potential source of misunderstanding regards the scope of coverage. The IRS's LOS measures (as well as TIGTA's LOA) are limited to calls routed to Accounts Management. While customer service representatives fielded about 18 million calls in 2023, the IRS received more than 50 million calls (IRS 2024a)—including calls to collections, the refund hotline, Taxpayer Protection Service, to establish installment agreements, or by practitioners seeking priority service.

Beyond the measurement issues, a single-minded focus on the number of calls answered does not give a full picture of the quality of telephone service. Another performance measure, based on a sample of calls, considers the accuracy of the information provided by the customer service representative, and the IRS sometimes separately reports on the wait time and the duration of the call in testimony and press releases. But as National Taxpayer Advocate Erin Collins (2023) points out, other helpful metrics are still missing, including the number of times a taxpayer hangs up because of the length of the wait, whether the taxpayer's issue is resolved during the call, and the taxpayer's perception of their experience with the customer service representative.

Nor do the current performance measures shed much light on the trade-offs in choosing the resources to devote to telephone services. As noted above, the lower year-round LOS estimates reflect shifts in customer service representatives' tasks during the year. The phones do not stop ringing (though likely in lower numbers) after Tax Day, but other delayed tasks take precedence when the filing season ends. Some insight into the output associated with those other tasks is provided by a second performance measure—the number of accounts management and correspondence work to be processed in inventory—but it is difficult to interpret without additional data on the composition of the inventory.

Context matters in other ways. Technological advances—such as more-accessible information on the IRS website and chatbots—may reduce reliance on telephone service. But if that means that a greater share of answered calls involves complicated questions, the LOS might fall as the waiting period and duration of calls lengthened. Changes in the tax code or unanticipated external events (such as a devastating hurricane or pandemic) may also pressure taxpayer services. Comparisons across years that do not account for factors outside the control of the IRS will make the IRS's performance look weaker.

## 7. Enforcement

Eleven of the GPRA performance measures concern the IRS's performance enforcing the tax code, but only one comes close to being an outcome measure—the repeat noncompliance rate. Still, there are three widely cited enforcement metrics—the tax gap (an outcome measure), audit rates (an output metric), and the return on investment (an efficiency measure). At various times, Congress applied restrictions to developing and using the tax gap and return-on-investment metrics. Both have become more visible after the substantive inflation-adjusted cuts to the IRS budget after 2010.

### *7.1 Outcome: Tax Gap*

Since 1964, the IRS has periodically conducted studies of tax noncompliance. To many, the tax gap—the difference between the taxes owed and the tax paid on time—may be viewed as the ultimate measure of the IRS's performance as an enforcer of the tax code. Yet, the compliance studies are not mandated, and the tax gap is not included in the performance measures. Indeed, Congress denied funding to continue the studies after 1988 because of concerns about the burdens

imposed on individuals selected at random for audits. Funding was restored in the early 2000s only after the IRS committed to revamping the studies to reduce the burden on individual taxpayers.

In its most recent study of noncompliance, the IRS estimated that the gross tax gap was $496 billion per year, on average, for tax years 2014 through 2016—or 15% of total tax liabilities owed by individuals and businesses (IRS 2022b). Late payments and enforcement revenue reduced the annual tax gap by $68 billion to $428 billion, on net, and the amount of unpaid taxes to 13% of the total owed.

The IRS estimates the tax gap using information from various sources, including a sample of taxpayers selected randomly for audits, operational audits, and household survey data. By far, the most dominant data source is the National Research Program (NRP) audits of individual taxpayers.

The NRP starts with a stratified random sample of individual income tax returns that are selected for audit. The scope of the audits, however, depends on the complexity of the tax return.

- For the simplest returns, if the IRS can reconcile reported amounts with information supplied by third parties (e.g., W-2s and 1099s) and there is no indication of any significant compliance issue, the IRS does not follow up with the taxpayer.
- For somewhat more complicated returns, the IRS will conduct correspondence audits that usually focus on just a few items on a return.
- For the most complicated returns, the IRS will conduct a face-to-face interview with the taxpayer at an IRS office, the taxpayer's home, place of business, or accountant's office.

At the end of the audit, the examiner makes a recommendation (additional tax, no change, or a refund).

The IRS's tax gap research reveals important sources of noncompliance and sheds light on the potential amounts of unpaid taxes that could be collected under current law. But while the NRP is generally well-designed, it may overstate some types of noncompliance while understating other types. Overstating may occur due to the NPR's reliance on the examiner's recommendation. After the audit is completed, taxpayers can appeal or take the dispute to court, but a resolution in the taxpayer's favor does not reduce the tax gap estimate.[26] While the IRS researchers can potentially monitor post-audit abatements (though those disputes may be lengthy), the more challenging task would be identifying when compliant taxpayers do not dispute the examiner's recommendation because of a lack of resources or fear of the IRS (Guyton, et al. 2024).

One unambiguous omission is unreported income from criminal activities. Noncompliance attributable to illegal-source income is excluded from the tax gap estimates, partly because of the extreme challenges of observing or estimating the gains from crime.

For some types of income, the IRS actively looks for underreporting but may lack sufficient information or resources to detect most of the noncompliance. Income from partnerships and foreign sources is particularly difficult to observe and verify. In both cases, the challenge of verifying income is compounded by difficulties in tracing the income to the owner. That complicated web makes it difficult to trace the partnership income from the entity to the actual partner who is liable for the tax.

To adjust for undetected unreported income, the IRS uses a methodology called detection-controlled estimation (DCE). The DCE is premised on the assumption that the "best" auditors detect the most underreported income and that those best auditors are the ones who recommend the largest upward adjustments in types of personal income, controlling for observable characteristics of the cases assigned to each examiner. But if the "best" examiners are also the ones who are the most aggressive and make questionable recommendations, the tax gap will be overstated.

The DCE adjustments can be substantial. Without the DCE adjustments, annual individual income was estimated to be underreported by $145 billion, on average, over the 2014–2016 period. The DCE adjustments nearly doubled

26 Data from the Treasury Inspector General for Tax Administration (TIGTA) indicate that only 63% of additional taxes recommended by examiners in operational audits in Fiscal Years 2015 through 2019 were ultimately assessed (after administrative appeals and abatements; TIGTA 2021). That figure is likely even lower due to further reductions on judicial review.

the estimate of underreported income—up to $278 billion. The estimate of underreported sole proprietors increased by 135%—from $34 billion to $80 billion—because of the DCE adjustments (GAO 2024). According to GAO, the IRS has not conducted a thorough analysis to determine the causes of these substantial adjustments but has embarked on pilots that may provide more insight. But concerns about the DCE methodology take on greater urgency because of reductions in the NRP sample size—from 14,200 individual income tax returns in 2015 to 4,000 in 2021.

Some countries have adopted other approaches to account for unobserved income. Before 2020, the United Kingdom's HM Revenue and Customs (2020) used the IRS's DCEs to correct for underreporting of income in their tax gap estimates. After an evaluation by the International Monetary Fund in 2013, the United Kingdom began investigating ways to develop multipliers that better reflected the British tax system. Because they found they did not have sufficient observations to build a DCE model, they developed an alternative approach that relies on a panel of experts—including experienced examiners—to estimate how much tax would be undetected in hypothetical audits, involving different types of issues, availability of third-party information, and the degree of cooperation from the taxpayer. This approach uses the Delphi technique, in which experts separately assess the hypothetical cases through a series of rounds, to reach a consensus on the multiplier.

A disadvantage of the UK's approach is that the results might not be replicated with a different group of experts. But it also may yield more information about the vulnerabilities in the tax system and how different types of taxpayers exploit those holes. Supplementing the current method with the UK approach might provide the most useful information.

Assuming sufficient resources, the ideal solution would be to improve the IRS's ability to detect errors on taxpayers' returns. Some of the shortfalls could be addressed with more resources. Improvements in detecting partnership and offshore income are already ongoing, as the IRS and academic researchers collaborate on the application of artificial intelligence techniques to compliance studies. Information about the final resolution of an audit—at least through appeals—could be tracked using the IRS's Enforcement Revenue Information System (ERIS). ERIS, however, does not follow a case after it enters the judicial system.

The impact of improvements in detection and data on resolutions could result in a higher or lower estimate of the tax gap than the current DCE-adjusted measures. It would likely be a more accurate measure, especially for specific areas of the tax code (such as partnerships) and shift the tax gap studies from being a score of the IRS's overall performance to an evaluation of the compliance vulnerabilities in the tax system.

None of those potential solutions, however, addresses a fundamental challenge in the tax system. The tax gap measures do not fully reflect the complexity of the tax code. There are many gray areas in the tax code, where complexity contributes to different interpretations of what legal avoidance is and what illegal evasion is (Hemel, et al. 2022). The gray areas are especially prominent in the tax provisions affecting high-income taxpayers, partnerships, and large businesses—taxpayers who are more likely than others to have the resources to hire tax advisers capable of designing an aggressive strategy open to different interpretations of its legality. A better understanding of the amounts of revenue lost due to aggressive tax avoidance strategies would complement the tax gap and provide a fuller picture of the IRS's ability to enforce the tax code.

### *7.2 Output: Audit Rates*

Audit rates took center stage in the debate about the IRS funding. Supporters of increased funding pointed to the overall reduction in audit rates—particularly among high-income taxpayers and large businesses. Others expressed concern about the potential burdens on compliant taxpayers if they were audited (erroneously) after the IRS's funding increased (Knefel 2022). Adding confusion to the discussion of audit rates was that the IRS changed its definition of audit rate as this debate was ongoing.

Before 2019, the IRS defined the audit rate as the ratio of closed audits in a fiscal year to the number of tax returns filed in the prior calendar year. That measure, however, assumed that within a year of filing, audits began and ended. Increasingly, though, the time gap between filing a return and the closure of an audit has extended beyond a year. As such, those audit rate measures did not measure taxpayers' likelihood of having been audited on their tax return for a specific year.

The IRS introduced a new measure of audit rates in 2019. The revised measure is the share of returns for a given tax year that are ever audited—a cumulative measure that increases over time as more returns from that year are selected for audit. That rate begins to flatten out once the statute of limitations on assessments is past—typically three years after filing.[27] For 2019 tax returns, that point was reached in 2023.

Consider audits of 2019 tax returns filed by taxpayers with $10 million of reported positive income: as of the end of September 2021, just 2% of those returns had been selected for audit; by 2023, that share had grown to 11% (IRS 2022a, IRS 2024c). The lag reflects the complexity of their returns and the challenges involved in determining which returns to audit in that income group.

Like the previous metrics, the audit rates should be viewed in a broader context. The quality and quantity of audits matters. Some insight into the quality of the audit is the "no-change" rate: A "no change" audit happens when a taxpayer can substantiate their claims of income, deductions, and credits—in theory, a signal that the IRS was not efficient in its selection of that return for audit. The percentage of cases closed is another indicator of efficiency.

Yet, neither a no-change rate nor a percentage of closed cases is sufficient to judge the quality of the audit selection technique. For example, as of 2023, 97% of audits of tax year 2019 individual income tax returns had closed (Table 2). Of the closed cases, 12% resulted in no change.

**TABLE 2. Individual Income Tax Audit Rates, Closed Cases, and No-Change Rates**

| Tax Year | All Individual Income Tax Returns | | | $1 Million or More of Positive Income | | |
|---|---|---|---|---|---|---|
| | Audit | Closure | No change | Audit | Closure | No change |
| 2010 | 1.0 | 99.8 | 15.1 | 9.1 | 96.7 | 37.0 |
| 2011 | 0.9 | 99.8 | 12.4 | 7.2 | 96.4 | 37.0 |
| 2012 | 0.8 | 99.8 | 13.3 | 5.5 | 96.7 | 31.9 |
| 2013 | 0.6 | 99.8 | 10.5 | 3.5 | 95.0 | 21.7 |
| 2014 | 0.6 | 99.5 | 9.3 | 3.1 | 90.1 | 21.6 |
| 2015 | 0.6 | 99.3 | 9.3 | 3.0 | 85.4 | 24.9 |
| 2016 | 0.5 | 98.6 | 10.2 | 3.1 | 74.7 | 24.8 |
| 2017 | 0.5 | 98.4 | 11.8 | 2.5 | 74.3 | 25.3 |
| 2018 | 0.3 | 97.7 | 11.8 | 1.8 | 82.4 | 30.4 |
| 2019 | 0.3 | 97.3 | 12.0 | 2.1 | 81.8 | 36.6 |

Source: Author's computations derived from the 2023 IRS Data Book for 2013-2019 tax returns; 2022 IRS Data Book for 2012 tax returns; 2021 Data Book for 2011 tax returns; and 2020 IRS Data Book for 2010 tax returns.

High-income tax returns take longer to audit, and the closure rate, as of 2023, was just 82% for audits of 2019 returns for taxpayers with income above $1 million with a no-change rate of 37%. That high no-change rate may reflect how much longer it takes to begin and complete audits of the type of complicated return that would ultimately have resulted in a change to the taxpayer's return. But even if all the remaining open audits of those very taxpayers led to an adjustment to the taxpayer's tax bill, the no-change rate for their audits would be 30%—more than twice the average no-change for all individual filers.

This example illustrates the limitations of performance measures. First, more than one quantitative metric can be necessary to evaluate the performance of a single activity. And second, more extensive research is needed to put the numbers into context. In this instance, do the higher-than-average no-change rates for high-income taxpayers indicate that the IRS's selection tools are inefficient, or does it mean that the wealthy have more resources to challenge the IRS? Conversely,

27 In some cases, the statute of limitations is extended beyond three years. Those exceptions include (1) failure to file a required tax return; (2) agreement between the IRS and taxpayer to extend the period; (3) taxpayer reported less than 25% of their income on the tax return; and (4) the taxpayer filed a false or fraudulent tax return with intent to avoid taxes. The three-year time limit can be suspended if the IRS issued a notice of deficiency (with the IRS's proposed assessment) or the taxpayer filed for bankruptcy.

do the lower no-change rates of low- and middle-income taxpayers indicate that the IRS selection tools are efficient or that they are too intimated, busy, or budget-constrained to challenge the agency's assessments (Guyton, et al. 2024)? Other types of research methods—such as focus groups or ethnographic studies—may yield information that can place audit rates and no-change rates into context.

### *7.3 Efficiency: Returns On Investment*

In discussions of the IRS's funding, the return on investment has typically been defined as the ratio of the additional tax receipts, interest, and penalties generated by new audit initiatives to the increase in expenditures for those activities. Historically, the ROIs have been limited to the relationship between collections and the salaries of the IRS employees directly involved in the enforcement actions (including examinations, appeals, and collections).

Until recently, the IRS's ROIs garnered little attention beyond the agency and a small circle of budget analysts and officials at the OMB, Treasury, and CBO. For many years, ROI estimates were viewed with some skepticism—partly because data and research were lacking to support the calculations. Another concern was that funding for IRS expansions had sometimes failed to materialize after the first year (as happened in Fiscal Year 1992 after a five-year expansion had been enacted the prior year) or was diverted to other uses (as happened in Fiscal Year 1994, when the initiative's funding was used to pay for unfunded but mandated cost-of-living allowances).

The third reason for the lack of focus on the ROIs was the limitation imposed by the Administration's and Congressional guidelines for inclusion of the effects of spending and revenue bills on the federal deficit in official estimates of the cost or savings of legislation. Those guidelines were formalized in the conference report for the Balanced Budget Act of 1997 and are occasionally updated upon agreement by the House and Senate Budget Committees, CBO, and the OMB. Scorekeeping Rule 14 is particularly relevant to the use of ROIs in budget considerations[28]:

> *No increase in receipts or decrease in direct spending will be scored as a result of provisions of a law that provides direct spending for administrative or program management activities.*

According to CBO, Rule 14 was adopted in part to avoid situations where hoped-for but quite uncertain savings are used to offset near-term certain spending increases or revenue decreases in the same legislation (CBO 2014). The rule applies to all direct spending and revenue proposals.

Nonetheless, legislation on budget processes sometimes permitted appropriators to score the revenues from special "program integrity" initiatives, which allowed an increase in enforcement funds for the IRS (and certain other agencies) above the statutory caps on domestic discretionary spending. For those limited purposes, economists at Treasury and CBO followed a broadly similar methodology for estimating collections from program integrity initiatives.

Both Treasury and CBO start with ROIs provided by the IRS. The IRS derives ROIs from the Enforcement Revenue Information System (ERIS), which was first developed in the 1990s and has expanded since then. The ERIS follows returns through enforcement and collections activities and contains information both on staff hours and the final amounts collected by the IRS. The time it takes to collect the outstanding tax liabilities after the enforcement action and appeal is based on other confidential IRS data. Those collection periods can stretch out over many years.

The IRS-produced ROIs are averages. Research by Holtzblatt and McGuire (2016) described other assumptions used by CBO to transform those averages into marginal ROIs—the amount of revenues attributable to an additional $1 of appropriations—for estimates of IRS program integrity proposals.

- The ROIs would not reach their peak until at least three years after implementation of an initiative because of the time it would take to hire and then train new employees.
- The marginal revenues from an initiative would decline over time as taxpayers discover new ways to avoid or evade tax liabilities at a faster pace than the IRS could develop counterstrategies.

28 https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/a11_current_year/app_a.pdf

- The marginal revenue from a new program would be smaller than the ROI for an earlier initiative because the IRS would first tackle the "low-lying fruit"—cases where detection and resolution of errors were easiest.
- Only the revenues directly resulting from audits and collections would be included in the estimates. The estimates omitted any additional revenues resulting from an increase in voluntary compliance.

The focus on ROIs intensified with the release of a paper in 2019 by Natasha Sarin and Larry Summers (2019). They presented a multifaceted plan to expand the IRS, estimating that a $100 billion infusion of funds over 10 years would generate $1.1 trillion of additional revenue for a net deficit reduction of $1 trillion. They diverged from the standard CBO methodology by including revenue from increasing information reporting and additional investments in technology and by excluding the impact of any type of audited taxpayers' responses. In the absence of access to internal IRS (such as the ERIS), Sarin and Summers extrapolated from published data and characterized their estimation as "naïve."

The Sarin and Summers's paper laid the groundwork for the IRA funding boost of $80 billion—and sparked a new discussion of how to estimate ROIs especially after Treasury's (2021) and CBO's initial estimates of the gross revenues raised by the Administration's $80 billion enforcement proposal differed by about one-third ($316 billion and $200 billion, respectively).[29] Surprisingly, CBO's much smaller estimate in 2021 accounted for two factors that would, on net, increase enforcement revenues but which were excluded from Treasury's estimates: first, an increase—albeit modest—in voluntary compliance and second, the interaction between the funding increase with another proposal (later dropped) to enhance the IRS's ability to detect noncompliance by requiring more information reporting from financial institutions.[30]

Since then, the IRS and Treasury have revamped their methodology for estimating ROIs and revenue effects of IRS funding (IRS 2024c). In combination, the revisions boost the Treasury revenue estimates of IRA's effects by 27%.

That revised estimate is somewhat remarkable because the IRS accounted for costs that were omitted in previous estimates. For example, ERIS understates labor costs because it does not include fringe benefits and the time spent by more than one employee on a case at each stage of the enforcement activity—such as supervisors or others who may be brought in to assist or review a case. Nor does ERIS include the fixed costs associated with new hires—rent for additional space, laptops, and so forth.

Some researchers have incorporated those additional costs into their ERIS computations. Holtzblatt and McGuire (2020) added the costs of fringe benefits, while research by Boning, et al. (2023) also included labor costs of supervisors and other employees who supported the work of those people directly involved in the enforcement activity. The research by Boning, et al. (2023) also incorporated expenditures attributable to office space and information technology costs as well as expenses incurred by other government agencies. Whereas CBO's estimate of the ROI peaked at about $7 for an additional $1 in funding, Boning, et al. determined that the overall ROI would be $2 once all costs are included. For those in the top 10%, the pre-voluntary compliance ROI would be $3 for an additional $1.

But the revised IRS's estimates reflect other factors that more than offset the additional costs. First, the revision accounts for improvements in the IRS's ability to detect noncompliance and efficiently allocate workloads. A second adjustment reflects changes in the assumptions about the IRS's productivity over time. The IRS disputes CBO's assumption that ROIs will decline over time—both because of enhancements in audit efficiency and the large backlog of unworked cases.

Finally, the IRS includes improvements in voluntary compliance after audits—though their analysis is far more optimistic than CBO's assumptions in 2021. After reviewing past compliance research, CBO (CBO 2020) concluded voluntary compliance increased overall in response to audits, but that compliance by higher-income individuals—one of the targets of the Administration's plan for enhanced enforcement—did not improve. The IRS's revised estimates are based on newer research by Boning, et al. (2023) that finds voluntary compliance to rise among taxpayers in all income groups. They find that the marginal ROI for taxpayers in the top 10% of the income distribution increases from 3:1 to 12:1 after accounting for the deterrence effect among audited taxpayers. The authors note that the ROI might be even higher if they also reflect-

29 Philip Swagel, "The effect of increased funding for the IRS," CBO Blog, Congressional Budget Office, September 2, 2021, https://www.cbo.gov/publication/57444.

30 Secretary of the Treasury Janet Yellen, Letter to Ways and Means Chairman Richard Neal, September 14, 2021. https://home.treasury.gov/system/files/136/Yellen_Neal_Congressional_Budget.pdf.

ed the indirect effects of audits—that is, the extent to which audits deter noncompliance by people who are not audited.

The IRS has also laid out a vision of future extensions of measures of the ROI to incorporate other types of activities, including non-audit enforcement actions, taxpayer services, and modernization of technologies. Because the IRS lacks data that explicitly links the costs of these activities to the resulting revenue, they provide documentation of the success of similar efforts—not limited to tax agencies—in other countries, states, and the private sector and, in some cases, the amount of revenue or cost-savings achieved by those activities. Based on those studies, the IRS estimated the potential savings for two of the initiatives—notices to prompt taxpayers to make estimated payments and improvements in information technologies. Including those projected savings would double their estimates of the revenue from IRA's funding. Those estimates, however, should be viewed as speculative because neither the full scope of the initiatives nor their costs is detailed in the IRS report.

Notably, the discussion regarding the evidence about the potential revenue gains from improvements in taxpayer services is sparse compared with the other potential activities, and the IRS did not provide an estimate of revenues resulting from an enhancement of taxpayer services. An analysis by Mazur and Sarin (2024) suggests that an additional dollar spent on taxpayer services—such as staffing telephone call centers and walk-in taxpayer assistance sites—could yield at least an additional $2 in revenue, though evidence to support that estimate is scarce.

Another direction is to recognize that some IRS's activities—especially in taxpayer services—have intangible benefits that may or may not affect voluntary compliance. Calling the IRS to get confirmation of one's interpretation of a tax law may not result in any change in reported taxes (especially if the taxpayer's initial interpretation was correct)—and thus show zero returns for the monetary costs incurred by the IRS. But a good interaction between the caller and the customer service agent can generate goodwill for the IRS with perhaps a positive spillover for trust in the federal government.

Finally, while attention lately has mainly focused on ROIs to support additional funding, the metric is also used to inform choices between specific activities. Research by Hodge et al. (2015) estimate ROIs for correspondence audits with different targets to determine which maximizes net direct revenues.

## 8. Equitable Tax Administration

The third goal identified in the IRS mission statement is for the agency to enforce the law with integrity and fairness to all. That goal is perhaps the most difficult to monitor. There are many dimensions of equity, but key elements—such as true income (which includes unreported income), race, and ethnicity—are not immediately observable by the IRS. Those barriers to measuring equitable tax treatment are compounded by the challenges detailed above for estimating the overall performance metrics. Several researchers have attempted to overcome those barriers to examine the distribution of compliance burdens, the tax gap, and audit rates.

### *8.1 Outcome: Distributing Compliance Burdens by Income*

Researchers at the IRS and the Tax Policy Center have used the IRS's individual compliance model to distribute the compliance burden by income. After monetarizing the time costs incurred by filers, research by Marcuss et al. (2013) shows that the average compliance cost as a share of adjusted gross income falls as income rises. Research by Berger, et al. (2018) generally shows similar results using a more comprehensive definition of income, but they also find that the average ratio of cost to income is equally high among families in the bottom income quintile and those in the top 95 to 99% of the income distribution.

Still, the challenges found in interpreting the aggregate compliance burden become even more problematic when distributing the costs by income. Compliance costs for low-income individuals might be understated because of the lack of data on nonfilers—some of whom might have tried to file to claim a refund but gave up because they did not understand how to fill out a return or how to seek out assistance. Similarly, the burden measures might not capture the costs incurred by filers who begin but do not complete a form or worksheet (for example, to determine if they should pay the alternative minimum tax).

### *8.2 Outcome: Distributing Noncompliance by Income*

Analyses related to the distribution of noncompliance have typically been the byproduct of research on the distribution of income. Some researchers have turned to the NRP to fill in gaps on unobserved income—income neither reported on tax forms nor on household surveys—and their findings provide some insight into the distribution of underpayments of taxes. But although the authors of the studies begin with the same data—the NRP—their results differ:

- Research by Johns and Slemrod (2010) found that the percentage of true income not reported to the IRS increases as "true" adjusted gross income grows but peaks among taxpayers in the top 99 to 99.5 percentile.
- Underreporting is relatively constant across most of the distribution of true income but declines among taxpayers with more than $5 million, according to research by DeBacker, et al. (2020).
- Research by Guyton, et al. (2023) estimate that underreported income as a fraction of true income rises from about 10% in the bottom 90% of the income distribution to 16% in the top 1% where it remains constant or falls.

In large part, those differences result from the authors' treatment of income that is neither reported nor detected by the IRS. While the Johns and Slemrod research distributes DCE-adjusted income, DeBacker, et al. (2020) argue that the DCE is, at best, an adjustment to aggregate income and was not designed to correct for underreporting by individual taxpayers. In their preferred analysis, they distribute unreported income prior to the DCE adjustment. Guyton, et al. start with DCE-adjusted income and add in their estimates of undetected income from partnerships and offshore accounts. Their distributional findings are driven not only by the addition of those two sources of unreported income but also by their assumptions that most of the undetected income from partnerships and offshore accounts is earned by the very highest-income taxpayers. Those assumptions are disputed in the paper by Auten and Splinter (2024).

Perhaps the most telling comment on this research is found in an appendix to the paper by Guyton, et al. (2023). They demonstrate how the results differ depending on various assumptions about the distribution of undetected income and conclude: "Finally, in light of all the uncertainty here, we can understand why some readers may wish to give up on DCE, at least for distributional analysis" (Guyton, et al. 2023, pp. 38). Nonetheless, the authors remain in the camp of distributing DCE-adjusted income.

What are the implications for understanding the distribution of the tax gap? Of the authors, only Johns and Slemrod estimate the distribution of unpaid taxes. They find that although the percentage of unreported income increases as true income grows, unpaid income taxes as a share of the actual tax liability fall as income rises.

In contrast, an analysis by Sarin (2021) begins with the DCE-adjusted unreported income—the result of a sensitivity test described in an appendix to DeBacker, et al. (2020)—and estimates that the top 1% of taxpayers are responsible for 28% of the tax gap.[31] Notably, though, she computes the distribution of the tax gap by applying the percentages of unreported individual income by filers in each decile to the aggregate tax gap—which also includes noncompliance from not filing an income tax return or underpaying income taxes as well as noncompliance attributed to payroll taxes, corporate income taxes, and estate taxes.

The distribution of noncompliance is an important outcome of a tax system that is equitable in its treatment of taxpayers. But the analysis of this question is deeply intertwined with the methodological design of tax gap studies.

### *8.3 Output: Distributing Audit Rates by Race*

Taxpayers are not asked to state their race or ethnicity on tax returns or other forms supplied to the IRS, and very few individuals interact in person with an IRS employee. Those factors contribute to a perception that the tax system is race-blind, but the lack of data also makes it difficult to determine whether the tax system treats individuals fairly, regardless of their race and ethnicity. To broaden the examination of disparities in the income tax system, researchers are developing methodologies to add race and ethnicity imputations to tax data.

31 Natasha Sarin, "The Case for a Robust Attack on the Tax Gap," Featured Stories, US Department of the Treasury, September 7, 2021, https://home.treasury.gov/news/featured-stories/the-case-for-a-robust-attack-on-the-tax-gap.

Those methodological breakthroughs enabled researchers at Stanford University and the Treasury Department to investigate racial disparities in the selection of tax audits (Elzayn 2023). They found that Black taxpayers were three to five times more likely to be audited than non-Black taxpayers—largely due to differences in the audit rates of claimants of the earned income tax credit (EITC).

That study also highlights at least two challenges for measuring racial disparities in the tax code. The first is the need to validate new methodologies. The Stanford and Treasury researchers relied on a technique called the Bayesian Improved First Name and Surname Geocoding (BIFSG), first developed to analyze racial disparities in health care. BIFSG imputes race and ethnicity based on first and last names and location (in this case, Census block). Research by Derby, et al. (2024) identified flaws in this approach that can lead to overstating the probability of non-White individuals being identified as White. In the case of the study of audit rates, their finding suggests that the Stanford-Treasury study understated the racial disparities in audit selection. Because Derby, et al. (2024) relied on a sample of low-income families that was not nationally representative, their findings are not conclusive.

Derby, et al. (2024) indicates that the development and implementation of race and ethnicity imputations on tax data is still evolving. In June 2024, the IRS announced an agreement with the Census Bureau, which will provide privacy-protected race and ethnicity data to the IRS and could replace the BIFSG method in the future (Anderson 2024).

The second challenge points again to understanding the context of the metrics. Without further analysis, the finding that audit rates are disproportionately higher among Black taxpayers than other filers is open to many interpretations. Since the release of the study, IRS researchers (Anderson, et al. 2024; Hertz, et al. 2023) have been delving deeper into the audit selection process to identify the sources for the racial disparities in EITC audits and have thus far found that contributing factors include incomplete data on eligibility criteria (including the child's relationship to the taxpayer), concentration of high-risk tax preparers in non-White neighborhoods, and the IRS's emphasis on selecting returns based on overclaimed refundable tax credits rather than understated income taxes.

The analysis of the racial disparities in audit rates demonstrates the important role played by a combination of a performance measure and in-depth analysis. Based on the findings to date, the IRS has taken steps to refine its audit selection methods—including expanding information on children's relationships to claimants and development of a new EITC risk scoring system (Anderson, et al. 2024). This is a promising area for future advances in practice based on evidence.

## 9. Conclusion

The IRS should—and will—be held accountable for the substantial infusion of IRA funds over the next decade. That requires the IRS to measure and report on its progress—especially with respect to meeting the broad goals set by its mission statement for taxpayer services, enforcement, and equitable and fair treatment of taxpayers. However, performance measures outlined in the IRS's 2023 Strategic Operating Plan and a 2024 update are incomplete. Moreover, other performance measures that predate the plan are a patchwork of sometimes unrelated items that were developed in response to legislative mandates—laws that typically applied to all government agencies. There are also shortcomings in some measures that can lead to misinterpretations by outside observers.

In this paper, I identify shortcomings of nine of the current metrics. Those metrics, however, form the foundation for measures that with key refinements would better inform evaluations of a transformed IRS.

***Taxpayer satisfaction.*** The IRS currently measures taxpayers' overall satisfaction with the agency, based on survey data collected by an outside organization that monitors consumers' attitudes in many different private and public sectors. But this overall measure provides no insight into the sources of taxpayers' satisfaction and dissatisfaction. User experience surveys, with more detailed questions, could yield useful information that would better guide the IRS in its interactions with taxpayers.

***Compliance burden.*** The IRS measures the costs to taxpayers of filling out forms issued by the agency. Those estimates satisfy the government-wide requirements of the Paperwork Reduction Act. But taxpayers interact with the IRS in many

ways—for example, waiting for their call to be answered or searching for information about notices on the IRS website—and those other costs are not included in the current measures. Moreover, the name "compliance burden" is a misnomer, because the measures do not distinguish between the costs of compliant and noncompliant taxpayers.

Expanding the survey to cover other types of activities would fill in some of the information about the costs incurred by taxpayers in their interactions with the IRS. Linking tax burden and tax gap data would provide insight into the trade-offs between the IRS's service and enforcement missions, informing the IRS's and lawmakers' decisions as how to allocate funds between the two sets of activities.

***Telephone calls.*** At least twice a year, the IRS releases information on the percentage of calls that are answered by customer service agents. But one figure covers just the filing season, and the other is for the entire year, and they can greatly differ because of changes in service priorities throughout the year. Providing information about the IRS's service performance throughout the year could reveal more about the trade-offs that the IRS makes—between answering the phones and, for example, opening and responding to the mail. Moreover, bundling the data on answer calls with information about the length of the call, the accuracy of the information, and whether the taxpayer's question was resolved would provide context for quality as well as for quantity.

***Tax gap.*** Examiners cannot observe all underreported income, nor do they typically have the incentives to uncover unclaimed tax benefits. However, the current methodology will overstate the noncompliance if adjustments are based on decisions made by the most aggressive examiners who may err in favor of the IRS. Improving the IRS's ability to detect noncompliance or unclaimed benefits—through trained examiners and technological improvements, such as artificial intelligence—would reduce the agency's reliance on its current statistical methods. The improvements in detection methodologies might increase or decrease the estimates of the tax gap but would potentially provide more insight into the sources and magnitudes of noncompliance.

***Audit rates.*** Audit rates are an oft-cited quantitative measure, but they provide no information on the quality of the activity. Two other measures can supplement audit rates, though rarely receive the same attention: the percentage of cases closed and the no-change rate (the percentage of audits resulting in no change to the tax return). Yet even the bundling of those three statistics is insufficient to evaluate the quality of audits. A high no-change rate may mean that the audit selection is not well-targeted or that taxpayers have the resources to successfully challenge an examiner's finding of underpaid taxes. Accompanying metrics with more in-depth research—possibly ethnographic studies of audited taxpayers—would shed more light on the quality of audits as well as their quantity.

***Return on investment.*** Until the recent debates over the budget shortfalls and funding boosts, little attention was paid to the returns on investments in the IRS. The new interest in ROIs has also focused on the shortcomings in the historic measures: not all returns to funding are included in the estimates, but neither have all costs. A more comprehensive measure of ROIs could better inform decisions about the level and allocation of the IRS's budget. But singling out the ROI does not acknowledge the nonmonetary returns to investments in the IRS.

***Fairness measures.*** As the discussion above reveals, it is challenging to measure the IRS's effectiveness in meeting its service and enforcement missions. Those challenges are compounded when analyzing the distribution of the services and enforcement metrics. Improving the aggregate measures and considering how to more accurately capture differences by income or racial groups will provide more insight into whether the IRS also meets its mission to treat taxpayers fairness.

Developing or refining mission-related, comprehensive IRS metrics is essential. It will require resources (or diversion of resources from other IRS tasks) when the IRS already faces many challenges to achieving its goals. But improving performance metrics is also an investment. As a first step, IRS could be more transparent when releasing performance metrics—for example, by identifying the omitted data in a metric or by discussing the relationships between metrics. Those first steps can build a foundation for more fundamental changes to the way the IRS measures its success and identifies its vulnerabilities, enabling the agency to better achieve its mission goals of providing support to taxpayers, enforcing the tax code, and treating all taxpayers fairly.

## References

American Customer Service Index. 2023. American Customer Service Index Federal Government Report 2023. Ann Arbor, MI: American Customer Service Index.

Anderson, Brandon, Keegan Brown, Damon Frezza, Zane Girouard, Alissa Graff, Tom Hertz, Navya Kambalapally, Piper Kurtz, Kara Leibel, Justin Nave, Mark Payne, Daniel Rodriguez, and Brian Sartain. 2024. "Research on Race and Ethnicity: 2024 Update." Presentation given at the 14th Annual IRS/TPC Joint Research Conference on Tax Administration, Washington, DC, June 13.

Auten, Gerald, and David Splinter. 2024. "Income Inequality in the United States: Using Tax Data to Measure Long-Term Trends." Journal of Political Economy 132 (7): 2179–227. https://www.journals.uchicago.edu/doi/full/10.1086/728741.

Berger, Daniel, Eric Toder, Victoria Bryant, John Guyton, and Patrick Langetieg. 2018. Estimating the Effects of Tax Reform on Compliance Burden. Washington, DC: Urban-Brookings Tax Policy Center.

Boning, William, Nathaniel Hendren, Ben Sprung-Keyser, and Ellen Stuart. 2023. "A Welfare Analysis of Tax Audits across the Income Distribution." Working Paper 31376. Cambridge, MA: National Bureau of Economic Research.

CBO (Congressional Budget Office) 2014. How Initiatives to Reduce Fraud in Federal Health Programs Affect the Budget. Washington, DC: CBO.

------. 2020. Trends in the Internal Revenue Service and Enforcement Funding. Washington, DC: CBO.

DeBacker, Jason, Bradley Heim, Ahn Tran, and Alexander Yuskavage. 2020. "Tax Noncompliance and Measures of Income Inequality." Tax Notes Federal: 1103–1118. https://www.taxnotes.com/tax-notes-federal/compliance/tax-noncompliance-and-measures-income-inequality/2020/02/17/2c3y5

Derby, Elena and Connor Dowd and Jacob Mortenson. 2024. "Statistical Bias in Racial and Ethnic Disparity Estimates Using BIFSG." Washington, DC: Joint Committee on Taxation.

Elzayn, Hadi, and Evelyn Smith, Thomas Hertz, Arun Ramesh, Robin Fisher, Daniel Ho, and Jacob Goldin. 2023. "Measuring and Mitigating Racial Disparities in Tax Audits." Working Paper. Palo Alto, CA: Stanford Institute for Economic Policy Research.

GAO (General Accounting Office). 2000. "IRS Is Working to Improve Its Estimates of Compliance Burden." GAO/GGD-01-11. Washington, DC: GAO.

------. 2024. "Tax Gap: IRS Should Take Steps to Ensure Continued Improvements in Estimates." GAO-24-106449. Washington, DC: GAO.

Guyton, John, and Ronald Hodge II. 2013. "The Compliance Costs of IRS Post-Filing Processes." Papers Given at the 2013 IRS-Tax Policy Center Research Conference, Urban Institute, Washington, DC, June 20.

Guyton, John and Patrick Langetieg, Daniel Reck, Max Risch, and Gabriel Zucman. 2023. "Tax Evasion at the Top of the Income Distribution: Theory and Evidence." Washington, DC: Internal Revenue Service.

Guyton, John and Kara Leibel, Dayanand Manoli, Ankur Patel, Mark Payne, and Brenda Schafer. 2024. "The Effects of EITC Correspondence Audits on Low-Income Earners." In Tax Policy and the Economy, vol. 38, edited by Robert Moffitt, 163–207. Chicago, Illinois: University of Chicago Press.

Hemel, Daniel, Janet Holtzblatt, and Steven Rosenthal. 2022. The Tax Gap's Many Shades of Gray. Washington, DC: Urban-Brookings Tax Policy Center.

Hertz, Tom, Brian Sartain, Kara Leibel, and Mark Payne. 2023. "Differences in Audit Rates by Race." Presentation given at the 13th Annual IRS/TPC Joint Research Conference on Tax Administration, Washington, DC, June 22.

HM Revenue and Customs. 2020. "Non-detection Multipliers for Measuring Tax Gaps." Working Paper. London, UK: HM Revenue and Customs.

Hodge, Ronald II, Alan Plumley, Kyle Richison, Getaneh Yismaw, Nicole Nisek, Matt Olson, and H, Sanith Wijesinghe. 2015. "Estimating Marginal Revenue/Cost Curves for Correspondence Audits." Presentation given at the 2015 IRS/TPC Research Conference on Tax Administration, Washington, DC, June 18.

Holtzblatt, Janet, and Jamie McGuire. 2016. "Factors Affecting Revenue Estimates of Tax Compliance Estimates." Working Paper 2016–5. Washington, DC: Congressional Budget Office.

------. 2020. Effects of Recent Reductions in the Internal Revenue Service's Appropriations on Returns on Investment. Washington DC: Urban-Brookings Tax Policy Center.

IRS (Internal Revenue Service). 2022a. Comprehensive Taxpayer Attitude Survey 2021 Executive Report. Publication 5296 (Rev. 4–2022). Washington D.C.: IRS.

------. 2022a. Internal Revenue Service Data Book, 2021. Publication 55-B. Washington, DC: IRS.

------. 2022b. Tax Gap Estimates for Tax Years 2014–2016. Publication 1415 (Rev. 10–2022). Washington, D.C.: IRS.

------. 2023a. Taxpayer Compliance Burden. Publication 5743 (Rev. 4–2023). Washington, DC: IRS.

------. 2023b. IRS Inflation Reduction Act Strategic Operating Plan. Publication 3744 (Rev. 4–2023). Washington, D.C.: IRS.

------. 2024a. Fiscal Year 2025 Congressional Budget Justification and Annual Performance Report and Plan. Publication 4450 (Rev. 2–2024). Washington, DC: IRS.

------. 2024b. Return of Investment: Re-Examining Revenue Estimates for IRS Funding. Publication 5901 (2–2024). Washington, DC: IRS.

------. 2024c. Internal Revenue Service Data Book, 2023. Publication 55-B. Washington, DC: IRS.

------. 2024d. IRA Strategic Operating Plan: Annual Update Supplement. Publication 3744-A (4–2024). Washington, DC: IRS.

Johns, Andrew, and Joel Slemrod. 2010. "The Distribution of Income Taxes Noncompliance." National Tax Journal 63 (3): 397–418. https://www.journals.uchicago.edu/doi/abs/10.17310/ntj.2010.3.01.

Marcuss, Rosemary, George Contos, John Guyton, Patrick Langetieg, Allen Lerman, Susan Nelson, Brenda Schafer, and Melissa Vigil. 2013. "Income Taxes and Compliance Costs: How Are They Related?" National Tax Journal 68 (4): 833–53. https://www.journals.uchicago.edu/doi/10.17310/ntj.2013.4.03.

National Taxpayer Advocate. 2023. Annual Report to Congress. Washington, DC: Taxpayer Advocate Service.

Rifaat, Alexander. 2024. "Conflicting IRS Level-of-Service Figures Raise Questions." Tax Notes. 183. April 22, https://www.taxnotes.com/tax-notes-today-federal/tax-system-administration/conflicting-irs-level-service-figures-raise-questions/2024/04/16/7jf9n.

Sarin, Natasha, and Mark Mazur. 2024. "The Inflation Act's Impact on Tax Compliance and Fiscal Sustainability." Tax Notes. February 19. https://www.taxnotes.com/featured-analysis/inflation-reduction-acts-impact-tax-compliance-and-fiscal-sustainability/2024/02/16/7j509.

Sarin, Natasha, and Lawrence Summers. 2019. "Shrinking the Tax Gap: Approaches and Revenue Potential." Tax Notes. November 18. https://www.taxnotes.com/special-reports/compliance/shrinking-tax-gap-approaches-and-revenue-potential/2019/11/15/2b47g

TIGTA (Treasury Inspector General for Tax Administration). 2021. Trends in Compliance Activities through Fiscal Year 2019. Report Number 2021-30-011. Washington, DC: TIGTA.

------. 2023. Final Results of the 2023 Filing Season. Report Number 2024-400-006. Washington, DC: TIGTA.

------. 2024. Quarterly Snapshot: The IRS's Inflation Reduction Act Spending Through March 31, 2024. Report Number 2024-IE-R015. Washington, DC: TIGTA. https://www.tigta.gov/sites/default/files/reports/2024-06/2024ier015fr.pdf

US Department of the Treasury. 2021. The American Families Plan Tax Compliance Agenda. Washington, DC: Treasury.

Werfel, Daniel. 2023. "The Filing Season and the IRS Budget." Written Testimony before the US House Ways and Means Committee, Washington, DC, April 27.

# Tools To Promote Trustworthiness in a Prototype AI System at the IRS

*Michael Szulczewski, Michael Feldman, and Steffani Silva (MITRE);*[1]
*Alissa Graff and Brandon Anderson (IRS, RAAS)*

## Abstract

The Internal Revenue Service (IRS) is exploring the use of artificial intelligence (AI) to better identify risks of tax noncompliance. While federal guidance directs agencies like the IRS to use AI in a manner that fosters public trust, there are few tools for assuring trustworthy AI that are standardized across the federal government and that can be implemented in AI projects. Here, we consider a prototype AI system we developed at the IRS and explore tools including documentation and software that promote trust in the system. We outline the system, identify stakeholders, define goals for AI trustworthiness based on their needs and federal guidance, and describe the adaptation of tools to satisfy those goals. This study informs and advances the adoption of trustworthy AI by identifying trustworthiness tools, explaining adoption challenges, and demonstrating an approach to overcome those challenges for a real-world use case.

## Introduction

The IRS has been using analytics to support tax administration for at least 50 years (1) and is currently exploring new techniques using AI[2] (5). In the 1970s (6), the IRS began using a modified linear discriminant analysis (LDA) called the Discriminant Function System (DIF) to "score income tax returns for examination potential" (1, 7, Section 4.1.2.6). In the mid-1980s, the IRS established the AI Lab to investigate methods such as artificial neural networks (ANNs) (1, 8), which are machine-learning models inspired by biological neural networks in brains (9). In 2023, the agency adopted a strategic operating plan that targets the improved use of analytics and AI (10, 11).

The increased use of AI by agencies like the IRS has been one of the motivations for the United States (U.S.) federal government to address AI trustworthiness. Since 2019, AI trustworthiness has been highlighted in at least three executive orders, four federal laws, and eight reports from government organizations (3). Executive Order 13960, "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," states that "Agencies must therefore design, develop, acquire, and use AI in a manner that fosters public trust" (12). The National Artificial Intelligence Initiative Act of 2020 states that "the Federal Government must provide sufficient resources and use its convening power" to "drive forward advances in trustworthy artificial intelligence" (4). The AI Accountability Framework for Federal Agencies and other Entities (13) recommends "foster[ing] public trust in responsible use" of AI. The IRS Strategic Operating Plan for Fiscal Years 2023-2031 embraces these directives and recommendations by instituting a project to "Establish trustworthy analytics practices and policies" (10).

Despite increased attention on AI trustworthiness, achieving it for actual AI systems remains a challenge. One challenge is an implementation gap: while there is abundant high-level federal guidance on AI trustworthiness (2, 13–19), there are few tools standardized across the U.S. federal government for achieving it. According to the General Services Administration's "AI Guide for Government," "U.S. agencies have already begun to create high-level AI principles, and

---

[1] Approved for Public Release; Distribution Unlimited. Public Release Case Number 24-2346. NOTICE: This technical data was produced for the U. S. Government under Contract Number TIRNO-99-D-00005, and is subject to Federal Acquisition Regulation Clause 52.227-14, Rights in Data—General, Alt. I, II, III and IV (MAY 2014) [Reference 27.409(a)]. No other use other than that granted to the U. S. Government, or to those acting on behalf of the U. S. Government under that Clause is authorized without the express written permission of The MITRE Corporation. For further information, please contact The MITRE Corporation, Contracts Management Office, 7515 Colshire Drive, McLean, VA 22102-7539, (703) 983-6000. © 2024 The MITRE Corporation. Direct correspondence to Michael Szulczewski (mszulczewski@mitre.org), 781-223-5492. This work reflects the contributions of many MITRE and IRS employees. We thank the following people for their involvement in developing and revising the approach and the data and model cards: Frank Cousin, Lisa Lakata, and Stephanie Needham. We thank Steve Dorton for an insightful discussion about AI trustworthiness.

[2] "[N]either the scientific community nor industry agree on a common definition" of AI (2). There are at least six definitions published by the Federal Government (3). We adopt the definition in the National Artificial Intelligence Initiative Act of 2020: "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine and human based inputs to (1) perceive real and virtual environments, (2) abstract such perceptions into models through analysis in an automated manner, and (3) use model inference to formulate options for information or action." (4).

even some policies around AI's responsible and trustworthy use…but next these principles must be translated into actionable steps that agencies can use" (2). Some steps are currently under development (19).

Here we describe steps to close the implementation gap for a prototype AI system we developed at the IRS. We first summarize the model used in the system and its inputs and outputs. Next, we describe our adaptation and use of three tools to foster trustworthiness in the system: model cards, data cards, and AI explainability methods. Model and data cards are documents that provide information about AI models and their training data in standardized, easy-to-read formats and are akin to nutrition labels (20), drug fact labels (21), or safety data sheets (22). AI explainability methods provide post hoc explanations of the predictions of AI models. After describing our adaptation and use of these tools, we explain lessons learned from the project and describe future challenges.

Our work provides two novel contributions to AI documentation and explainability. For AI documentation, our novel contribution is surveying and assessing AI-documentation needs for IRS stakeholders and combining and adapting previous versions of data and model cards from government, industry, and academia into versions that meet those needs (see Appendix). Even though we developed the cards for IRS stakeholders, we believe they are sufficiently general to be useful for other agencies. Our novel contribution for explainability is surveying and assessing the explainability needs of IRS stakeholders and identifying and testing methods to meet those needs.

## AI System

We addressed AI trustworthiness for a prototype AI system we developed under the Research, Applied Analytics, and Statistics organization of the IRS. The system's goal was to better identify tax-noncompliance risk for enterprises. An enterprise is "a network of flowthrough entities and their owners whose economic activity is under the control (defined as 50 percent or more direct or indirect ownership) of a single taxpayer or married couple" (23). A flow-through entity, also known as a passthrough-entity (PTE), is a tax entity such as a partnership, subchapter S corporation (S corporation), or trust that generally has the right to pass net income or losses untaxed to their partners, shareholders, or beneficiaries (23, 24). We focused on enterprises controlled by high-income individuals, who we defined as taxpayers whose total positive income exceeded several million U.S. dollars.

The AI system focused on enterprises because their tax-noncompliance risk is poorly quantified but likely significant. According to the IRS's Strategic Operating Plan for Fiscal Years 2023–2031, the "IRS does not know what portion of the $290 billion net tax gap is network related" (10). Despite this uncertainty, however, it is likely that PTE networks represent a significant percentage of the gap. For Tax Years 2014 to 2016, the IRS estimated that 9% of the tax gap for individual income tax underreporting and 5% of the gross tax gap were derived from passthrough income (25). The percentage of the tax gap attributed to PTEs may be rising because audits for PTEs have significantly decreased: "While the IRS audited 4.4% of passthrough entities in 2010, that number fell to 0.1% in 2017 (the most recent tax year with nearly all audits closed), and audits have continued to decrease." (10). As a result, Guyton et al. state that "Understanding noncompliance involving pass-throughs is … essential" (26).

To quantify tax-noncompliance risk for enterprises, the AI system used machine learning models called graph neural networks (GNNs). GNNs are types of ANNs that are uniquely suited to modeling network data (27–30). One GNN in the system classified whether the exam of a high-income individual that controls an enterprise would result in a change to the individual's tax liability. A second GNN predicted the additional tax an exam of the controlling individual would recommend or assess.

The GNNs were useful for testing the trustworthiness of AI tools because, as with other ANNs, people cannot explain their predictions (31–34). These types of models are called black-box models to signify that they "hide their internal logic to the user" (35). The opposite types of models are called glass-box (36), white-box (37), interpretable (31), or self-interpretable (16) models, and they expose "how variables are jointly related to form the final prediction" (31). Since black-box models are commonly used and since the ability to explain a model's predictions is often a "gold standard for building trust" (38), the GNNs in our AI system presented a common yet challenging use case.

## Development and Use of Tools for AI Trustworthiness

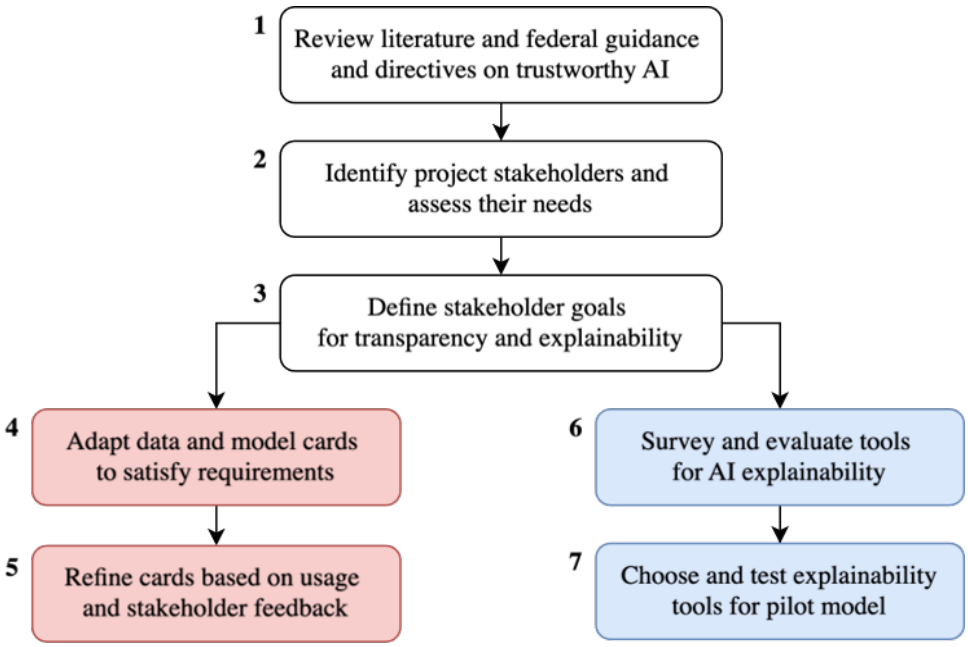
To develop and implement tools for AI trustworthiness, we used practices from systems engineering (39), federal frameworks for AI (e.g., (13, 14)), and trustworthy AI literature (e.g., (32)). We performed the steps shown in Figure 1. For Step 1, 'Review literature and federal guidance on trustworthy AI,' we reviewed over 60 journal articles on trustworthy AI (e.g., (32, 33, 40–42)) and the 23 sources of U.S. federal guidance shown in Table 1.

### *Review of Trustworthy AI Literature and Federal Guidance*

AI literature and federal guidance identify more than 20 characteristics of AI systems that promote trustworthiness (43)[3]. As of 2019, these characteristics have been described in more than 84 documents (46). They include fairness, accuracy, robustness, resiliency, transparency, and explainability (32). They are often ambiguously defined (33, 48), may not be mutually exclusive, and may exhibit complicated dependencies (32).

**FIGURE 1:**

We performed the first three steps below (white boxes) to identify AI trustworthiness goals for the prototype AI system. We performed the subsequent steps to adapt and use data and model cards (red boxes), and to identify and test explainability tools (blue boxes).




While there are many characteristics of trustworthy AI, we focus on transparency and explainability because they are frequently identified as important (13, 32, 42, 43, 45, 46, 49–52) (Figure 2). Researchers have studied explainability since the 1970s and transparency since at least the early 1990s as key components of trustworthy AI systems (53, 54). The Ethics Guidelines for Trustworthy AI (47) identifies transparency and explainability as "requirements that AI systems should meet in order to be deemed trustworthy." The Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence supports AI regulation "related to the transparency of AI models and regulated entities' ability to explain their use of AI models" (55). The Artificial Intelligence Risk Management Framework (AI RMF) states "Characteristics of trustworthy AI systems include: … transparent, explainable" (14). The IRS described "trustworthy analytics" as "yield[ing] transparent, explainable … outcomes" (10).

[3] Characteristics of trustworthy AI (14) may be described as principles (12, 32, 44, 45) or guidelines (43, 46, 47) for trustworthy AI.

While transparency and explainability are important for trustworthy AI, their definitions are often ambiguous or overlapping. For example, explainability is sometimes used synonymously with interpretability, and transparency is sometimes used synonymously with comprehensibility—but other times these terms represent distinct concepts (54, 56, 57). Here, we adopt definitions that are succinct and published by an authority on standards, the International Organization for Standardization (ISO). According to the ISO, transparency is the "property of a system that appropriate information about the system is made available to relevant stakeholders" (51). Explainability is the "property of an AI system to express important factors influencing the AI system results in a way that humans can understand … It is intended to answer the question 'Why?' without actually attempting to argue that the course of action that was taken was necessarily optimal" (51).

While transparency and explainability are important, choosing or adapting appropriate tools to achieve them is challenging. One challenge is a proliferation of options (2, 42). There are at least 22 tools that support transparency by documenting AI datasets or models and new tools are published every year (42). There are at least 17 tools for explainability of GNNs (68)—the type of model used in our system (see AI System)—and many more for other types of models (40, 56, 56, 69, 70). Another challenge is limited standardization. Tools for documenting datasets have different names despite overlapping use cases: they are called datasheets (71), augmented datasheets (72), Data Cards (73), dataset cards (74), Dataset Nutrition Labels (20), data briefs (75), and data statements (76).[4] Of the tools for documenting models, at least three have the same name of 'model card' (77–79), but each has different structure and content. Documentation tools may be document-based (71, 80, 80) or software-based (81); may include questionnaires, checklists, or figures; and often vary in length, detail, and intended stakeholders (42). Due to limited standardization and the growing number of tools, choosing appropriate tools requires a time-consuming evaluation of many different options. If available tools do not meet project needs, adapting existing tools requires additional time to select appropriate parts from them, customize the parts, integrate the parts, and test the adapted tools.

**TABLE 1: Resources for U.S. federal guidance and directives we reviewed to create data and model cards**

| Title | Publication Year | Organization |
|---|---|---|
| "AI Risk Management Framework" (14) | 2023 | NIST |
| "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (58) | 2022 | NIST |
| "Four Principles of Explainable Artificial Intelligence" (16) | 2021 | NIST |
| "Trust and Artificial Intelligence" (49) | 2020 | NIST |
| "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools" (50) | 2019 | NIST |
| "Artificial Intelligence: Emerging Opportunities, Challenges, and Implications" (59) | 2018 | GAO |
| "Artificial Intelligence: Agencies Have Begun Implementation but Need to Complete Key Requirements" (3) | 2023 | GAO |
| "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" (19) | 2024 | OMB |
| "Guidance for Regulation of Artificial Intelligence Applications" (18) | 2020 | OMB |
| "Open Data Policy-Managing Information as an Asset" (60) | 2013 | OMB |
| "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government" (12) | 2020 | EOP |
| "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (55) | 2023 | EOP |
| "Maintaining American Leadership in Artificial Intelligence" (61) | 2019 | EOP |
| "Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People" (15) | 2022 | OSTP |
| "AI Guide for Government" (2) | 2022 | GSA |
| "Treasury Strategic Plan 2022-2026" (11) | 2023 | TREAS |

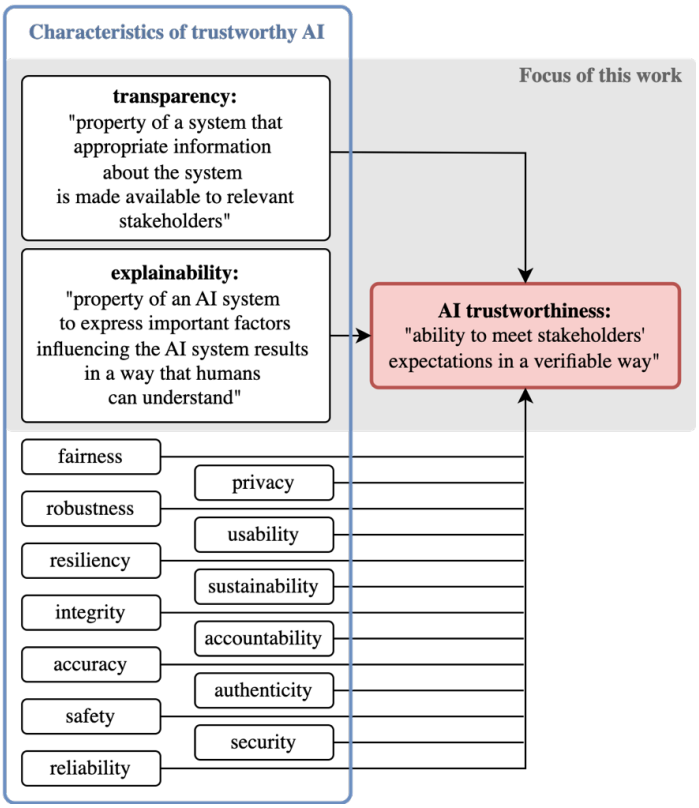
[4] We call our dataset document a data card because the term is succinct and symmetric with the term model card, and because the term is used in the OMB's proposed memorandum, "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" (19)

| Title | Publication Year | Organization |
|---|---|---|
| “Federal Data Strategy 2021 Action Plan” (62) | 2021 | OMB, OSTP, DOC, SBA |
| “Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem” (63) | 2023 | NAIRR |
| “National Artificial Intelligence Advisory Committee Year 1 Report” (64) | 2023 | NAIAC |
| “Data, Analytics, and Artificial Intelligence Adoption Strategy Accelerating Decision Advantage” (65) | 2023 | DOD |
| “Ethical Principles for Artificial Intelligence” (66) | 2020 | DOD |
| “National Artificial Intelligence Initiative Act of 2020” (4) | 2020 | Congress |
| “AI in Government Act of 2020” (67) | 2020 | Congress |

Abbreviations: NIST (National Institute of Standards and Technology); GAO (Government Accountability Office); GSA (General Services Administration); DOD (Department of Defense); OMB (Office of Management and Budget); EOP (Executive Office of the President); OSTP (Office of Science and Technology Policy); TREAS (Department of Treasury); DOC (Department of Commerce); SBA (Small Business Administration); NAIRR (National Artificial Intelligence Research Resource Task Force); NAIAC (National Artificial Intelligence Advisory Committee).

**FIGURE 2:**

Federal guidance and AI research identify many characteristics of trustworthy AI (black-outlined boxes) (13, 14, 18, 43, 44, 46, 47, 51, 82). We focus on two that are frequently described as important: transparency and explainability. We adopt the quoted definitions from the International Organization for Standardization (51).




## *Stakeholders and Goals*

To determine the appropriate tools for our prototype, we first identified project stakeholders (Figure 1, Step 2). We compiled potential AI stakeholder roles from more than 40 roles proposed in AI literature (40, 42, 53) and federal guidance

(2, 13, 14)[5]. We focused on roles described in the AI RMF (14) because they are specific[6] and are mapped to stages of an AI lifecycle. We selected roles for the AI lifecycle stages involved in our prototype and excluded or grouped some roles to align with our scope and personnel (Table 2). After selecting roles, we identified relevant stakeholders for the roles in the IRS. For roles external to the IRS (called external entities in Table 2), we considered stakeholders such as AI researchers and the U.S. federal government, including other agencies.

After identifying stakeholders, we defined their goals (Figure 1, Step 3). We inferred the goals of external stakeholders from AI literature (e.g., (32, 33, 40–42)) and federal guidance (Table 1).[7] We defined the goals of internal IRS stakeholders from interviews and our previous experience. While the goals of IRS and external stakeholders were frequently consistent, we prioritized goals of internal stakeholders. The goals are shown in Table 3.

**TABLE 2:**

We identified 9 stakeholder roles for our AI system. The roles and tasks derive mostly from the AI RMF (14).

| Stakeholder roles | Tasks |
|---|---|
| Leadership | Ensure alignment of AI projects with organizational goals |
| Domain experts (e.g., PTE experts) | Provide deep knowledge about a field |
| Data, software, and AI model engineers | Process data, write software, develop models, and test models |
| AI model-development managers | Ensure data, software, and model engineering meet requirements and communicate with stakeholders |
| Operations and monitoring engineers | Operate and monitor AI systems |
| Operations managers | Manage the deployment and use of an AI system |
| Users (e.g., classifiers, auditors) | Use deployed AI systems to perform IRS duties |
| AI impact assessors | Evaluate AI assurance[8] |
| External entities (e.g., GAO, U.S. Treasury, Inspector General for Tax Administration) | Provide guidance or directives for specifying, managing, or reporting AI risks |

### *Transparency*

To meet stakeholder transparency goals, we developed custom data and model cards (Figure 2, Step 4). Data and model cards are often used for AI documentation in research (84, 85) and have been suggested for use in the federal government (19) (see Transparency). To develop the cards, we started with versions created for federal agencies by the Interagency AI Community of Practice: Responsible AI Working Group (86, 87). From these, we changed the card structures, modified and reorganized their content, and added new content. The changes were informed by and adapted components of data and model cards published by academic and corporate research groups (20, 71, 73, 74, 76, 78–81). The cards from the Responsible AI Working Group were written in Microsoft Word, and we continued using Word because it was familiar to all stakeholders and did not require the procurement, installation, or management of new software. We show our cards in the Appendix.

After making initial versions of the cards, we refined them over the course of a year (Figure 1, Step 5). First, we filled out the cards for the prototype AI system. We found some components were difficult to complete due to ambiguous wording and found some sections were missing relevant questions. We updated the cards to address these limitations. Second, we met with stakeholders periodically to solicit and incorporate feedback. We met with engineers, domain experts, AI

[5] A stakeholder "is a group or individual that is affected by or has a stake in the product or project" (39). Stakeholders in AI may also be referred to as personas (42), AI actors (14), or audiences (40).

[6] Stakeholder roles are often defined at different levels of abstraction. For example, Pushkarna et al. (73) define stakeholder roles for data cards at a high level of abstraction: producers, agents, and users. Preece et al. (53) similarly define stakeholders roles for AI explainability at a high level of abstraction: developers, theorists, ethicists, and users. The AI RMF (14), however, defines stakeholder roles at a lower level of abstraction: data engineers, model engineers, developers, governance experts, organizational management, C-suite executives, trade associations, advocacy groups, etc.

[7] We adopt the definition of stakeholder goals from the NASA Systems Engineering Handbook (39): a goal is "an elaboration of the need, which constitutes a specific set of expectations for the system. Goals address the critical issues identified during the problem assessment. Goals need not be in a quantitative or measurable form, but they should allow us to assess whether the system has achieved them."

[8] AI assurance is "A process that is applied at all stages of the AI engineering lifecycle ensuring that any intelligent system is producing outcomes that are valid, verified, data-driven, trustworthy and explainable to a layman, ethical in the context of its deployment, unbiased in its learning, and fair to its users." (83)

impact assessors, project managers, and leadership about every two weeks. At the end of the project, we distributed the cards to stakeholders in eight other AI projects and incorporated further feedback.

### *Explainability*

While our data and model cards can support explainability, the explainability goals of the IRS stakeholders required considering additional tools. Specifically, Goal 15 in Table 3—"Explain why specific inputs to the model create specific predictions"—could not be satisfied with model and data cards for our GNNs because they are black-box models.[9] As a result, we surveyed and evaluated four explainability methods to provide explanations (88–91) (Figure 1, Step 6). We assessed the methods based on the following criteria: the tasks that could be explained (e.g., node prediction, link prediction, or graph prediction; multiclass classification); the permissible node types (e.g., heterogenous or homogeneous) and link types (e.g., directed or undirected); the extent of testing by the model creators or external parties; the availability and quality of documentation; the ease of integration in our system; and the ease of producing and interpreting explanations. We selected GNNExplainer (91) for implementation.

GNNExplainer is a perturbation-based method (68) that works by assessing the change in a GNN's prediction due to perturbing its inputs. For each prediction, it can identify node features and edges in the input graph that more strongly impact the prediction than others. This information is relevant to the IRS because it indicates tax entities (e.g., partnerships, S corporations) and tax-form line items (e.g., line items in 1065 and 1120-S) in enterprise networks that stakeholders may further evaluate if the GNN predicts tax-noncompliance for an enterprise (see AI System).

**TABLE 3:**

We identified the following stakeholder goals from stakeholder interviews, federal guidance, AI literature, and previous experience.

| | Goal |
|---|---|
| 1 | Use plain language that can be quickly and easily understood by stakeholders with different technical expertise. |
| 2 | Provide brief high-level summaries of the dataset and model. |
| 3 | Describe intended uses and users of the dataset and model. |
| 4 | Describe the composition of the dataset including any quality issues. |
| 5 | Describe how the dataset was collected. |
| 6 | Describe any processing performed on the data and reference applicable code. |
| 7 | Provide locations for the dataset and any additional documentation. |
| 8 | Describe plans for actively maintaining the dataset. |
| 9 | Describe model inputs and outputs and cite any relevant documentation (e.g., data card, metadata). |
| 10 | Describe risks of using the model and how the risks can be mitigated. |
| 11 | Evaluate and contextualize model performance. |
| 12 | Describe any known limitations of the model. |
| 13 | List documents and resources relevant to the model. |
| 14 | Ensure the model is explainable, either by being self-interpretable or by using tools that can produce post-hoc explanations. |
| 15 | Explain why specific inputs to the model create specific predictions. |
| 16 | Ensure model explanations are meaningful to relevant stakeholders. |

To evaluate whether GNNExplainer met the explainability goals (Figure 1, Step 7; Table 3), we presented information it produced about feature and edge importance to IRS stakeholders with filled-out model and data cards as context. The feedback was positive. Stakeholders stated that the information provided support for the GNN's predictions and that they understood how the information could be used in practice. Stakeholder goals did not include explanation accuracy or knowledge limits in this project (16), but these goals would likely be critical to consider in the future.

[9] One possible approach for providing explanations of a black-box model in a model card is to approximate the black-box model with a self-explainable model, and report explainability information for the approximate model in the card. We did not pursue that approach due to concerns about the accuracy of explanations.

## Discussion

### *Lessons Learned*

During the project, we learned several lessons that may be useful for other agencies or other projects at the IRS. One lesson is highlighted in Datasheets for Datasets (71): it is useful to fill out data and model cards at the beginning of a project and then update them regularly until project completion. While team members may not have sufficient information at the project's start to complete the cards, the exercise of filling them out can raise questions related to AI trustworthiness that can inform development (71). Updating the cards regularly supports routine documentation practices and ensures that information gleaned throughout the project is not forgotten.

Another benefit of filling out data and model cards throughout a project is improved communication between engineers and other stakeholders. Engineers have expertise in areas such as wrangling data and developing models that other stakeholders may not, but engineers may not have expertise in communicating technical concepts to less-technical stakeholders. By completing the cards during a project, engineers can communicate their work in standardized, easy-to-read formats. Stakeholders, in turn, can use the cards to ask questions and set expectations for development.

Another lesson learned was that, at the end of a project, data and model cards can streamline project delivery. When completing a project involves transferring ownership of datasets or models between stakeholders, the stakeholders accepting ownership must assimilate information from the delivering team. Data and model cards consolidate and standardize project documentation, which can help both project teams economize their communications and facilitate the transfer of ownership. This can be particularly important for projects that involve contractors who may be difficult to contact after a project ends.

### *Future Challenges and Opportunities*

An ongoing challenge for AI trustworthiness is establishing a standardized framework for measuring it. There is limited consensus in AI research on metrics for trustworthiness or its characteristics; limited consensus on how to measure them; limited consensus on their quantitative relationships; and limited consensus on how to set appropriate target values or thresholds (13, 14, 16, 32, 44, 48, 57, 92–94). According to Benk et al., "the measurement of trust in AI and the relationship between different types of measures of it remain open research issues" (48). Federal guidance recognizes this challenge but provides few solutions. For example, the AI RMF states, "Today, the ability to understand and analyze the decisions of AI systems and measure their trustworthiness is limited" (14), but identifies risk measurement as a key function for risk management. GAO's AI Accountability Framework emphasizes the need for metrics, but leaves organizations to define them based on their own program objectives (13).

The absence of a standardized, quantitative framework for AI trustworthiness could cause several problems. First, agencies could develop trustworthiness metrics that do not measure trustworthiness. The AI RMF recognizes this problem, stating that "development of metrics is often an institutional endeavor and may inadvertently reflect factors unrelated to the underlying impact" (14). Second, agencies could develop metrics and target values that undermine AI trustworthiness. For example, an agency might develop metrics and target values for transparency and explainability that decrease the perception of AI trustworthiness. This is possible because transparency and explainability can support or hinder trustworthiness depending on context and degree (53, 92, 95, 96). Third, agencies might develop metrics and target values that are not easily comparable. As a result, assessing the degree to which different agencies or different groups within the same agency satisfy the same federal directives for trustworthiness would be difficult. If multiple agencies adopt the same metrics but different measurement methods, comparison would also be difficult.

In addition to the absence of a quantitative framework, another challenge is balancing tradeoffs in trustworthiness when choosing models. These tradeoffs exist because models exhibit different trustworthiness characteristics such as accuracy, transparency, and explainability to different degrees (see Figure 2 for a list of characteristics). As a result, selecting a model with appropriate trustworthiness requires trading off trustworthiness characteristics that are less important in the target application for those that are more important.

Trustworthiness tradeoffs can be illustrated with the choice between black-box and self-interpretable models. Choosing between these types of models often involves tradeoffs in trustworthiness characteristics such as accuracy and explainability. Black-box models, including the GNNs in our system, are often perceived to be more accurate than self-interpretable models (70)[10] but are less explainable. For example, if our GNN predicted that an exam of an enterprise's controlling owner would assess additional tax, IRS staff using the model would not be able to understand why by inspecting the model itself. In contrast, if the prediction were made by a self-interpretable model such as a logistic regression or decision tree, IRS staff could explain what enterprise features led to the prediction by comparing weights in the regression or tracing branches in the decision tree. While black-box models can be explained using post hoc methods, the methods can introduce other problems for trustworthiness such as providing inaccurate explanations that mislead users[11] (16, 33, 54, 97, 98). As a result, choosing a black-box model may prioritize accuracy as important for trustworthiness at the expense of explainability.

If agencies choose to use black-box models, an additional challenge may be enabling non-experts to use explainability methods. Explainability methods such as the ones surveyed and tested for our system can require technical expertise to use, so technical stakeholders such as engineers may need to operate them for non-technical stakeholders such as managers and aid in interpreting the results. Alternatively, agencies could procure or develop explainability software that is accessible to a variety of stakeholders.

In addition to supporting explainability, another challenge is promoting trustworthiness for an AI system composed of multiple models. For these multi-model systems, trustworthiness characteristics like transparency and explainability could be assessed for individual models, the whole system, or both. These approaches have different tradeoffs. For example, writing individual model cards facilitates model re-use in different systems but incurs higher documentation costs. It may also lead to inaccurate documentation if models are used on significantly different data sets than those on which they were trained and tested, but their cards do not report the differences. On the other hand, writing cards for a collection of models may require less documentation, but likely requires a new type of document that focuses on the system level, such as IBM's FactSheets (80).

Regardless of whether AI systems are documented at the model or system level, document version control is a challenge. Version control is the practice of tracking and managing changes to different versions of files. Tracking and managing changes will likely be critical for trustworthiness documents like data and model cards because they will likely be modified by many stakeholders (see Table 2) at many different times. One set of modifications will likely be made to card templates. Card templates like those published in this work (see Appendix) will probably be updated continuously by stakeholders such as leadership, management, and AI-impact assessors as the federal government issues new guidance on AI trustworthiness, and as an organization's internal policies evolve. Another set of modifications will likely be made to the cards' content for specific datasets and models as datasets grow or are modified, and as models are retrained on growing datasets or are otherwise updated. Version control can record and systematize the process of making these changes by tracking what changes were made, when they were made, and who made them. It can also provide mechanisms for changes to be reviewed and approved before being integrated into the documents.

While version control will likely be critical, several steps are required for adoption. One step is selecting version-control tools. Common options include software such as GitLab and Microsoft SharePoint (99), and tools embedded in word-processing software like Microsoft Word. These tools provide different features and interfaces, and organizations will need to choose the option that best meets their needs. This choice is complicated by the fact that features may be impacted by file types: for example, versioning information in Word's Track Changes feature is not saved in plain text files, and some GitLab features are available for plain text files but not Word files. A second step in adopting version control is educating users in how to use it. While users may be familiar with version-control features in Word such as Track Changes, more feature-rich systems like GitLab are more complicated to use and would likely require user training. A third step in adopting version control is ensuring it is used in compliance with organization policies. While software systems can support

[10] Some researchers argue that black-box models are not inherently more accurate than self-interpretable models: "It is a myth that there is necessarily a trade-off between accuracy and interpretability" (54).

[11] C. Rudin argues that post hoc explanations "must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation" (54).

good practices for version control, they do not ensure them. Organizations using version control would likely benefit from defining and encouraging good practices and confirming their use.

## Conclusion

Despite increasing guidance and directives from the federal government on trustworthy AI, there are few tools standardized across the government that are available to use in projects. To acquire tools, stakeholders will likely face challenges such as navigating ambiguous and inconsistent terminology in trustworthy AI; selecting and customizing tools to support diverse stakeholder goals; assessing the degree to which tools meet the goals; and managing the lifecycle of the tools.

The case study outlined in this work represents our attempt to meet some of these challenges. We hope it acts as a steppingstone for future efforts by providing a review of existing resources, an approach to aligning those resources with federal and project requirements, and a discussion of lessons learned and future challenges.

## References

1. J. Butler, Analytical Challenges in Modern Tax Administration: A Brief History of Analytics at the IRS Symposium on Artificial Intelligence & the Future of Tax Law: AI in Tax Compliance and Enforcement. *Ohio St. Tech. L. J.* **16**, 258–277 (2020).
2. "AI Guide for Government" (U.S. Government Services Administration, 2022); https://coe.gsa.gov/coe/ai-guide-for-government/print-all/index.html.
3. "Artificial Intelligence: Agencies Have Begun Implementation but Need to Complete Key Requirements" (GAO-24-105980, U.S. Government Accountability Office, 2023); https://www.gao.gov/products/gao-24-105980.
4. E. B. Johnson, *National Artificial Intelligence Initiative Act of 2020, Division E of the William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021* (2020; https://www.congress.gov/bill/116th-congress/house-bill/6216/text).
5. "IRS announces sweeping effort to restore fairness to tax system with Inflation Reduction Act funding; new compliance efforts focused on increasing scrutiny on high-income, partnerships, corporations and promoters abusing tax rules on the books" (News Release IR-2023-166, U.S. Internal Revenue Service, 2023); https://www.irs.gov/newsroom/irs-announces-sweeping-effort-to-restore-fairness-to-tax-system-with-inflation-reduction-act-funding-new-compliance-efforts.
6. "How the Internal Revenue Service Selects and Audits Individual Income Tax Returns" (U.S. Government Accountability Office, 1976); https://www.gao.gov/products/100316.
7. "Internal Revenue Manuals" (Internal Revenue Service); https://www.irs.gov/irm.
8. T. J. Beckman, "AI in the IRS" in *Proceedings of the Annual AI Systems in Government Conference* (1989; https://ieeexplore.ieee.org/abstract/document/47329), pp. 226–232.
9. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016).
10. "Internal Revenue Service Inflation Reduction Act Strategic Operating Plan, FY 2023-2031" (U.S. Internal Revenue Service, 2023); https://www.irs.gov/about-irs/irs-inflation-reduction-act-strategic-operating-plan.
11. "Treasury Strategic Plan 2022-2026" (2023); https://home.treasury.gov/about/budget-financial-reporting-planning-and-performance/strategic-plan.
12. "Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government" (Executive Order 13960, Executive Office of the President, 2020); https://www.federalregister.gov/d/2020-27065.
13. "Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities" (GAO-21-519SP, U.S. Government Accountability Office, 2021); https://www.gao.gov/products/gao-21-519sp.
14. "Artificial Intelligence Risk Management Framework" (NIST AI 100-1, National Institute of Standards and Technology, 2023); https://airc.nist.gov/AI_RMF_Knowledge_Base/AI_RMF.

15. "Blueprint for an AI Bill of Rights" (The White House, 2022); https://www.whitehouse.gov/ostp/ai-bill-of-rights/.
16. P. J. Phillips, C. A. Hahn, P. C. Fontana, A. N. Yates, K. Greene, D. A. Broniatowski, M. A. Przybocki, "Four Principles of Explainable Artificial Intelligence" (NISTIR 8312, National Institute of Standards and Technology, 2021); https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf.
17. R. Schwartz, A. Vassilev, K. Greene, L. Perine, A. Burt, P. Hall, "Proposal for Identifying and Managing Bias in Artificial Intelligence" (SP 1270, National Institute of Standards and Technology, 2022); https://www.nist.gov/artificial-intelligence/proposal-identifying-and-managing-bias-artificial-intelligence-sp-1270.
18. R. T. Vought, "Guidance for Regulation of Artificial Intelligence Applications" (Memorandum For The Heads Of Executive Departments And Agencies M-21–06, Executive Office of the President, Office of Management and Budget, 2020).
19. S. Young, "Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence" (Proposed Memorandum for the Heads of Executive Departments and Agencies OMB-2023-0020-0001, Executive Office of the President, Office of Management and Budget, Washington, D.C. 20503, 2023); https://www.regulations.gov/document/OMB-2023-0020-0001.
20. K. S. Chmielinski, S. Newman, M. Taylor, J. Joseph, K. Thomas, J. Yurkofsky, Y. C. Qiu, The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. arXiv arXiv:2201.03954 [Preprint] (2022). https://doi.org/10.48550/arXiv.2201.03954.
21. Sample Drug Facts Label, *US Food and Drug Administration* (2016). https://www.fda.gov/drugs/special-features/sample-drug-facts-label.
22. Safety Data Sheets (2012). https://www.osha.gov/laws-regs/regulations/standardnumber/1910/1910.1200AppD.
23. "Partnerships and S Corporations: IRS Needs to Improve Information to Address Tax Noncompliance" (Report to the Chairman, Committee on Finance, U.S. Senate GAO-14-453, U.S. Government Accountability Office, 2014); https://www.gao.gov/products/gao-14-453.
24. "Tax Gap: IRS Can Improve Efforts to Address Tax Evasion by Networks of Businesses and Related Entities" (Report to the Committee on Finance, U.S. Senate GAO-10-968, U.S. Government Accountability Office, 2010); https://www.gao.gov/products/gao-10-968.
25. Internal Revenue Service, "Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2014–2016" (Publication 1415 (Rev. 10-2022), Department of the Treasury, Washington, D.C., 2022).
26. J. Guyton, P. Langetieg, D. Reck, M. Risch, G. Zucman, "Tax Evasion at the Top of the Income Distribution: Theory and Evidence | NBER" (Working Paper 28542, National Bureau of Economic Research, 2021); https://www.nber.org/papers/w28542.
27. M. M. Bronstein, J. Bruna, T. Cohen, P. Veličković, Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv:2104.13478 [cs, stat]* (2021).
28. W. L. Hamilton, *Graph Representation Learning* (Morgan & Claypool Publishers, 2020).
29. L. Wu, P. Cui, J. Pei, L. Zhao, *Graph Neural Networks: Foundations, Frontiers, and Applications* (Springer Nature, 2022).
30. J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications. *AI Open* **1**, 57–81 (2020).
31. C. Rudin, J. Radin, Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review* **1**, 10–1162 (2019).
32. B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy AI: From Principles to Practices. *ACM Comput. Surv.* **55**, 177:1-177:46 (2023).
33. A. Adadi, M. Berrada, Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).
34. D. Castelvecchi, Can we open the black box of AI? *Nature News* **538**, 20 (2016).

35. R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **51**, 93:1-93:42 (2018).
36. B. L. Garrett, C. Rudin, The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice. 4275661 [Preprint] (2023). https://doi.org/10.2139/ssrn.4275661.
37. O. Loyola-González, Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View. *IEEE Access* **7**, 154096–154113 (2019).
38. R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, R. Ranjan, Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* **55**, 194:1-194:33 (2023).
39. "NASA Systems Engineering Handbook" (NASA SP-2016-6105 Rev2, National Aeronautics and Space Administration, 2019); https://www.nasa.gov/reference/systems-engineering-handbook/.
40. A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58**, 82–115 (2020).
41. D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy Artificial Intelligence: A Review. *ACM Comput. Surv.* **55**, 39:1-39:38 (2022).
42. M. Micheli, I. Hupont, B. Delipetrev, J. Soler-Garrido, The landscape of data and AI documentation approaches in the European policy context. *Ethics Inf Technol* **25**, 56 (2023).
43. T. Hagendorff, The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines* **30**, 99–120 (2020).
44. S. Thiebes, S. Lins, A. Sunyaev, Trustworthy artificial intelligence. *Electron Markets* **31**, 447–464 (2021).
45. "Recommendation of the Council on Artificial Intelligence" (OECD/LEGAL/0449, Organisation for Economic Co-operation and Development, 2023); https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
46. A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines. *Nat Mach Intell* **1**, 389–399 (2019).
47. "Ethics Guidelines for Trustworthy Artificial Intelligence" (High-Level Expert Group on AI, 2019); https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
48. M. Benk, S. Tolmeijer, F. von Wangenheim, A. Ferrario, The Value of Measuring Trust in AI - A Socio-Technical System Perspective. arXiv [Preprint] (2022). https://arxiv.org/abs/2204.13480v1.
49. B. Stanton, T. Jensen, "Trust and Artificial Intelligence" (Draft NISTIR 8332, National Institute of Standards and Technology, 2020).
50. "U.S. Leadership in AI: A Plan for Federal Engagement in Developing Technical Standards and Related Tools" (National Institute of Standards and Technology, 2019); https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.
51. "Overview of trustworthiness in artificial intelligence" (International Standard (ISO/IEC) TR 24028:2020, International Organization for Standardization, 2020); https://www.iso.org/obp/ui/en/#iso:std:iso-iec:tr:24028:ed-1:v1:en.
52. J. M. Wing, Trustworthy AI. *Commun. ACM* **64**, 64–71 (2021).
53. A. Preece, D. Harborne, D. Braines, R. Tomsett, S. Chakraborty, Stakeholders in Explainable AI. arXiv arXiv:1810.00184 [Preprint] (2018). https://doi.org/10.48550/arXiv.1810.00184.
54. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).
55. "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence" (Executive Order 14110, Executive Office of the President, 2023); https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.
56. A. Rawal, J. McCoy, D. B. Rawat, B. M. Sadler, R. St. Amant, Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives. *IEEE Transactions on Artificial Intelligence* **3**, 852–866 (2022).

57. L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* **106**, 102301 (2024).
58. R. Schwartz, A. Vassilev, K. K. Greene, L. Perine, A. Burt, P. Hall, "Towards a Standard for Identifying and Managing Bias in Artificial Intelligence" (Special Publication 1270, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2022); https://www.nist.gov/publications/towards-standard-identifying-and-managing-bias-artificial-intelligence.
59. "Artificial Intelligence: Emerging Opportunities, Challenges, and Implications" (GAO-18-142SP, U. S. Government Accountability Office, 2018); https://www.gao.gov/products/gao-18-142sp.
60. S. M. Burwell, S. VanRoekel, T. Park, D. J. Mancini, "Open Data Policy-Managing Information as an Asset" (Memorandum For The Heads Of Executive Departments And Agencies M-13–13, Executive Office of the President, Office of Management and Budget, 2013).
61. "Maintaining American Leadership in Artificial Intelligence" (Executive Order 13859, Executive Office of the President, 2019); https://www.federalregister.gov/documents/2019/02/14/2019-02544/maintaining-american-leadership-in-artificial-intelligence.
62. "Federal Data Strategy 2021 Action Plan" (Office of Management and Budget, Office of Science and Technology Policy, Department of Commerce, and Small Business Administration, 2021); https://strategy.data.gov/.
63. "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource" (National Artificial Intelligence Research Resource Task Force, 2023); https://www.whitehouse.gov/ostp/news-updates/2023/01/24/strengthening-and-democratizing-the-u-s-artificial-intelligence-innovation-ecosystem/.
64. "National Artificial Intelligence Advisory Committee (NAIAC) Year 1 Report" (National Artificial Intelligence Advisory Committee, 2023); https://www.ai.gov/wp-content/uploads/2023/05/NAIAC-Report-Year1.pdf.
65. "Data, Analytics, and Artificial Intelligence Adoption Strategy Accelerating Decision Advantage" (Department of Defense, 2023).
66. "Ethical Principles for Artificial Intelligence" (Department of Defense, 2020); https://www.ai.mil/docs/Ethical_Principles_for_Artificial_Intelligence.pdf.
67. J. McNerney, *AI in Government Act of 2020* (2020; https://www.congress.gov/bill/116th-congress/house-bill/2575).
68. H. Yuan, H. Yu, S. Gui, S. Ji, Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**, 5782–5799 (2022).
69. A. Das, P. Rad, Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. arXiv arXiv:2006.11371 [Preprint] (2020). https://doi.org/10.48550/arXiv.2006.11371.
70. R. Caruana, S. Lundberg, M. T. Ribeiro, H. Nori, S. Jenkins, "Intelligible and Explainable Machine Learning: Best Practices and Practical Challenges" in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, New York, NY, USA, 2020; https://dl.acm.org/doi/10.1145/3394486.3406707)*KDD '20*, pp. 3511–3512.
71. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, K. Crawford, Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).
72. O. Papakyriakopoulos, A. S. G. Choi, W. Thong, D. Zhao, J. Andrews, R. Bourke, A. Xiang, A. Koenecke, "Augmented Datasheets for Speech Datasets and Ethical Decision-Making" in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York, NY, USA, 2023; https://dl.acm.org/doi/10.1145/3593013.3594049)*FAccT '23*, pp. 881–904.
73. M. Pushkarna, A. Zaldivar, O. Kjartansson, "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI" in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York, NY, USA, 2022; https://dl.acm.org/doi/10.1145/3531146.3533231)*FAccT '22*, pp. 1776–1826.

74. Dataset Cards, *Hugging Face* (2024). https://huggingface.co/docs/hub/en/datasets-cards.
75. A. Fabris, S. Messina, G. Silvello, G. A. Susto, "Tackling Documentation Debt: A Survey on Algorithmic Fairness Datasets" in *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Association for Computing Machinery, New York, NY, USA, 2022; https://dl.acm.org/doi/10.1145/3551624.3555286) *EAAMO '22*, pp. 1–13.
76. E. M. Bender, B. Friedman, Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* **6**, 587–604 (2018).
77. D. K. Chen, Y. Modi, L. A. Al-Aswad, Promoting Transparency and Standardization in Ophthalmologic Artificial Intelligence: A Call for Artificial Intelligence Model Card. *The Asia-Pacific Journal of Ophthalmology* **11**, 215 (2022).
78. M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, "Model Cards for Model Reporting" in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019; http://arxiv.org/abs/1810.03993), pp. 220–229.
79. Model Cards, *Hugging Face* (2024). https://huggingface.co/docs/hub/en/model-cards.
80. M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K. R. Varshney, FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* **63**, 6:1-6:13 (2019).
81. A. Crisan, M. Drouhard, J. Vig, N. Rajani, "Interactive Model Cards: A Human-Centered Approach to Model Documentation" in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery, New York, NY, USA, 2022; https://dl.acm.org/doi/10.1145/3531146.3533108)*FAccT '22*, pp. 427–439.
82. "OECD Framework for the Classification of AI systems" (OECD Digital Economy Papers No. 323, Organisation for Economic Co-operation and Development, Paris, 2022); https://doi.org/10.1787/cb6d9eca-en.
83. F. A. Batarseh, L. Freeman, C.-H. Huang, A survey on artificial intelligence assurance. *Journal of Big Data* **8**, 60 (2021).
84. W. Liang, N. Rajani, X. Yang, E. Ozoani, E. Wu, Y. Chen, D. S. Smith, J. Zou, What's documented in AI? Systematic Analysis of 32K AI Model Cards. arXiv:2402.05160 [Preprint] (2024). https://arxiv.org/abs/2402.05160v1.
85. X. Yang, W. Liang, J. Zou, Navigating Dataset Documentations in AI: A Large-Scale Analysis of Dataset Cards on Hugging Face. arXiv arXiv:2401.13822 [Preprint] (2024). https://doi.org/10.48550/arXiv.2401.13822.
86. "AI Model Cards" (Interagency AI Community of Practice: Responsible AI Working Group, 2022).
87. "Data Docs for AI" (Interagency AI Community of Practice: Responsible AI Working Group, 2022).
88. M. Sundararajan, A. Taly, Q. Yan, "Axiomatic Attribution for Deep Networks" in *Proceedings of the 34th International Conference on Machine Learning* (PMLR, 2017; https://proceedings.mlr.press/v70/sundararajan17a.html), pp. 3319–3328.
89. D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, X. Zhang, "Parameterized Explainer for Graph Neural Network" in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2020; https://proceedings.neurips.cc/paper/2020/hash/e37b08dd3015330dcbb5d6663667b8b8-Abstract.html)vol. 33, pp. 19620–19631.
90. H. Yuan, H. Yu, J. Wang, K. Li, S. Ji, "On Explainability of Graph Neural Networks via Subgraph Explorations" in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021; https://proceedings.mlr.press/v139/yuan21c.html), pp. 12241–12252.
91. R. Ying, D. Bourgeois, J. You, M. Zitnik, J. Leskovec, GNNExplainer: Generating Explanations for Graph Neural Networks. *arXiv:1903.03894 [cs, stat]* (2019).
92. C. Seifert, S. Scherzinger, L. Wiese, "Towards Generating Consumer Labels for Machine Learning Models" in *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)* (2019; https://ieeexplore.ieee.org/abstract/document/8998974), pp. 173–179.
93. K. Haresamudram, S. Larsson, F. Heintz, Three Levels of AI Transparency. *Computer* **56**, 93–100 (2023).

94. H. Felzmann, E. Fosch-Villaronga, C. Lutz, A. Tamò-Larrieux, Towards Transparency by Design for Artificial Intelligence. *Sci. Eng. Ethics* **26**, 3333–3361 (2020).

95. L. Kästner, M. Langer, V. Lazar, A. Schomäcker, T. Speith, S. Sterz, "On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness" in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)* (2021; https://ieeexplore.ieee.org/abstract/document/9582305), pp. 169–175.

96. L. R. Marusich, J. Z. Bakdash, E. Onal, M. S. Yu, J. Schaffer, J. O'Donovan, T. Höllerer, N. Buchler, C. Gonzalez, Effects of Information Availability on Command-and-Control Decision Making: Performance, Trust, and Situation Awareness. *Hum. Factors* **58**, 301–321 (2016).

97. I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, N. C. Bouaynaya, Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Processing Magazine* **39**, 73–84 (2022).

98. E. Dai, S. Wang, "Towards Self-Explainable Graph Neural Network" in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Association for Computing Machinery, New York, NY, USA, 2021; https://dl.acm.org/doi/10.1145/3459637.3482306)*CIKM '21*, pp. 302–311.

99. Versioning in SharePoint, *Microsoft 365* (2024). https://learn.microsoft.com/en-us/microsoft-365/community/versioning-basics-best-practices.

100. S. Donovan, "Preparing for and Responding to a Breach of Personally Identifiable Information" (Memorandum for Heads of Executive Departments and Agencies M-17–12, Executive Office of the President, Office of Management and Budget, Washington, D.C. 20503, 2017); https://osec.doc.gov/opog/privacy/Memorandums/OMB_M-17-12.pdf.

101. K. R. Boeckl, N. B. Lefkovitz, "NIST Privacy Framework: A Tool for Improving Privacy Through Enterprise Risk Management, Version 1.0" (CSWP 01162020, National Institute of Standards and Technology, 2020); https://www.nist.gov/publications/nist-privacy-framework-tool-improving-privacy-through-enterprise-risk-management.

# Appendix

| | |
|---|---|
| **Data card authors(s)** | [name(s)] |
| **Contact info.** | [email address, phone number, etc.] |
| **Data card release date** | [year-month-day] |
| **Data card version** | [VX.X] |

## Data Card

Write the dataset name here (and any aliases or acronyms)

### 1. Dataset Identification

1.1. **Summary.** Briefly describe the dataset in plain language.

1.2. **Creator.** Who created the dataset (i.e., what research group, agency, or division)? Provide contact information if available (e.g., name, affiliation, email address, website).

1.3. **Points of Contact (POCs).** Identify POCs who can answer questions about the dataset or provide assistance (list names, email addresses, departments, etc.).

1.4. **Release date.** When was the dataset made available?
YYYY-MM-DD

1.5. **Size.** What is the size of the dataset in bytes?

1.6. **Version.** Provide the version number of the dataset or other identifying information.

1.7. **Format.** Describe the format of the dataset.
Example: The dataset consists of 3 CSV files and 1 JSON file.

1.8. **Sensitivity.**

| 5.1 Does the dataset contain sensitive data[12]? | 5.2 Does this data card contain sensitive data? |
|---|---|
| ☐ No | ☐ No |
| ☐ Yes | ☐ Yes |

1.9. **Personally identifiable information (PII).**

| 5.3 Does the dataset contain PII[13]? | 5.4 Does this data card contain PII? |
|---|---|
| ☐ No | ☐ No |
| ☐ Yes | ☐ Yes |

### 2. Motivation

2.1. **Initiator.** What entity (e.g., division, team, agency, or external party) ordered the creation of the dataset?

2.2. **Purpose.** Why was the dataset created? What were the intended uses? If the dataset was modified or aggregated from other datasets, include their intended uses.

[12] Sensitive data includes data about people such as "disability-related data, genomic data, biometric data, behavioral data, geolocation data, data related to interaction with the criminal justice system, relationship history and legal status such as custody and divorce information, and home, work, or school environmental data"; and "data that "have the reasonable potential to be used in ways that are likely to expose individuals to meaningful harm, such as a loss of privacy or financial harm due to identity theft. Data and metadata generated by or about those who are not yet legal adults is also sensitive data" (*15*).

[13] "PII refers to information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual" (*100*).

| | |
|---|---|
| **Data card authors(s)** | [name(s)] |
| **Contact info.** | [email address, phone number, etc.] |
| **Data card release date** | [year-month-day] |
| **Data card version** | [VX.X] |

2.3. **Users.** Who were the intended users of the dataset (i.e., what person, research group, agency, or division)?

## 3. Composition

3.1. **Instance counts.** How many instances (e.g., cases, objects, observations) are in the dataset? List the count for each type of instance.

3.2. **Instance descriptions.** What do the instances in the dataset represent?

3.3. **Metadata.** Is there metadata[14]?
- ☐ No
- ☐ Yes. Provide a link or location and describe how to obtain access permission if applicable.

3.4. **Completeness.** Is the dataset a sample from another dataset?
- ☐ No
- ☐ Yes. Describe the parent dataset, why the sampling was performed, and whether the dataset is representative of the parent dataset.

3.5. **Missingness**. Are any data missing (e.g., null or NA)?
- ☐ No
- ☐ Yes. Describe the missing data and explain why it's missing.

3.6. **Data Quality.** Is there an erratum for the dataset?
- ☐ No
- ☐ Yes. Provide a link or its location.

Are there known data quality issues not described in an erratum (e.g., errors, noise, redundancies, or imbalances)?
- ☐ No
- ☐ Yes. Describe them.

3.7. **Sensitivity.** If the dataset contains sensitive information (including PII), describe it.

## 4. Collection

4.1. **Acquisition.** How was the dataset acquired?
Example: The 3 CSV files were downloaded from www.data.com. The JSON file was acquired from John Smith in the Research, Applied Analytics and Statistics Division of the Internal Revenue Service.

4.2. **Method.** What process was used to collect the data (e.g., web scraping, surveying, etc.)?

[14] Metadata is "Information describing the characteristics of data. This may include, for example, structural metadata describing data structures (i.e., data format, syntax, semantics) and descriptive metadata describing data contents" (*101*).

| | |
|---|---|
| **Data card authors(s)** | [name(s)] |
| **Contact info.** | [email address, phone number, etc.] |
| **Data card release date** | [year-month-day] |
| **Data card version** | [VX.X] |

4.3. **Data Sources.** If the dataset was derived from other datasets, list the parent datasets, provide links or locations, and provide links or locations to their data cards if possible. Describe the derivation process (e.g., sampling or modifications of parent datasets).

## 5. Processing

5.1. **Method.** Has any data been processed (e.g., cleaned, labeled, imputed, deleted)?

☐ No
☐ Yes. Complete the following:

***Raw Data.*** Is the raw dataset available?

☐ No
☐ Yes. Provide a link or its location. If access permission is required, describe how to obtain it.

***Software.*** Is the processing code available?

☐ No. Describe the processing.
☐ Yes. Provide a link or its location. If access permission is required, describe how to obtain it.

## 6. Distribution

6.1. **Location.** Where is the dataset located? Provide a link if possible and explain how to obtain access permission if applicable.

6.2. **Distribution restrictions.** Are there restrictions on distributing the dataset (e.g., due to export control, regulations, intellectual property, terms of use, etc.)?

6.3. **Documentation.** Provide links or locations for any additional dataset documentation (e.g., websites or articles related to the dataset's creation or intended use).

## 7. Maintenance

7.1. **Status.** Is the dataset actively maintained at the time of this data card's creation?

☐ No
☐ Yes. Provide contact information for the maintainers.

7.2. **Updates.** Are there plans to update the dataset?

☐ No
☐ Yes. Describe planned updates, the updating cadence, and whether older versions of the dataset will be supported.

| | |
|---|---|
| **Model card authors(s)** | [name(s)] |
| **Contact info.** | [email address, phone number, etc.] |
| **Model card release date** | [year-month-day] |
| **Model card version** | [VX.X] |

# Model Card

Write the model name here (and any aliases or acronyms)

## 1. Model Identification

1.1. **Model type.** What is the model type (e.g., linear regression, convolutional neural network, etc.)?

1.2. **Task.** What is the model task (e.g., regression, classification, anomaly detection, etc.)?

1.3. **Creator.** Who created the model (i.e., what research group, agency, or division)? Provide contact information if available (e.g., name, affiliation, email address, website).

1.4. **Points of Contact (POCs).** Identify POCs who can answer questions about the model or provide assistance (list names, email addresses, departments, etc.).

1.5. **Creation date.** When was the model created?
YYYY-MM-DD

1.6. **Version.** Provide the version number of the model or other identifying information.

Is there a commit ID for the model in a version-control system?
☐ No
☐ Yes. Provide the ID.

## 2. Motivation

2.1. **Initiator.** What entity (e.g., division, team, agency, or external party) ordered the creation of the model?

2.2. **Purpose.** Why was the model created? What were the intended uses?

2.3. **Users.** Who were the intended users of the model (i.e., what person, research group, agency, or division)?

## 3. Data

3.1. **Source.** Is there a data card for the dataset used to develop the model?
☐ No. Create one if possible.
☐ Yes. Provide a link or location and describe how to obtain access permission if applicable.

3.2. **Inputs summary.** Summarize the input features to the model in plain language.
Example: The inputs are demographic data about a person's health.

3.3. **Output summary.** Summarize the model outputs in plain language.
Example: The model outputs a classification score for the likelihood a person has diabetes.

| | |
|---|---|
| **Model card authors(s)** | [name(s)] |
| **Contact info.** | [email address, phone number, etc.] |
| **Model card release date** | [year-month-day] |
| **Model card version** | [VX.X] |

3.4. **Input details.** Provide a link or location for the input features and describe how to obtain access permission if applicable.

Is there metadata describing the features?
☐ No. Describe the features.

☐ Yes. Provide a link or location and describe how to obtain access permission if applicable.

3.5. **Output details.** Provide the possible values of the model outputs and their meanings.
Example for categorical values: 0 means no disease; 1 means disease
Example for continuous values: Outputs are real numbers greater than or equal to zero. They represent a person's predicted income.

3.6. **Training, validation, and test sets.** Describe how the dataset was split into training, validation, and test sets.

## 4. Risks of use

4.1. **Risks.** Are there known potentially negative impacts of using the model, such as adverse effects on the civil rights, civil liberties, or privacy of individuals or groups?
☐ No.
☐ Yes. Describe potential impacts. Identify the groups of people who may be impacted and, if feasible, estimate the likelihood and magnitude of the impacts.

4.2. **Mitigation strategy.** Describe how the risks can be mitigated.

## 5. Performance

5.1. **Results.** For each performance metric[15], list the metric and value obtained for the training, validation, and test sets.

5.2. **Metrics rationale.** Describe why the performance metrics were chosen.

5.3. **Decision thresholds.** Were any decision thresholds used (e.g., for logistic regression, outputs greater than .7 are classified as having disease)?
☐ No
☐ Yes. Explain how the threshold was chosen.

5.4. **Baselines.** Are there baselines or other reference points against which the model can be compared?
☐ No
☐ Yes. Describe them and include their values if available.

[15] Metrics quantitatively measure the degree to which the model performance is "consistent with the program goals and objectives" (*13*). The metrics "should extend beyond assessing for accuracy, safety, and security and include bias, equity, and other societal considerations" (*13*).

| | |
|---|---|
| **Model card authors(s)** | [name(s)] |
| **Contact info.** | [email address, phone number, etc.] |
| **Model card release date** | [year-month-day] |
| **Model card version** | [VX.X] |

## 6. Explainability

6.1. **Audience.** List the groups of people who may require knowing how and why the model produces particular results (e.g., analysts, managers, external parties, etc.).

6.2. **Model opacity**. Is the model self-interpretable[16]?

☐ No. List tools used for post-hoc explanations and describe their functions[17]. Identify if the tools produce local explanations, global explanations, or both[18].

☐ Yes. Explain in plain language how the model produces outcomes.

6.3. **Explanation suitability.** Describe how the chosen explainability methods or tools meet the needs of the audiences.

## 7. Limitations and recommendations

7.1. **Limitations.** Are there known limitations of the model such as out-of-scope use cases?

☐ No
☐ Yes. Describe them.

7.2. **Caveats.** Is there information not recorded elsewhere in this documentation that may be useful to model users?

☐ No
☐ Yes. Explain.

7.3. **Recommendations.** Are there known methods to improve the development or use of the model?

☐ No
☐ Yes. Describe them.

## 8. References

8.1. List references to relevant documents, repositories, or other resources that may be useful to understand the model. Explain why the reference is useful if appropriate.

---

[16] Self-interpretable models "are models that are themselves the explanations. Not only do they explain the entire model globally, but by walking through each input through the model, the simulation of the input on the self-interpretable model can provide a local explanation for each decision. Some of the most common self-interpretable models include decision trees and regression models (including logistic regression)" (*16*).

[17] "Post-hoc explanations are explanations, often generated by other software tools, that describe, explain, or model the algorithm to give an idea of how the algorithm works. Post-hoc explanations often can be used on algorithms without any inner knowledge of how the algorithm works, provided that it can be queried for outputs on chosen inputs" (*16*).

[18] "A local explanation explains a subset of decisions or is a per-decision explanation. A global explanation produces a model that approximates the non-interpretable model" (*16*).

# 4



## Simplifying the Filing Burden

**Gogani-Khiabani ◆ Dewangan ◆ Trivedi**
**Olson ◆ Tizpaz-Niari**

**Thomas**

**Chen ◆ Cornwall ◆ Herlache ◆ Leary**
**Schafer ◆ Vigil ◆ Javaid**

# Technical Challenges in Maintaining Tax Prep Software with Large Language Models

*Sina Gogani-Khiabani, Saeid Tizpaz-Niari[1] (University of Texas, El Paso),*
*Varsha Dewangan, Ashutosh Trivedi (University of Colorado, Boulder),*
*and Nina Olson (Center for Taxpayer Rights)*

## 1. Introduction

The growing complexity of U.S. income tax laws has made manual tax return preparation burdensome and susceptible to errors. According to the IRS, 90% of tax filers submitted their taxes electronically in 2020 (IRS (2020c)). Additionally, the use of software for tax preparation is on the rise, with more than 72 million individuals preparing their taxes independently, without tax professionals, marking a 24% increase from 2019 (Internal Revenue Service (2020a)). As a result, the industry revenue for tax preparation services has grown to an estimated $13.9 billion in 2023 (IBIS World (2023)). Recently, the IRS introduced Direct File, an online software tool that provides free online tax filing in 12 states (Internal Revenue Service (2024)). The development of socio-legal critical software is known to be challenging (Escher and Banovic (2020)), as it requires combined expertise in mission-critical software development practices and legal framework interpretation. The ever-changing nature of tax regulations further aggravates this challenge due to the need to keep tax software artifacts accurate and up to date. These constant changes require continuous revisions and updates to ensure compliance and functionality. Consequently, the current state-of-the-art method remains time-consuming and susceptible to errors. A National Science Foundation (NSF) program on Designing Accountable Software Systems (DASS) provides support to the authors in developing principled software engineering tools to improve the accountability of tax preparation software.

In this paper, we discuss key technical challenges in maintaining tax preparation software by leveraging recent advances in Large Language Models (LLMs).

We posit that the precise and formal language used in tax amendments, as outlined in IRS publications, is amenable to automatic translation into executable software code via LLMs. In addition to natural language processing, LLMs have demonstrated significant potential in generating code (Hindle et al. (2016); Fan et al. (2023); Li et al. (2023)), thanks to the naturalness of software and the availability of extensive training datasets from software repositories. Our work explores the opportunities and challenges of leveraging LLMs to maintain tax preparation software as it responds to changes in tax laws.

### *Testing and Debugging of Tax Prep Software*

The authors in their prior works (Tizpaz-Niari et al. (2023); Srinivas et al. (2023)) focused on the trustworthiness of tax prep software. One important obstacle to validate the correctness of tax prep software against the tax code, as outlined in the various publications by the IRS, is the oracle problem (Barr et al. (2015)): the class of correctness requirements for tax preparation systems are not explicitly available since the correct tax-filing is highly subjective to individual taxpayers. Given the relevant information about an individual, resolving the correct decision for that individual requires accounting and legal expertise. The authors made a critical observation connecting the principle of common law and stare decisis to the metamorphic specifications: the correctness of tax preparation software must also be viewed in comparison with similarly situated taxpayers. One key contribution of these prior works is to explicate formal representations of these properties from the latest Internal Revenue Service (IRS) documents (Srinivas et al. (2023)). In addition, we presented a framework, called TenForty (Tizpaz-Niari et al. (2023)) that automatically generates test cases from these metamorphic specifications to ensure the trustworthiness of tax prep software.

1 This project has been supported by the NSF under grants CCF-2317206 and CCF-2317207. Corresponding Author: saeid@utep.edu.

### *LLMs for Maintainability*

Following the new tax legislation and the IRS publications of new regulations, the tax prep software needs to be updated to reflect the changes in the software artifacts. However, as the tax law has evolved over different years, updating the corresponding software manually is error-prone and tedious. We study the following research question:

> *Can we leverage recent breakthroughs in pre-trained LLMs to assist software developers in automatically updating the implementation of tax law in software artifacts?*

LLMs produce a probability distribution over their outputs and thus, can generate several candidate solutions with potentially widely differing characteristics. Our ability to rank solutions based on their fitness is a key challenge in employing LLMs for correct software implementations of the tax code. Equipped with a reliable ranking mechanism, one can invoke LLMs in what is known as chain-of-thought reasoning (Wei et al. (2022)) to iteratively improve a candidate solution. In this paper, we focus on the problem of ranking candidate solutions generated by the LLMs.

### *Experimental Setup and Results*

For our experiments, we focus on generating functions to compute three key tax calculations: 1) tax brackets, 2) tax deductions, and 3) Earned Income Tax Credits (EITC) through LLMs for Tax Year 2021. We use two variants of OpenAI's LLM, ChatGPT (i.e., GPT-4.0 and GPT-3.5), and prompt them with descriptions from the tax publications under two distinct scenarios: 1) with the reference implementation from Tax Year 2020, and 2) without any reference implementations (direct prompting). In response to these prompts, the LLMs generated 10 candidate code implementations. We first use well-established ranking metrics such as CodeBERTScore (Zhou et al. (2023)) and compare their ranking outcomes to the ground truth implementations. We found that the existing metrics often fail to rank the candidate implementations in a way that the highest-ranked candidates have the lowest errors compared to the ground truth implementations. Then, we introduce a new metric, MajorityVote, where we take the majority votes of candidates in ranking them (i.e., an implementation that agrees the most with other candidates is considered a high-rank candidate). Our experiments show that a combination of CodeBERTScore and MajorityVote outperformed each metric in isolation.

Our results show that when the LLMs are prompted without the reference code of the prior year, the top ranked candidates, generated by GPT-3.5, achieved accuracy between 0 to 2% whereas those by GPT-4.0 achieved accuracy between 43 and 100%. However, when prompted with the implementation of tax prep software from the prior year (given as the context to the LLMs), GPT-3.5 and GPT-4.0 achieved accuracy between 21 and 100% and 48 to 100%, respectively. Rather than considering the absolute accuracy, we also study the accuracy of our ranking methods with some threshold where a solution within $\delta$-percent of ground truth is considered correct. We observe that without the prior year code context, GPT-3.5's candidates are considerably off and barely achieved 10% accuracy with a $\delta$ of 10. GPT-4.0's candidates, however, when prompted without the contexts from the prior years, achieved 100% accuracy in all cases when the $\delta$ is at least 7%. Interestingly and somehow surprisingly, when prompted with the code context, GPT-3.5's candidates achieved a similar or better accuracy, compared to GPT-4.0's candidates, when $\delta$ sets to at least 4%.

We share our best practices on prompting LLMs to generate implementations of a tax code, giving a code context to the LLMs, and providing the implementations from the prior code. While our current research focuses on a robust ranking system for identifying the most promising candidates from the LLM-generated tax prep software code, the next phase will focus on validation and refinement of (top-ranked) candidates to understand the extent to which the LLMs can be used to update tax prep software and maintain them automatically. For validation, we plan to integrate the ranking system with the metamorphic specifications and testing framework to ensure the correctness of updated code. If the candidate failed over the correctness requirements, then the Feedback Prompt Generator (FPG) will analyze the specific errors and create targeted prompts to guide the LLM in generating more accurate code in the next iteration.

## 2. Maintainability Challenges in Tax Prep Software

In this section, we briefly review prior work (Tizpaz-Niari et al. (2023)) that leverages metamorphic relations to test the functional correctness of tax preparation software. Using this approach, we have demonstrated how an open-source tax

preparation software project failed to correctly update the code to account for new tax legislation. This underscores the importance of automatic methodologies to update tax preparation software. The key research question in approaching the trustworthiness of tax prep software is the following:

> *How can one ensure that the tax prep software faithfully implements the tax law code as outlined by various IRS publications such as Form 1040, Publication 596 (EITC), Schedule 8812 (Qualifying Dependents), and Form 8863 (Education Credits)?*

### *Challenges*

Due to the lack of explicit correctness requirements, one can recourse to a pre-existing dataset to test and debug the software. Unfortunately, it is hard to obtain a meaningful labeled dataset—individuals and their "optimal" tax returns—due to obvious privacy and legal concerns. Even when one can learn a good generative model (Mehta et al. (2022)) to produce a synthetic population, tax software suffers from what is known as the oracle problem (Barr et al. (2015)) in software engineering: determining the correct output of an individual decision is time-consuming, expensive, and error-prone due to its highly subjective nature (as discussed next). A key observation made in this preliminary work is that several compliance specifications can be expressed relating an individual with a counterfactual one. We proposed a formal (first-order) logic (metamorphic relation) to express such compliance properties.

### *Metamorphic Relations*

We characterized 33 metamorphic specifications (Srinivas et al. (2023)) from 5 domains of the U.S. Individual Income Tax Return: (1) Credit for the Elderly or the Disabled (Internal Revenue Service (2021b)), a credit for taxpayers who are aged 65 or older or who are retired on permanent and total disability; (2) Earned Income Tax Credit (EITC) (Internal Revenue Service (2021c)), a refundable tax credit for lower-income households; (3) Child Tax Credit (CTC), a nonrefundable credit to reduce the taxes owed based on the number of qualifying children under the age of 17 (Internal Revenue Service (2021e)); (4) Educational Tax Credit (ETC) that help students with the cost of higher education by lowering their owed taxes or increasing their refund (Internal Revenue Service (2021a)); and (5) Itemized Deduction (ID) that is an option for taxpayers with significant tax deductible expenses (Internal Revenue Service (2021d)). Some examples are:

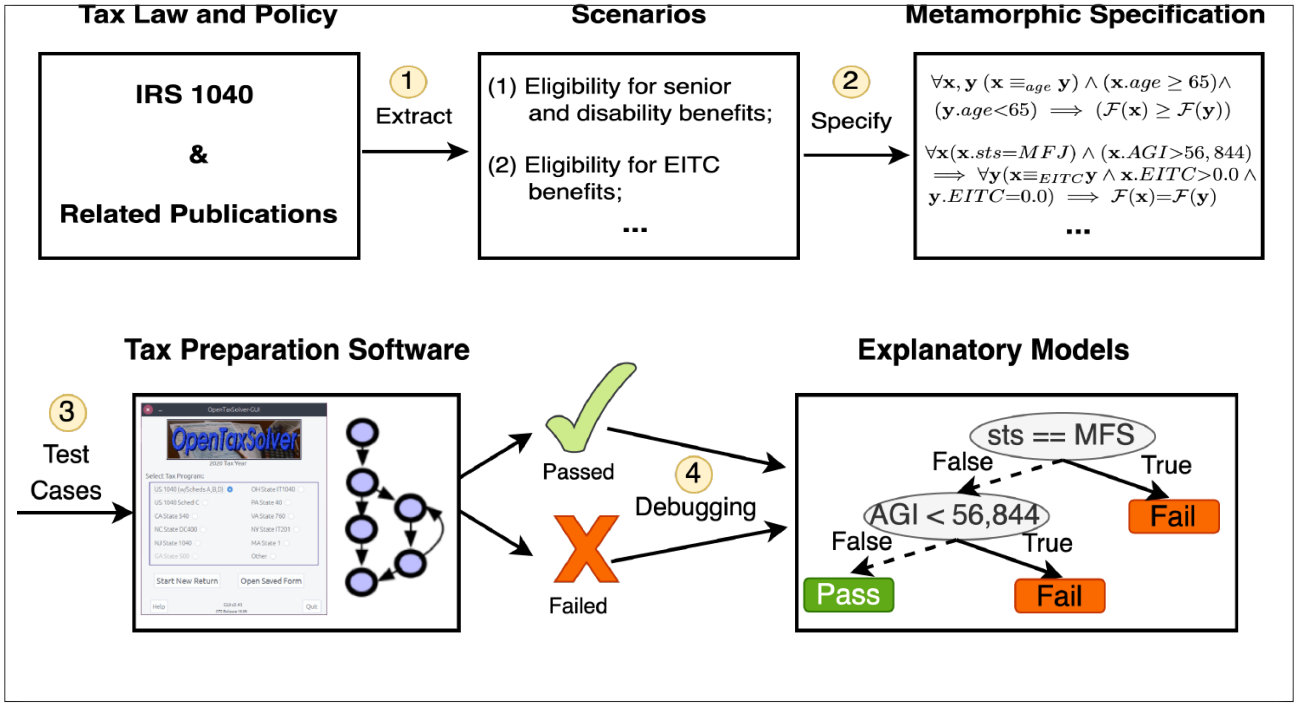
- A blind individual must receive similar or better tax benefits when compared to a sighted person. This is due to higher standard deductions for blind individuals. This equity specification can be expressed as a metamorphic relation:

$$\forall\, \mathbf{x}, \mathbf{y}((\mathbf{x} \equiv \mathit{blind}\, \mathbf{y}) \wedge (\mathbf{x}.\, \mathit{blind} \wedge \neg\, \mathbf{y}.\, \mathit{blind})) \Rightarrow F(\mathbf{x}) \geq F(\mathbf{y})$$

- An individual who qualifies for EITC (e.g., income below $56,844) must receive a higher or equal return than a similar unqualified one.

$$\begin{aligned} &\forall\, \mathbf{x}\ (\mathbf{x}.\ \text{sts=MFJ}) \\ &\qquad\qquad \Rightarrow \forall \mathbf{y}\ (\mathbf{x} \equiv \text{AGI}\ \mathbf{y} \wedge \mathbf{x}.\ \text{AGI} \leq 56{,}844 \wedge \mathbf{y}.\ \text{AGI} > 56{,}844) \\ &\vee \qquad\quad (\mathbf{x} \equiv \text{L27}\ \mathbf{y} \wedge \mathbf{x}.\ \text{L27} > 0.0 \wedge \mathbf{y}.\ \text{L27} = 0.0) \vee (\mathbf{x} \equiv \text{QC}\ \mathbf{y} \wedge \mathbf{x}.\ \text{QC} \geq \mathbf{y}.\ \text{QC}) \\ &\qquad\qquad \Rightarrow \text{F}(\mathbf{x}) \geq \text{F}(\mathbf{y}) \end{aligned}$$

**FIGURE 1. TenForty Framework**




Notes: General Framework using Disability and EITC benefits as examples. Our approach specifies the correctness requirements from relevant tax policies. Then, it generates random test cases and infers decision trees to localize circumstances under which the software fails to satisfy metamorphic requirements.

### *TenForty Framework*

We develop an open-source software, TenForty (Tizpaz-Niari et al. (2023)) (Figure 1), designed to test and debug tax software. While it currently focuses on an open-source tax preparation software, OpenTaxSolver (Roberts (2021)) for the accompanied case study, it can be readily extended to other tax prep software. TenForty allowed us to study the compliance of OpenTaxSolver (Tax Years 2018 to 2021), a popular open-source tax preparation software (Reddit Linux Community (2019); Cherry (2020)), in the domains of disability, credits, and deductions that are known to be challenging and error-prone (Internal Revenue Service (2020b)), leveraging the metamorphic relations. TenForty generates tens of thousands of random test cases using a given compliance requirement as a metamorphic relation. Furthermore, it explains the circumstances under which the software has failed to comply using an explainable ML model (based on CART decision tree algorithm (Breiman et al. (1984)). Our tool has already revealed three types of failures in OpenTaxSolver: missing some eligibility conditions (e.g., married people filing separately status is not eligible to take earned income credits); software fails to satisfy the correctness requirements when the computed tax returns get very close to zero (small non-zero values); and the updated software (e.g., 2021 version updated from 2020 version) that allows users to explicitly opt for an option does not satisfy some correctness requirements in the corner cases.

## 3. Overview: Generating Software Code from Tax Code via LLM

In this section, we overview the LLM-based code generation and our ranking system using some intuitive examples. To illustrate the key concepts, we will use simplified examples focusing on snippets of the generated code and the relevant portions of the input context. The full context provided to the LLMs includes detailed instructions, tax policy updates, and, in some cases, the previous year's tax code. However, for brevity, the figures will only display the code snippets as well as the contextual elements directly related to those snippets.

### *Updating Tax Brackets without Prior Software Code*

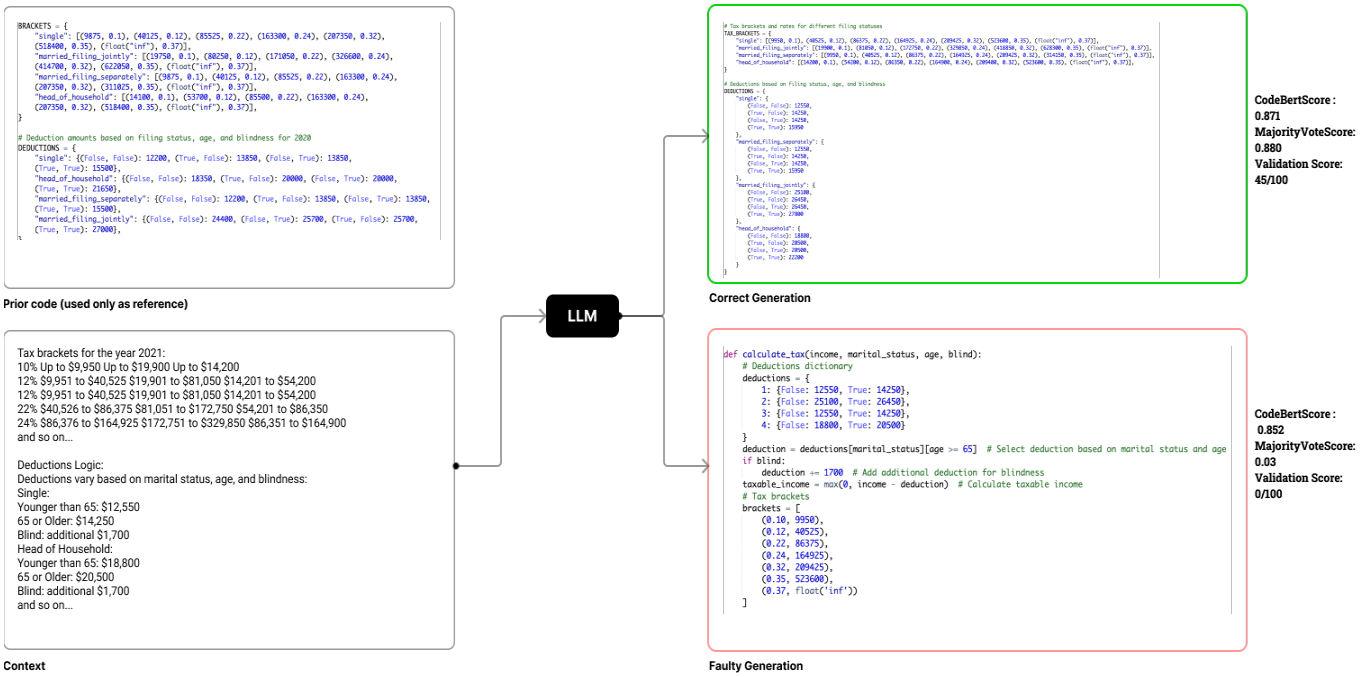
Figure 2 presents a visual comparison of two code snippets generated by an LLM when provided with the 2021 tax law updates and a prompt, but without the context of the previous year's code. The code snippet "Faulty Generation" exhibits several flaws, most notably the incomplete definition of the 'TAX BRACKETS' dictionary. This error results in a syntactically incorrect program and would likely lead to runtime errors. The CodeBERTScore for this snippet is 0.852, reflecting its lower semantic similarity to the reference code and task instructions due to this significant structural error.

In contrast, the code snippet "Correct Generation" demonstrates a better understanding of the task and the required code structure. The 'TAX BRACKETS' dictionary is defined correctly, along with the 'DEDUCTIONS' dictionary. This snippet is syntactically correct and closer to a functional implementation. The CodeBERTScore for this snippet is 0.871, indicating a much stronger semantic alignment with the reference code and the instructions. The ground truth score (or validation score) is 0.45, reflecting the need for further refinement of this candidate solution.

### *Updating Tax Brackets with Prior Software Code*

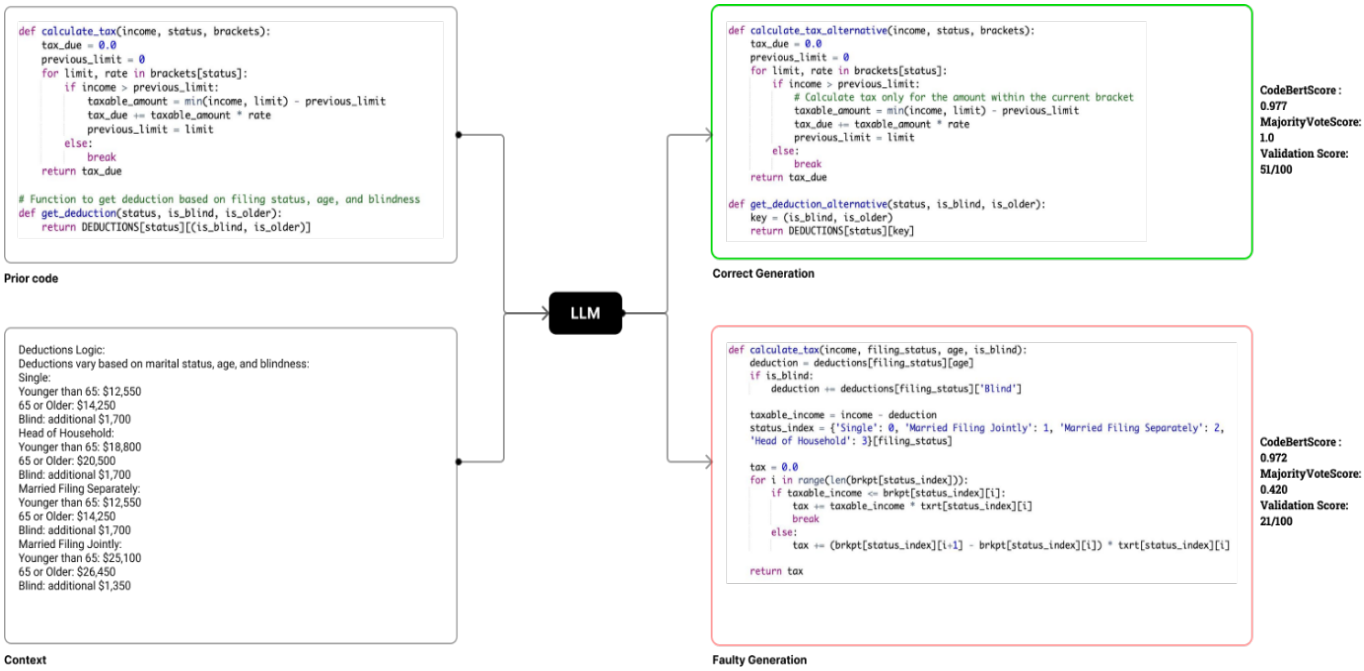
Figure 3 displays two code snippets generated by an LLM when provided with the 2021 tax law updates, the 2020 tax code, and a prompt instructing the LLM to update the code. This scenario demonstrates that even with prior code context, LLMs can generate code with logical errors that might not be immediately apparent.

## FIGURE 2. Updating tax brackets without prior software code




Notes: Prior code is listed only for clarity to understand CodeBERTScore calculation logic; it does not impact the code generation process.

## FIGURE 3. Updating Tax Brackets with Prior Software Code



While both code snippets appear structurally similar to the reference code, the snippet "Faulty Generation" contains several logical errors. For instance, there's a potential misalignment between the tax brackets and their corresponding rates, leading to incorrect tax calculations. Additionally, the code incorrectly calculates the blindness deduction by adding a constant value to an already established deduction, potentially causing a double-counting error. These errors would result in incorrect tax outputs for certain inputs, making the code functionally incorrect. Despite these errors, this snippet achieves a CodeBERTScore of 0.972, demonstrating that semantic similarity alone is insufficient to guarantee code correctness.

The snippet "Correct Generation", however, accurately updates the tax calculation logic. It aligns the tax brackets and rates correctly, and it avoids the double-counting error in the blindness deduction. This snippet achieves a CodeBERTScore of 0.977, slightly higher than the faulty code due to its better semantic alignment.

These comparisons underscore the importance of our multi-faceted ranking approach, which incorporates both CodeBERTScore and MajorityVoteScore. CodeBERTScore provides insights into the semantic quality of the generated code, assessing its alignment with the task instructions and reference code (if provided). However, as demonstrated in Figure 3, semantic similarity alone is not always sufficient to guarantee functional correctness. MajorityVoteScore plays a crucial role in detecting logical errors that might not be evident from the code's structure or syntax. By combining these metrics, our ranking system effectively distinguishes between code candidates with varying levels of quality and correctness, enabling us to select the most promising candidates for further validation and refinement stages of our framework.

## 4. Methodology

**FIGURE 4. AI-assisted framework to update tax software following the updated tax policies.**

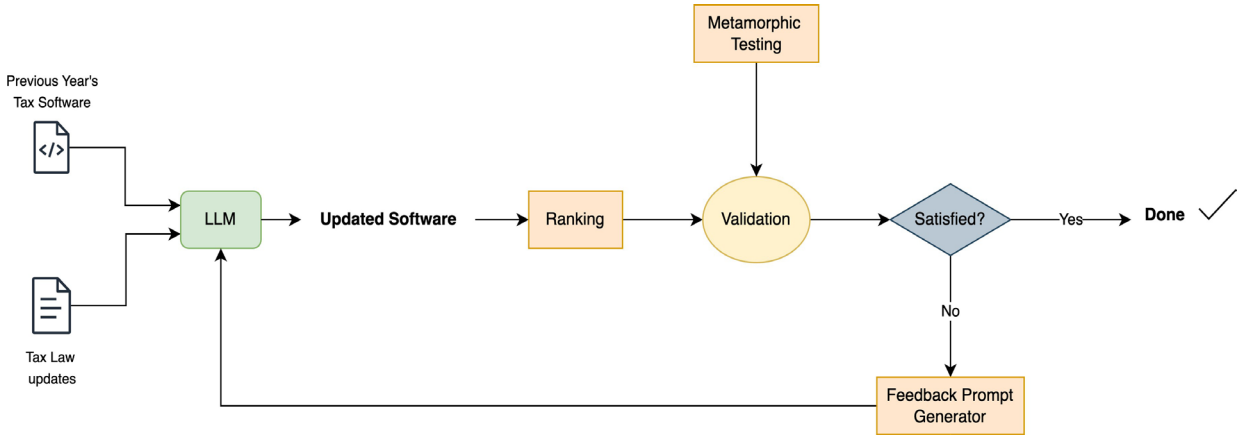



Figure 4 illustrates our proposed framework for automatically updating tax preparation software using Large Language Models (LLMs). This framework tackles the challenge of adapting software to the annual revisions in IRS tax policies, aiming to reduce manual effort and increase the trustworthiness. The framework operates as a cyclical process consisting of several key stages:

1. Input and Analysis: The process begins by providing the LLM with two essential inputs:

   - Previous Year's Tax Software Code: The source code of the existing tax software serves as the base for the update. To understand its importance, we perform experiments without including the previous year's code.
   - Latest Tax Policy Updates: The LLM receives the official IRS publications detailing the changes in tax laws for the current year.

2. Code Generation: Leveraging the provided inputs, the LLM generates multiple candidate versions of the updated tax software. The LLM is guided by a series of prompts that provide context, instructions, and previous year's tax calculation code to maximize the alignments of generated code to the desired functionality. We experiment with two LLMs, ChatGPT-3.5 and GPT-4.0, to explore their effectiveness in this code generation task.
3. Ranking and Selection: Since the LLMs can generate many candidate codes, the main focus of this paper is to come up with a ranking criterion to identify the most promising candidates. We consider the following ranking mechanisms:

   - **CodeBERTScore:** We leverage CodeBERT (Zhou et al. (2023)), a pre-trained model specializing in understanding code, to assess the semantic similarity of the generated code. This metric calculates the cosine similarity between the generated code and both the reference code from the previous year and the IRS policy updates. A higher CodeBERTScore indicates a stronger alignment between the generated code and the intended meaning and structure expressed in the reference code and the new tax regulations.
   - **MajorityVoteScore:** We execute each candidate code with a set of random inputs to quantify the majority vote. These random input profiles cover a diverse range of income levels, filing statuses, and other relevant parameters. For each input profile, we run all generated code versions and record their outputs. We then determine the most frequent output across all versions, assuming this "majority vote" output to be the correct answer. The majority vote score of each code version is then calculated as the percentage of inputs for which its output matches the majority vote output.
   - **WeightedScore:** To determine the overall ranking of the generated code candidates, we employ a weighted average that combines both CodeBERTScore and MajorityVoteScore. This approach allows us to prioritize candidates that excel in two key aspects: semantic similarity to the task instructions and reference code (CodeBERTScore) and functional correctness in producing accurate tax calculations (MajorityVoteScore). We perform various experiments and find that assigning a weight of 0.6 to CodeBERTScore and 0.4 to the MajorityVoteScore works well in practice. The setup also depends on the capabilities of LLMs. More capable LLMs (like GPT-4.0) consistently generate high-quality code, making the majority vote score a reliable indicator of correctness whereas less capable LLMs (such as GPT-3.5) may exhibit greater inconsistency in code quality, relying too heavily on the majority vote score, which could lead to misinterpretations.

4. Metamorphic Testing: To further validate the top-ranked code candidates, we employ metamorphic testing (Tizpaz-Niari et al. (2023); Srinivas et al. (2023)). We previously leverage metamorphic specification and testing to validate the correctness of tax prep software (see Section 2). After we choose a top-ranked candidate, we use the metamorphic testing paradigm to validate its correctness or obtain failed test-cases to guide a refinement process.
5. Feedback Loop:

   - Success: If a code candidate successfully passes the metamorphic testing stage without any failures, we deem it correct and return the solution to the tax software developers as the correct updated software.
   - Refinement: In cases where the code fails one or more metamorphic test cases, our framework initiates a feedback loop for iterative refinement. The Feedback Prompt Generator (FPG) analyzes the specific test failures and generates targeted prompts to guide the LLM in rectifying the identified issues.

6. Iteration: The process of code generation, ranking, metamorphic testing, and feedback- driven refinement may iterate multiple times. This process succeeds if the generated software passes all metamorphic tests, and we obtain some statistically or formal guarantees on the correctness.

Overall, the framework in Figure 1 provides a means to update and maintain tax prep software automatically via LLMs. This paper only focuses on the ranking systems for the LLM-generated code and discusses the technical challenges in using LLMs to update tax prep software as the tax law changes each year. While our prior works used metamorphic testing (Tizpaz-Niari et al. (2023); Srinivas et al. (2023)) to ensure the correctness of general tax prep software,

more work is needed to integrate it as the validation component in the TenForty framework. Also, the feedback prompt generators may not be trivial and require extensive future works to guide LLMs in generating candidate code.

## 5. Experiments and Results

### *5.1 Updating tax prep software without prior code context via LLMs*

This section explores the performance of LLMs in updating tax preparation software when no context about the previous year's code is provided. This scenario examines whether the LLMs are capable in generating tax prep software code from scratch.

#### *Procedure*

We follow the same general framework outlined in the methodology section but omit the initial input of the previous year's code. The LLM receives only the following:

- Tax Policy Updates: The official IRS publications describing the changes in tax laws for the current year (2021 in our experiments).
- Prompt Engineering: A set of instructions guiding the LLM to generate the updated software.

The LLM then generates multiple candidate code versions. These versions are ranked using the CodeBERTScore and majority vote accuracy metrics. The top-ranked candidates undergo metamorphic testing to validate their correctness.

#### *Prompt Engineering*

Here's a specific prompt used to guide the LLM in generating code for the "Brackets Only" scenario:

- Objective: Develop a Python script to calculate federal income tax for the year 2021. The script should accurately compute tax based on the user's annual income and marital status, incorporating the 2021 tax brackets.
- Data Structures:
  - Use dictionaries to map tax brackets for different filing statuses (Single, Married Filing Jointly, Married Filing Separately, Head of Household).
  - Ensure keys are accurately used to prevent KeyError and validate their presence before access.
- User Inputs:
  - `'income': Collect as a float using input(), representing the user's annual income in USD.`
  - `'marital_status': Integer (1-4); 1=Single, 2=Married Filing Jointly, 3=Married Filing Separately, 4=Head of Household.`
- Requirements:
  - The script must compute the tax using the provided tax brackets.
  - Output the tax amount in dollars formatted to two decimals (e.g., `print(f"Tax amount: $tax:.2f")`).
  - Include error handling for user inputs to ensure they are within valid ranges and formats.

[2021 Tax Brackets (concrete numbers should be provided)]

### *General Prompt Template*

We adapt the following template for different scenarios, modifying the specific instructions and data as needed:

- Objective: [Clearly state the task, e.g., "Develop a Python script to calculate federal income tax for the year 2021."]
- Data Structures: [Specify the expected data structures, e.g., dictionaries for tax brackets and deductions.]
- User Inputs: [List the required user inputs and their data types.]
- Requirements: [Outline the functional requirements of the code, e.g., tax calculation logic, output format, error handling.]

[Provide any relevant tax policy data, e.g., tax brackets, deduction amounts, EITC rules.]

**TABLE 1. Results for top 4 ranked code generations out of 10 without prior code.ST:**

| Scenario | LLM | Version | CBS | MVS | Weighted Score | GTS |
|---|---|---|---|---|---|---|
| Brackets | GPT-4.0 | 7 | 0.9 | 1 | 0.935 | 100/100 |
| | | 2 | 0.899 | 1 | 0.934 | 100/100 |
| | | 4 | 0.899 | 1 | 0.934 | 100/100 |
| | | 5 | 0.899 | 1 | 0.934 | 100/100 |
| | GPT-3.5 | 9 | 0.894 | 0.94 | 0.910 | 0/100 |
| | | 2 | 0.892 | 0.94 | 0.909 | 0/100 |
| | | 6 | 0.903 | 0.06 | 0.608 | 0/100 |
| | | 8 | 0.894 | 0.06 | 0.602 | 0/100 |
| Brackets + Deductions | GPT-4.0 | 3 | 0.871 | 0.88 | 0.875 | 45/100 |
| | | 2 | 0.866 | 0.88 | 0.871 | 45/100 |
| | | 5 | 0.861 | 0.88 | 0.869 | 45/100 |
| | | 10 | 0.887 | 0.12 | 0.580 | 0/100 |
| | GPT-3.5 | 2 | 0.859 | 1 | 0.916 | 1/100 |
| | | 1 | 0.859 | 1 | 0.916 | 1/100 |
| | | 6 | 0.858 | 1 | 0.915 | 1/100 |
| | | 10 | 0.858 | 1 | 0.915 | 1/100 |
| Brackets + Deductions + EITC | GPT-4.0 | 7 | 0.883 | 0.79 | 0.827 | 43/100 |
| | | 1 | 0.863 | 0.7 | 0.765 | 25/100 |
| | | 5 | 0.87 | 0.61 | 0.714 | 32/100 |
| | | 6 | 0.857 | 0.61 | 0.709 | 32/100 |
| | GPT-3.5 | 6 | 0.852 | 1 | 0.941 | 2/100 |
| | | 2 | 0.851 | 0.98 | 0.929 | 0/100 |
| | | 10 | 0.845 | 0.98 | 0.926 | 0/100 |
| | | 3 | 0.845 | 0.5 | 0.638 | 0/100 |

Notes: CBS=CodeBERTScore, MVS=MajorityVoteScore, GTS=Ground Truth Score.

### *Results and Discussion*

Table 1 presents the results for the top-performing code candidates generated by GPT-4.0 and GPT-3.5 in each scenario without prior code context. As shown in the table, LLMs demonstrate a varied level of performance depending on the complexity of the scenario and the specific LLM model used.

- **Overall Lower Performance:** We observe a general trend of lower performance across all scenarios when the LLM is not provided with the previous year's code. Both GPT-4.0 and GPT-3.5 exhibit lower MajorityVoteScore and CodeBERTScore compared to when they have the reference code for guidance, because the LLM must come up with the logic of the code itself as opposed to when they are presented with the previous code and can use it as guidance.

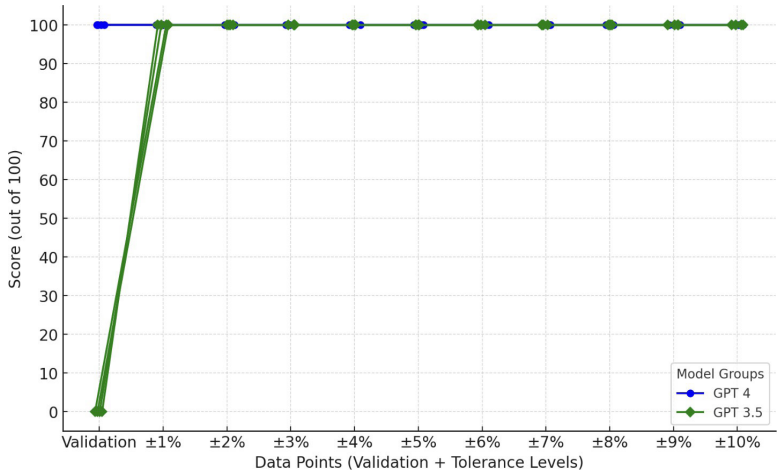
- **GPT-4's Continued Superiority:** Despite the lack of prior code, GPT-4.0 consistently outperforms GPT-3.5. This suggests that GPT-4.0 possesses a stronger capacity for understanding instructions and generating correct code from scratch.
- **Accuracy Decline with Complexity:** As the scenarios become more complex with the inclusion of deductions and EITC, the accuracy of both LLMs drops noticeably, particularly for GPT-3.5. This underscores the challenges LLMs face in generating intricate tax logic from scratch without the benefit of a reference code to guide the process.
- **MajorityVoteScore vs. Ground Truth Discrepancies:** A crucial observation is the occasional disparity between high majority vote accuracy and significantly lower ground truth matching scores. This discrepancy suggests that LLMs can sometimes generate code that consistently produces the most common output, but still contains subtle errors that cause deviations from the ideal calculation. This finding emphasizes the importance of considering other ranking metrics and their combinations.

While Table 1 presents the absolute accuracy of the generated code candidates, it doesn't provide insights into how close their outputs are to the true tax calculations. To gain a more nuanced understanding of the code's correctness, we analyze the accuracy with respect to acceptable error margins from the ground truth. Furthermore, the scatter plots in Figure 5 illustrate the accuracy of generated code candidates (y-axis) based on an acceptable threshold of error margins from the ground truth. The plots show the consistency of ChatGPT-4.0 compared to ChatGPT-3.5. Even if the output of code is not the exact ground truth, the output from ChatGPT-4.0 generated codes is almost always within the 10% error margins of ground truth values which suggests that the logic of the code is sound but there might be some small problems. This also emphasizes that the ranking part of our framework works well in finding codes that have the potential for fixing. ChatGPT-3.5 shows that it cannot generate sound, high-quality code from scratch consistently. ChatGPT-3.5 is especially fragile when it does not generate at least two good codes that can have consensus on outputs. While Table 1 presents the absolute accuracy of the generated code candidates when LLMs are not provided with prior year code, it does not reveal how close their outputs are to the true tax calculations. To understand the proximity of generated outputs to the ground truth, we analyze the accuracy within an acceptable error margin.

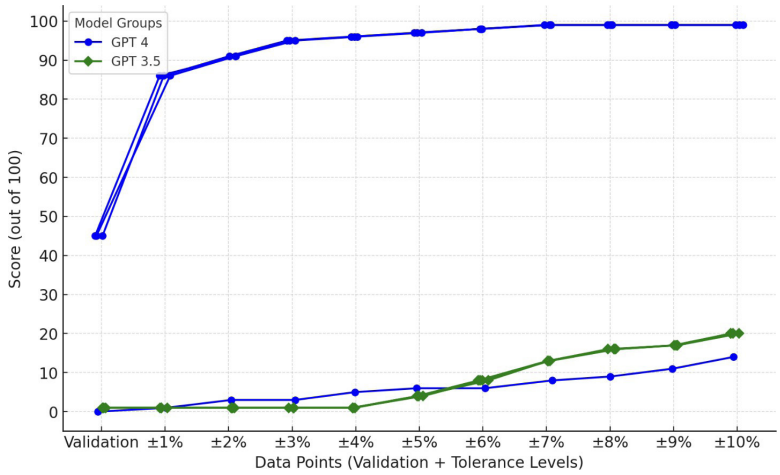
When prompted without the reference code, ChatGPT-3.5's top-ranked candidates struggle to achieve high accuracy, even with a generous error margin. They barely reach 10% accuracy even when allowing a $\delta$ of 10%. This suggests that ChatGPT-3.5, when generating code from scratch, often produces codes that may have a wrong logic. Conversely, ChatGPT-4.0's top-ranked candidates, even without prior code context, exhibit better performance. They consistently achieve 100%-accuracy when considering a $\delta$ threshold of at least 7%. However, it's important to note that achieving perfect accuracy at a 7% error margin still indicates the presence of errors that require refinement.

**FIGURE 5. Scenarios without prior code for 4 top ranked candidates per ChatGPT-3.5/4.0.**

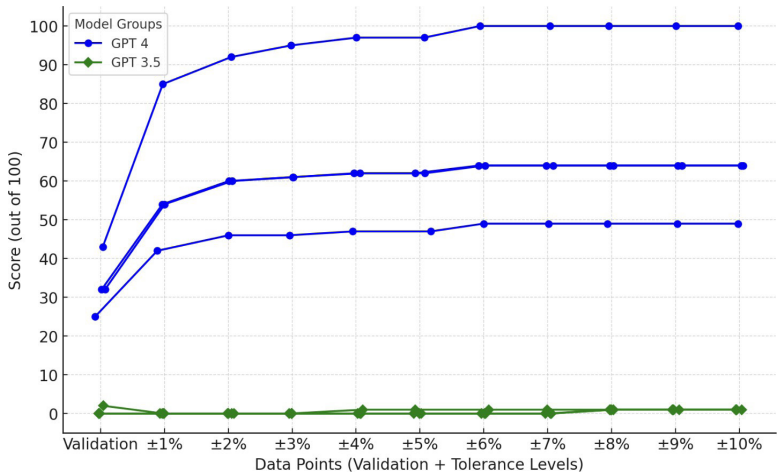
(a) Brackets




(b) Brackets + Deductions




(c) Brackets + Ded + EITC



## 5.2 Updating tax prep software with prior code context via LLMs

This section investigates the performance of LLMs in updating tax software when provided with the previous year's code as context. This scenario emulates a more realistic use case where the LLM can leverage existing code structure and logic as a foundation for incorporating tax policy changes.

### *Procedure*

Following the framework outlined in the methodology section, the LLM receives the following inputs:

- Previous Year's Tax Software Code: The source code of the existing tax software (for 2020 in our experiments) acts as a basis for the update.
- Tax Policy Updates: The official IRS publications detailing the changes in tax laws for the current year (2021 in our case).
- Prompt: A set of instructions that guide the LLM in modifying the provided code to reflect the new tax policy.

The LLM generates multiple updated code versions, which are then ranked using CodeBERTScore and MajorityVoteScore.

### *Prompt Engineering*

Here's a specific prompt used to guide the LLM in generating code for the "Brackets + Deductions" scenario:

- Objective: Update the provided Python code to calculate federal income tax for the year 2021. The updated script should accurately compute tax based on the user's annual income, marital status, age, and blindness status, incorporating the 2021 tax brackets and standard deductions.
- Reference Python Code from 2020:

```
# Constants for tax brackets and rates for 2020
BRACKETS = {
      "single": [(9875, 0.1), (40125, 0.12), (85525, 0.22),
            (163300, 0.24), (207350, 0.32), (518400, 0.35),
            (float("inf"), 0.37)],
      # ... [Other filing statuses] ...
}

# Deduction amounts based on filing status, age, and blindness for 2020
DEDUCTIONS = {
      "single": {(False, False): 12200, (True, False): 13850,
            (False, True): 13850, (True, True): 15500},
      # ... [Other filing statuses] ...
}

# ... [Rest of the 2020 code] ...
```

- Instructions (User Inputs + Requirements):
  - ► Update the BRACKETS dictionary to reflect the 2021 tax brackets.
  - ► Update the DEDUCTIONS dictionary to incorporate the 2021 standard deduction.
  - ► Ensure the script accurately calculates tax based on income, filing status, age, and blindness status.
  - ► Maintain the same user input format (income, marital status, age, blindness).

► Output the tax amount in dollars formatted to two decimals.

[The 2021 tax brackets and deduction amounts.]

### *General Prompt Template with Code Context:*

- Objective: [Clearly state the task, including the year of the provided code and the desired year for the updated code.]
- Reference Python Code: [Previous year's code.]
- User Inputs: [Provide specific instructions on how to update the provided code, referencing variable names or functions as needed.]
- Requirements: [Specify any changes in user input format or output requirements.]

[Provide the necessary tax policy data for the target year.]

**TABLE 2. Results for top 4 ranked code generations out of 10 with prior code.**

| Scenario | LLM | Version | CBS | MVS | Weighted Score | GTS |
|---|---|---|---|---|---|---|
| Brackets | GPT-4.0 | 3 | 0.914 | 1 | 0.944 | 100/100 |
| | | 5 | 0.911 | 1 | 0.942 | 100/100 |
| | | 9 | 0.911 | 1 | 0.592 | 100/100 |
| | | 4 | 0.910 | 1 | 0.941 | 100/100 |
| | GPT-3.5 | 1 | 0.941 | 1 | 0.962 | 100/100 |
| | | 2 | 0.939 | 1 | 0.960 | 100/100 |
| | | 7 | 0.937 | 1 | 0.959 | 100/100 |
| | | 8 | 0.936 | 0.59 | 0.815 | 59/100 |
| Brackets + Deductions | GPT-4.0 | 7 | 0.972 | 1 | 0.983 | 51/100 |
| | | 5 | 0.972 | 1 | 0.983 | 51/100 |
| | | 3 | 0.972 | 1 | 0.983 | 51/100 |
| | | 6 | 0.972 | 1 | 0.983 | 51/100 |
| | GPT-3.5 | 4 | 0.976 | 1 | 0.990 | 21/100 |
| | | 3 | 0.976 | 1 | 0.990 | 21/100 |
| | | 6 | 0.975 | 1 | 0.990 | 21/100 |
| | | 5 | 0.975 | 1 | 0.990 | 21/100 |
| Brackets + Deductions + EITC | GPT-4.0 | 6 | 0.978 | 1 | 0.991 | 48/100 |
| | | 8 | 0.978 | 1 | 0.991 | 48/100 |
| | | 10 | 0.976 | 1 | 0.991 | 48/100 |
| | | 3 | 0.976 | 1 | 0.991 | 48/100 |
| | GPT-3.5 | 1 | 0.986 | 1 | 0.994 | 56/100 |
| | | 2 | 0.977 | 0.92 | 0.943 | 56/100 |
| | | 7 | 0.977 | 0.56 | 0.727 | 35/100 |
| | | 3 | 0.977 | 0.56 | 0.727 | 35/100 |

Notes: CBS=CodeBERTScore, MVS=MajorityVoteScore, GTS=Ground Truth Score.

### *Results and Discussion*

Table 2 shows the results for the top-ranked code candidates generated by ChatGPT-3.5 and ChatGPT-4.0 when provided with prior code context. When comparing Table 2 and Table 1, it is evident that providing prior code as a base significantly enhances the performance of LLMs in updating tax software.

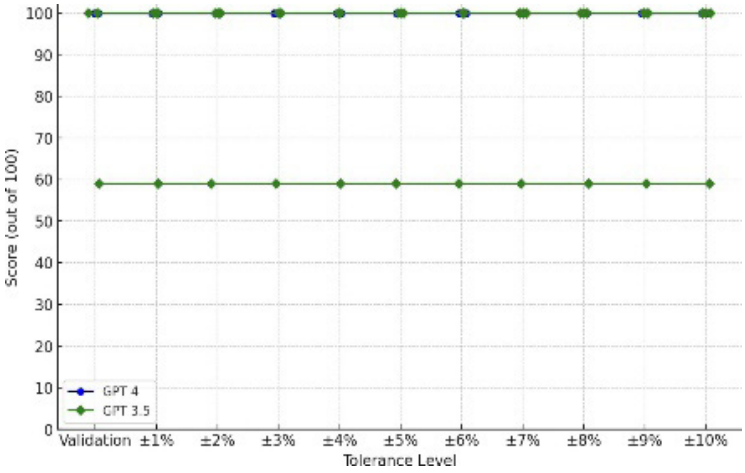
- Overall Strong Performance: Both ChatGPT-4.0 and ChatGPT-3.5 exhibit good performance across all scenarios when provided with the previous year's code. The CodeBERTScore are generally high, and MajorityVoteScore is often near-perfect, particularly for GPT-4. This suggests that LLMs can effectively leverage existing code structure to incorporate new tax policy updates.
- ChatGPT-4.0's Consistent Excellence: ChatGPT-4.0 consistently achieves a higher CodeBERTScore and greater accuracy compared to ChatGPT-3.5. In many cases, ChatGPT-4.0 generates code that achieves both perfect MajorityVoteScore and a perfect match with the ground truth outcomes.
- ChatGPT-3.5's Stroke of Genius: Surprisingly, ChatGPT 3.5 showed great performance for the most complicated scenario when given the prior year's code. We can see it has better top ranked codes than ChatGPT-4.0. Upon further investigation by looking at charts in Figure 6, we can see that, although ChatGPT-3.5 has generated better performing top ranks, but it lacks consistency as it also generated codes that have wrong logic as opposed to ChatGPT-4.0, where, if you only look at the Ground Truth Score, it performs worse than ChatGPT-3.5. However, by looking at charts, we can see that the logic of the generated codes via ChatGPT-4.0 are more sound and robust.
- Brackets Only—Near-Perfect Results: In the simplest "Brackets Only" scenario, both LLMs excel, with GPT-4.0 consistently achieving perfect results. This indicates that LLMs can easily adapt existing code to update tax brackets with high precision.
- Lower Ground Truth Matching: As scenarios become more complex, the ground truth matching scores decrease, even when MajorityVoteScore remains high. This reveals the presence of subtle errors that might not affect the most frequent output but still deviate from the ideal tax calculation. This suggests that, although LLM might be more confident in its generation, that does not mean it will generate a code that produces the exact tax in each scenario, emphasizing the need for comprehensive testing methods to detect such nuanced errors.
- EITC Complexity: The "Brackets + Deductions + EITC" scenario presents the most significant challenge. While GPT-4.0 maintains high MajorityVoteScore, the ground truth score drops, indicating that EITC logic is still difficult for LLMs to implement accurately, even with prior code context. This suggests that complex tax calculations might require more sophisticated prompting strategies or the integration of additional knowledge sources to guide the LLMs effectively.

Table 2 provides a snapshot of the absolute accuracy and ranking scores of the top code candidates. However, to assess the robustness of the generated code and its potential for refinement, we analyze the accuracy across a range of error tolerance thresholds. The scatter plots in Figure 6 visualize the percentage of matching outputs for various tolerance levels when the LLMs are provided with prior year code. Once again, we observe a striking difference in the consistency of ChatGPT-4.0 when compared to ChatGPT-3.5. The generated code by ChatGPT-4.0 consistently achieves near-perfect or perfect accuracy, even at stringent tolerance levels. In fact, as highlighted in the introduction, ChatGPT-4.0 achieves 100% accuracy when allowing an error margin ($\delta$) of at least 7, demonstrating its ability to produce code that aligns closely with the ground truth calculations.

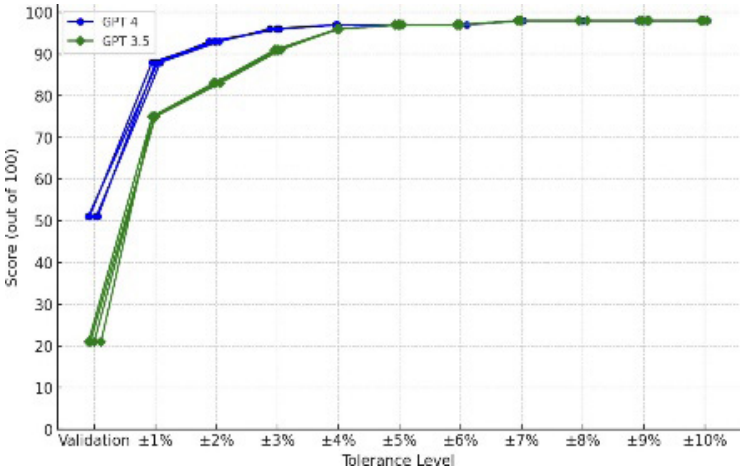
However, a closer look at the scatter plots reveals a nuanced trend: while ChatGPT-4.0 excels at stricter tolerances, ChatGPT-3.5 often exhibits better performance for some generations as the margin of error increases. For instance, at a tolerance level of 5% or higher, ChatGPT-3.5 consistently achieves the same or even better accuracy when compared to GPT-4.0. Although it may not produce good quality code as often as ChatGPT-4.0, this suggests that ChatGPT-3.5 can leverage the provided code context to generate code that is more robust to larger error margins. This observation has important implications for our framework. ChatGPT-3.5, when guided by prior code, might be particularly well-suited for scenarios where a higher tolerance for error is acceptable. Its ability to consistently generate code within a broader acceptable range could be valuable in specific applications. Conversely, GPT-4.0 remains the preferred choice when precision is paramount, as it consistently produces outputs that closely match the ground truth, even at stringent tolerance levels.

**FIGURE 6. Scenarios with prior code contexts for 4 top ranked candidates per ChatGPT-3.5/4.0.**

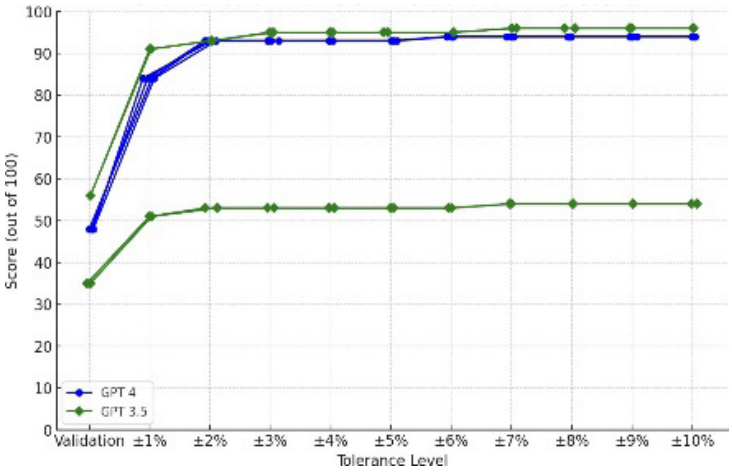
(a) Brackets




(b) Brackets + Deductions




(c) Brackets + Ded + EITC



These results further validate the effectiveness of our ranking approach. Even in scenarios where the generated code is not perfectly accurate, the ranking system successfully identifies candidates, especially those generated by GPT-4.0, that exhibit high potential for being refined into fully correct implementations. The scatter plots, by visualizing the accuracy across different error margins, provide insights into the robustness of the generated code and the need for the validation and feedback prompts to achieve the desired level of accuracy.

## 6. Discussion

Since the completion of our initial study, which was conducted in Fall 2023 and Spring 2024 and focused on ranking code (primarily in C programming language), we have continued to refine our approach to automating tax preparation software updates using Large Language Models (LLMs). A significant advancement is the improvement in model capabilities; even smaller LLMs are now able to accurately modify existing Python code by incorporating new values and adapting to recent tax policy updates. Although this progress is not related to ranking code, it underscores the potential of these models not only in replicating prior implementations but also in generating reliable updates with minimal human intervention. Considering these advancements, we are developing a more robust framework aimed at improving the reliability of software updates for tax calculations. This new framework is being tested via symbolic executions across a wider range of LLMs and more complex tax scenarios to assess their ability to autonomously manage code modifications and additions. Preliminary results indicate that even smaller LLMs can achieve high accuracy in updating and extending Python code, which is promising for the future of automated tax software maintenance. The ongoing work is expected to significantly aid tax prep software developers to update their code as the tax law evolves every year.

## 7. Conclusion

The ever-growing complexity of tax law and policies has significantly increased the role of tax preparation software in navigating the intricacies of legal accountability and compliance. As the tax law gets updated, maintaining the compliance and trustworthiness of tax prep software is challenging. As part of a wider NSF-sponsored project, our goal is to develop principled techniques and tools to support software programmers in maintaining tax preparation software. Our framework combines best practices from formal methods (metamorphic specifications), software engineering (automated testing and debugging), and AI (LLMs for code-generation) to ensure that the software not only adheres to the latest tax regulations but also remains easy-to-maintain. By leveraging this integrated approach, we aim to reduce the time and effort required for updates, enhance the accuracy of tax calculations, and ultimately improve the reliability and user trust in tax preparation software. This will enable programmers to more effectively respond to legislative changes and meet the needs of taxpayers efficiently.

## References

Barr, Earl T., Mark Harman, Phil McMinn, Muzammil Shahbaz, and Shin Yoo (2015). The oracle problem in software testing: A survey. IEEE Transactions on Software Engineering, 41(5):507– 525, 2015.

Breiman, L., J.H. Friedman, R.A. Olshen, and C.I. Stone (1984). Classification and regression trees. Wadsworth: Belmont, CA, 1984.

Cherry, Jessica (2020). Use opentaxsolver as an open-source alternative to TurboTax. https://opensource.com/article/20/2/open-source-taxes, 2020. Online.

Escher, Nel and Nikola Banovic (2020). Exposing error in poverty management technology: A method for auditing government benefits screening tools. Proc. ACM Hum. Comput. Interact., 4(CSCW):064:1–064:20, 2020.

Fan, Angela, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy, Shubho Sengupta, Shin Yoo, and Jie M Zhang (2023). Large language models for software engineering: Survey and open problems. arXiv preprint arXiv:2310.03533, 2023.

Hindle, Abram, Earl T. Barr, Mark Gabel, Zhendong Su, and Premkumar Devanbu (2016). On the naturalness of software. Communications of the ACM, 59(5):122–131, 2016.

IBIS World (2023). Tax preparation services in the us — market size, industry analysis, trends and forecasts. 2023. Online. https://www.ibisworld.com/united-states/market-research-reports/tax-preparation-services-industry/

Internal Revenue Service (2020a). Filing statistics for week ending December 11, 2020. 2020. Online. https://www.irs.gov/newsroom/filing-season-statistics-for-week-ending-december-11-2020.

Internal Revenue Service (2020b). Filing taxes 101: Common errors taxpayers should avoid. https://www.irs.gov/newsroom/filing-taxes-101-common-errors-taxpayers-should-avoid, 2020. Online.

Internal Revenue Service (2020c). Six reasons 90% of people will e-file their tax returns. Online.

Internal Revenue Service (2021a). Form 8863, Education Credits (American Opportunity and Lifetime Learning Credits). https://www.irs.gov/pub/irs-prior/f8863--2021.pdf, 2021. Online.

Internal Revenue Service (2021b). Publication 524, Credit for the Elderly or the Disabled. https://www.irs.gov/pub/irs-prior/p524--2021.pdf, 2021. Online.

Internal Revenue Service (2021c). Publication 596, Earned Income Credit. https://www.irs.gov/pub/irs-prior/p596--2021.pdf, 2021. Online.

Internal Revenue Service (2021d). Schedule A (Form 1040), Itemized deductions. https://www.irs.gov/pub/irs-prior/f1040sa--2021.pdf, 2021. Online.

Internal Revenue Service (2021e). Schedule 8812 (Form 1040), Credits for qualifying children and other dependents. https://www.irs.gov/pub/irs-prior/f1040s8--2021.pdf, 2021. Online.

Internal Revenue Service (2024). Direct file, file your taxes for free. https://directfile.irs.gov/, 2024. Online.

Li, Raymond, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim (2023). Starcoder: may the source be with you! arXiv preprint arXiv:2305.06161, 2023.

Mehta, Priya, Sandeep Kumar, Ravi Kumar, Ch. Sobhan Babu, (2022). Enhancement to training of bidirectional GAN: An approach to demystify tax fraud. arXiv preprint arXiv:2208.07675, 2022.

Reddit Linux Community (2019). Open-source alternative to TurboTax called open-source tax solver. https://www.reddit.com/r/linux/comments/bhp3cq/open_source_alternative_ to_turbotax_called/, 2019. Online.

Roberts, Aston (2021). Open tax solver. https://sourceforge.net/projects/opentaxsolver/, 2021. Online.

Srinivas, Dananjay, Rohan Das, Saeid Tizpaz-Niari, Ashutosh Trivedi, and Maria Leonor Pacheco (2023). On the potential and limitations of few-shot in-context learning to generate meta- morphic specifications for tax preparation software, 2023. The Proceedings of the Natural Legal Language Processing Workshop, EMNLP 2023.

Tizpaz-Niari, Saeid, Verya Monjezi, Morgan Wagner, Shiva Darian, Krystia Reed, and Ashutosh Trivedi (2023). Metamorphic testing and debugging of tax preparation software. In 2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS), pages 138–149, 2023.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022.

Zhou, Shuyan, Uri Alon, Sumit Agarwal, and Graham Neubig (2023). CodeBERTScore: Evaluating code generation with pretrained models of code. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 13921–13937, Singapore, December 2023. Association for Computational Linguistics.

# More Information or More Frequent Information? A Proposal for Quarterly 1099s

*Kathleen DeLaney Thomas (University of North Carolina)*[1]

## 1. Introduction

Third-party information reporting enhances tax compliance. When substantial information reporting is present, the compliance rate reaches 94 percent, compared to just 45 percent when there is little or no information reporting.[2] Accordingly, policymakers have expanded information reporting requirements over the past several decades to enhance revenue collection. More recently, Congress has expanded information reporting requirements for third-party settlement organizations ("TPSOs") by significantly lowering the reporting threshold from $20,000 to $600 as well as eliminating the requirement for 200 or more transactions. While such a shift will subject more taxpayers to information reporting, it will also create additional burden on the IRS to process the influx of new information returns. Although the new $600 threshold was enacted in 2021, the IRS has announced it will delay enforcement until at least 2025, using the old $20,000/200 transactions threshold for 2023, and a phased $5,000 reporting threshold for 2024.[3]

The IRS's delayed implementation of the new $600 reporting threshold for TPSOs illustrates a tension in tax administration. More information is generally better for tax enforcement, as subjecting more taxpayers to information reporting means that more individuals should be deterred from cheating and should report their income accurately. However, casting a wider net imposes costs. First, more information returns impose a greater burden on the IRS to process those returns, as well as greater costs on the third parties that must issue the returns. Second, casting a wider net among taxpayers will likely increase the chances that nonreportable income shows up on information returns (for example, gross proceeds from casual sales that do not exceed basis), increasing complexity and confusion among taxpayers. Third, and relatedly, if information returns become too prevalent (particularly for nontaxable income), taxpayers may perceive they are not meaningful and they may lose some of their deterrent effect. Thus, in setting a threshold for information reporting, policymakers face a tradeoff between these costs and the foregone revenue that results from unreported income.

The current approach to enhancing tax enforcement through information reporting has been to expand its use through either lowering the reporting threshold (as in the recent case of TPSOs) or widening the scope of third parties required to report. Either approach generally results in more information returns issued to more taxpayers. However, there is a third approach that has received virtually no attention in the United States: policymakers could also increase the frequency and efficacy of tax information sent to taxpayers. More specifically, Congress could require information returns to be sent quarterly to align with taxpayers' estimated tax payment deadlines. While receiving quarterly tax information would likely help taxpayers make timely estimated tax payments, this approach is also not without costs. Third parties would have an increased burden to compile and distribute tax information four times rather than once a year. And although the IRS would not have to process quarterly information returns (which would be sent only to taxpayers), it would have to enforce a requirement to send quarterly returns (for example, by imposing penalties on third parties who fail to do so). This article will explore the tradeoffs between the current approach of expanding the scope of information reporting with an approach that requires more frequent information.

## 2. Background on Information Reporting

One of the government's most effective tools for encouraging tax compliance is information reporting, which is when a third party (i.e., not the taxpayer or the IRS) reports the taxpayer's income on an information return. The information return is sent to both the taxpayer and to the IRS after the end of the year, and the IRS then uses the form to monitor whether the taxpayer has accurately reported the income.

[1] Aubrey L. Brooks Distinguished Professor of Law, University of North Carolina School of Law. I am grateful to Emily Cauble, Robert Weinberger, participants at the 2024 IRS/TPC Joint Research Conference on Tax Administration, and participants at the 2024 Mid-Level Tax Conference in Chicago for helpful feedback on this paper.

[2] IRS Publication 5869, Tax Gap Projections for Tax Years 2020 and 2021, Figure 4.

[3] IR-2023-221, IRS announces delay in Form 1099-K reporting threshold for third party platform payments in 2023; plans for a threshold of $5,000 for 2024 to phase in implementation, Nov. 21, 2023.

IRS compliance data illustrates the important role that information plays in promoting compliance. The overall rate of compliance in the United States, measured by the ratio of taxes collected versus taxes owed, is about 85 percent.[4] Much of that high compliance rate is attributable to income that is subject to information reporting. For employee wages, which are both reported on a Form W-2 and subject to withholding, compliance is nearly perfect (99 percent). Income that is not subject to withholding but subject to substantial information reporting (e.g., interest and dividends) is also reported accurately at very high rates (94 percent). On the other hand, compliance is significantly lower when information reporting is not present. The IRS estimates the compliance rate for income not subject to information reporting to be 45 percent.

Commentators have suggested two main reasons that information reporting is so effective. First, providing the IRS with information about taxpayers corrects information asymmetries and allows the agency to pursue those who underreport their income. Second, information reporting acts as a deterrent because taxpayers likely know the IRS is receiving information about their income and are therefore less likely to underreport.[5]

Whether a taxpayer is subject to third-party information reporting depends on the source of their income and how much the taxpayer earns. Employee wages are generally reportable on Form W-2 and are also subject to withholding. Other forms of income, such as interest, dividends, and sales of securities by brokers, as well as certain payments to independent contractors, are reportable on a Form 1099 (though not subject to withholding). This article focuses on payments made to independent contractors, including online platform workers.

Certain payments made by businesses to independent contractors are reportable on Form 1099-MISC if the payments exceed $600 during the year, including payments for services.[6] In 2008, Congress expanded information reporting for some independent contractors that are paid through "third party settlement organizations" (TPSOs).[7] A TPSO generally serves as an intermediary to facilitate online transactions between buyers and sellers, charging a fee for its services.[8] TPSOs include online payment platforms, like Venmo and PayPal, as well as other types of platforms on which taxpayers earn income from performing services or selling goods, like Uber or Etsy. The 2008 legislation required the online platform to issue a Form 1099-K to any taxpayer paid more than $20,000 and accumulating payments from more than 200 transactions on the platform during the tax year.[9] The higher $20,000 threshold "trumped" the $600 threshold under the 1099-MISC rules if both applied, meaning independent contractors paid through online platforms were subject to a much higher threshold for information reporting.

In response to criticism that the disparate reporting thresholds for independent contractors ($600 versus $20,000) were confusing and arbitrary,[10] and due to growing concern about lack of tax compliance in the gig economy, Congress amended the reporting threshold for Form 1099-K in 2021.[11] The new rule, enacted as part of the American Rescue Plan, unifies the reporting threshold between the 1099-MISC rules and the 1099-K rules. Under the new statute, online platforms must issue a Form 1099-K to any taxpayer who earns more than $600 from the platform during the tax year; there is no minimum number of transactions required. This new reporting threshold is expected to substantially increase the number of taxpayers subject to information reporting and raise an estimated $8.4 billion in additional revenue over the next decade.[12]

In sum, the new Form 1099-K rule, which lowers the reporting threshold for payments from online platforms from $20,000 to $600, reflects the general trend in expanding information reporting over the past several decades. That trend is to require more year-end 1099s, which means more taxpayers will receive them.

---

4 See Tax Gap Projections for Tax Years 2020 and 2021: https://www.irs.gov/pub/irs-pdf/p5869.pdf

5 Professor Leandra Lederman compares this effect to red light cameras that catch drivers running red lights: "[T]he taxpayer is aware the government is watching." Leandra Lederman, Reducing Information Gaps to Reduce the Tax Gap: When is Information Reporting Warranted? 78 Fordham L. Rev. 1733, 1737-38. (2010); see also Jay A. Soled, Homage to Information Returns, 27 Va. Tax Rev. 371, 371 (2007).

6 I.R.C. § 6401(a).

7 I.R.C. § 6050W; Treas. Reg. 1.6050W-1. The 1099-K reporting rule did not take effect until 2012.

8 Congressional Research Service, Payment Settlement Entities and IRS Reporting Requirements 1, https://crsreports.congress.gov/product/pdf/IF/IF12095.

9 I.R.C. § 6050W(a), (e).

10 See, *e.g.,* Shu-Yi Oei & Diane Ring, Can Sharing Be Taxed? 93 Wash. U. L. Rev. 989, 1034-38 (2016).

11 The American Rescue Plan Act of 2021 (ARPA; P.L. 117- 2), Section 9674.

12 Congressional Research Service, Payment Settlement Entities and IRS Reporting Requirements 2, https://crsreports.congress.gov/product/pdf/IF/IF12095

## 3. Quarterly Information Returns

Another approach to expanding information reporting, which has received considerably less attention by policymakers and scholars, would be to require that more frequent information be sent to taxpayers.[13] More specifically, Congress could require third parties to send information returns to taxpayers every quarter, which would line up with taxpayers' obligations to pay estimated taxes. While quarterly information reporting could be mandated for many types of income, this article focuses on requiring quarterly information returns for independent contractors receiving a Form 1099-K from TPSOs. This proposed quarterly return—a Form 1099-ES— is explored in more detail in earlier work.[14]

Under quarterly information reporting, third parties otherwise required to issue a Form 1099-K to a taxpayer at the end of the year would be required to send the taxpayer a Form 1099-ES at the end of every quarter once a certain payment threshold was reached during the tax year. The Form 1099-ES would be sent only to the taxpayer, and not the IRS. Both the taxpayer and the IRS would still receive an annual Form 1099-K.

The quarterly payment threshold for Form 1099-ES could be set at an amount equal to the annual reporting threshold for Form 1099-K (e.g., $600 under the current rules), or at a fraction of the annual threshold. At the current $600 threshold, it is likely not cost effective to set the quarterly 1099-ES threshold any lower. But if Congress were to raise the Form 1099-K reporting threshold, the quarterly 1099-ES threshold might be set at 25 or 50 percent of the annual threshold. Consider, for example, if Congress were to raise the 1099-K reporting threshold to $10,000.[15] A quarterly information return might be required as soon as a taxpayer's gross earnings reach $2,500 (25 percent of the annual threshold). If an independent contractor earned $3,000 in the first quarter of the year, the third-party payer would be required to send them a 1099-ES at the end of the first quarter and for all following quarters reporting gross payments. If an independent contractor earned only $700 in the first quarter and $2000 in the second quarter, the taxpayer would receive their first quarterly 1099-ES in the second quarter, showing both year-to-date gross earnings and earnings for that quarter. The taxpayer would continue to receive a quarterly Form 1099-ES for the remainder of the year.

To assist independent contractors in meeting their estimated tax payment obligations, Form 1099-ES could be sent after the quarterly payment period ends but before estimated tax payments are due. The Internal Revenue Code breaks up the tax year into four payment periods ending March 31, May 31, August 31, and December 31, with quarterly estimated taxes due 15 days after the end of each payment period, on April 15, June 15, September 15, and January 15, respectively.[16] Since TPSOs presumably keep electronic records of taxpayers' earnings, Form 1099-ES could be delivered to taxpayers electronically shortly after the payment period ends, leaving time for the taxpayer to calculate and pay estimated taxes based on the reported earnings. For example, TPSOs could be required to issue a Form 1099-ES within 5 days from the end of the payment period, leaving the taxpayer with 10 days to make an estimated tax payment before the deadline, as reflected below:

**TABLE 1. Sample Schedule for Quarterly Form 1099-ES Deadlines**

| End of Payment Period | 1099 ES Due Date | Estimated Tax Payment Due Date |
|---|---|---|
| March 31 | April 5 | April 15 |
| May 31 | June 5 | June 15 |
| August 31 | September 5 | September 15 |
| December 31 | January 5 | January 15 |

[13] For a detailed proposal advocating quarterly information returns, *see* Kathleen DeLaney Thomas, Rethinking Tax Information: The Case for Quarterly 1099s, 97 So. Cal L. Rev. 1527 (2024).

[14] *Id.* See also Improving the Tax System for Independent Contractors: Quarterly 1099s, 183 TAX NOTES FEDERAL 79 (2024).

[15] See, *e.g.,* S. 1725, Red Tape Reduction Act of 2023, https://www.govinfo.gov/app/details/BILLS-118s1725is.

[16] I.R.C. § 6654(c).

The Form 1099-ES would resemble the year-end Form 1099-K in many respects and would contain the same identifying information for the taxpayer and the third party. The only substantive change would be to Box 1: rather than reporting "Gross Amount of Payment Card/Third Party Network Transactions," the Form 1099-ES would break gross payment reporting into two parts, shown in Figure 1.

**FIGURE 1**

| |
|---|
| Box 1a.<br>Gross Payments for the Payment Period January 1- March 31*<br>$_____<br>*Or adjusted accordingly for the relevant payment period |
| Box 1b.<br>Gross Payments Year to Date<br>$_____ |

Further, while the bottom half of Form 1099-K is blank, the bottom half of Form 1099-ES could contain: 1) the exact deadline for making an estimated tax payment; 2) the website or application for making the payment online (or physical mailing address); and 3) simple information for how to calculate the estimated tax payment based on the gross payments reported.

A final feature of quarterly 1099s could be providing taxpayers with a simplified safe harbor method of calculating estimated taxes. While the Form 1099-ES itself would be a reminder of the obligation to pay, some taxpayers may not understand how to calculate their estimated taxes. Including a simple formula for calculating estimated taxes on the Form 1099-ES (rather than referring taxpayers to a website or publication) would provide independent contractors with a statement of their quarterly earnings along with instructions for remitting taxes all in one place.

The Internal Revenue Code already provides taxpayers with safe harbor rules for avoiding estimated tax penalties: taxpayers may either pay 100 percent of their prior year tax liability or 90 percent of their current tax liability.[17] However, these rules may be difficult for some taxpayers to understand and require information the taxpayer does not have readily available. The simplest safe harbor rule for Form 1099-ES would be one that taxpayers could apply without needing to access any additional information other than what is reflected on the form. Such a safe harbor rule could allow taxpayers to calculate estimated taxes for the payment period as a fixed percentage of the gross receipts reported for that period. For example, Form 1099-ES might provide that the taxpayer may calculate their estimated taxes as 5 percent of the gross payments for that period.[18] Use of the safe harbor formula would be optional: taxpayers who wished to pay more or less could do so, but taxpayers who relied on the 5 percent formula would avoid estimated tax penalties even if they owed additional tax at the end of the year.

As discussed above, third-party information returns have the double benefit of correcting informational asymmetries between taxpayers and the government and deterring noncompliance. An additional benefit of information returns may

[17] I.R.C. § 6654 (d)(1)(B). For taxpayers with adjusted gross income over $150,000, the payments must equal 110% (rather than 100%) of the prior year tax liability. I.R.C. § 6654 (d)(1)(c).

[18] The 5% would be intended to approximate the taxpayer's total tax liability for that period, considering both income tax and self-employment tax. Further study may reveal a more accurate percentage. For further discussion of the 5% rate and examples of its application, *see* Thomas, *supra* note 12, at Part IV.B. For a rough calculation, consider that a taxpayer's net earnings after business deductions might be approximately 30 to 40% of their gross earnings. Further consider that a taxpayer's marginal income tax rate plus self-employment tax rate (15%) is likely to result in a total tax rate in the range of 15 to 37% for low-to-middle income earners. Taking the low end of this range, 15% (tax rate) x 35% (net profit ratio) would be approximately 5% of gross receipts.

be that they reduce compliance burdens for taxpayers because they provide a total gross payment amount (from a particular payer), along with notice of the obligation to report the income. However, under current law, independent contractors receive tax information only once a year, yet they have the obligation to remit and pay taxes quarterly. Quarterly information returns could fill a crucial information gap by providing taxpayers with notice of their quarterly earnings as well as their obligation to make an estimated tax payment.

Studies indicate that many gig economy workers, particularly those who are young and/or inexperienced at reporting self-employment income, struggle with estimated taxes. A 2016 survey of platform workers found that a third of such workers did not know whether they had to pay quarterly estimated taxes, and nearly half did not know how much they would owe in taxes and did not set aside money for taxes.[19] The Government Accountability Office (GAO) has also reported that saving for and remitting estimated taxes is one of the top tax compliance challenges faced by platform workers, based on stakeholder interviews.[20] The goal of quarterly information returns would be to assist taxpayers in meeting their estimated tax payment obligations. The form would provide taxpayers with notice of their obligation to remit taxes quarterly, provide simple instructions for how to estimate income and self-employment taxes, and instructions for how to make a payment. However, the approach is not without costs, and the compliance benefits of more frequent information reporting remain to be seen.

## 4. The Future of Form 1099-K Reporting Remains Uncertain

As the platform economy has expanded, more taxpayers are earning self-employment income. This expansion of independent contractors creates compliance challenges because, in the absence of information reporting, such taxpayers often fail to report their income and/or pay estimated taxes. And because a significant number of platform workers (and other independent contractors) earn less than $20,000 per year in self-employment income, the old $20,000 threshold meant that many of these workers were not subject to any information reporting.[21] The new $600 threshold for 1099-K reporting is intended to subject more taxpayers to information reporting, ideally improving compliance.

However, the $600 threshold for TPSOs has been subject to criticism for being too low, overbroad, and confusing for taxpayers. Some have suggested that $600 is an outdated threshold—one that has never been adjusted for inflation—and should be significantly higher.[22] Others have argued that casting a wider net for 1099s will inevitably result in nontaxable transactions showing up on these forms (for example, a gift or reimbursement paid via a payment platform like Venmo), leading to confusion for taxpayers.[23] Further, since Form 1099-K reports gross receipts, commentators have expressed concerns that taxpayers may overreport taxable income or fail to understand how to convert the number on a 1099-K to (net) taxable income.[24] Finally, commentators have suggested that a $600 threshold will result in too many information returns for the IRS to process at its current capacity.[25] In response to criticisms that the $600 threshold is too low, several proposals in Congress have suggested a compromise threshold such as $5,000 or $10,000. In a similar vein, the IRS has announced it will continue to allow third parties to use the $20,000 threshold for 2023, and then will phase in the new rule by allowing a temporary $5,000 threshold for 2024.[26]

Two additional points related to the Form 1099-K threshold merit consideration. First, the Form 1099 threshold has no bearing on substantive tax liability. So, while casting a wider net of information returns may, indeed, capture more nontaxable transactions, it does not create new tax burdens for independent contractors who earn taxable self-employment income. The question of where to set the threshold, then, is one of administration and enforcement, not substantive tax

[19] See Caroline Bruckner, Shortchanged: The Tax Compliance Challenges of Small Business Operators Driving the On-Demand Platform Economy, KOGOD Tax Policy Center (2016).

[20] Government Accountability Office, Taxpayer Compliance: More Income Reporting Needed for Taxpayers Working Through Online Platforms, GAO-20-366, 14 (May 2020).

[21] *Ibid.*

[22] See, e.g., Steven Chung, The Form 1099's Minimum $600 Reporting Requirement is Almost 70 Years Old Without Adjusting for Inflation, Above the Law (Dec. 29, 2021), https://abovethelaw.com/2021/12/the-form-1099s-minimum-600-reporting-requirement-is-almost-70-years-old-without-adjusting-for-inflation/ ("This has resulted in ordinary payments to be subject to a rule presumably meant for large transactions at the time the law was enacted.").

[23] See, e.g., Carol Miller, Fixing Another Liberal Tax Burden, The Hill (Oct. 13, 2022), https://thehill.com/opinion/congress-blog/3687308-fixing-another-liberal-tax-burden/.

[24] See, e.g., The Coalition for 1099K Fairness, https://1099kfairness.org/issue.

[25] *Ibid.*

[26] *Ibid.,* for a summary of some of the Congressional proposals..

law. In other words, the question is, at what level does the compliance benefit from third-party information reporting outweigh the administrative cost?

Second, even in the presence of Form 1099-K reporting, it appears that independent contractors still exhibit significant rates of noncompliance with respect to quarterly estimated taxes and other tax obligations. For example, a report on noncompliance in the gig economy by the Treasury Inspector General for Tax Administration (TIGTA) found that 25 percent of taxpayers who received a Form 1099-K and filed a Form 1040 did not correctly report the income from the 1099-K, and 13 percent did not report and pay self-employment taxes.[27] Thus, it remains to be seen whether expanding information reporting to more platform workers will result in compliance rates as high as those traditionally seen in the presence of third-party information reporting.

Given the significant criticism of the $600 threshold and the IRS's delay in enforcing it before 2025, the status of Form 1099-K reporting is uncertain. Accordingly, the remainder of this discussion will consider three possibilities with respect to the Form 1099-K threshold: the new $600 takes effect; the new threshold is repealed, and the previous $20,000 threshold is restored; or a compromise threshold of $5,000 is enacted by Congress. The discussion further considers combining quarterly information reporting with these possible annual reporting thresholds.

## 5. Weighing the Approaches: A Lower Threshold vs. Higher Frequency

Third party information reporting presents a complex tradeoff between additional tax revenue generated and costs imposed on third parties and the IRS. IRS administration and enforcement can also impact taxpayers' perceptions of fairness of the tax system, which may also impact voluntary compliance. The discussion below considers five factors in weighing the Form 1099-K threshold and the potential addition of quarterly information reporting: 1) income reported and revenue collected; 2) misreporting and taxpayer confusion; 3) administrative costs to the IRS; 4) costs to third parties; and 5) perceptions of fairness.

### *5.1 Income Reported and Revenue Collected*

There is substantial evidence that more third-party information reporting leads to higher compliance rates in reporting income. Accordingly, the lower the threshold for Form 1099-K reporting, the more income that should be reported by independent contractors. A $600 threshold is likely to generate significantly more income reported than a $5,000 threshold, and a $20,000 threshold should generate the least amount of the three alternatives.

In general, when compliance is high because information reporting is present, compliance rates for taxes paid are also high. This is reflected in the IRS's tax gap estimates, which show that the underreporting gap is by far the biggest source of individual noncompliance, while the underpayment gap is comparatively small.[28] Put more simply, most noncompliance comes from taxpayers not reporting their income, and most unreported income is income that was not subject to third-party information reporting. When income is reported on a Form 1099, historical tax gap data indicates that taxpayers tend to both report it and pay tax on it. This result is intuitive; once the IRS knows about income, taxpayers have little incentive not to pay tax on it as failing to do so will subject them to penalties.

However, the tax payment rate (and revenue generated) from expanded Form 1099-K reporting may not track historic compliance rates. This is because many of the new information returns, by design, will be sent to platform workers and other independent contractors who may not realize they have tax obligations and therefore may not appropriately budget for taxes or remit estimated taxes (including self-employment tax). These taxpayers may fail to pay taxes on reported income because they simply do not have the funds. As the GAO observed in a 2020 report, "Because earnings of some platform workers may be low and earnings and expenses may fluctuate, they can have difficulty estimating their taxes owed and setting aside money to pay the taxes….To the extent these burdens and difficulties confuse workers, they are less likely

[27] TIGTA, Expansion of the Gig Economy Warrants Focus on Improving Self-Employment Tax Compliance (Feb. 14, 2019), at 8, https://www.tigta.gov/sites/default/files/reports/2022-02/201930016fr.pdf.

[28] IRS Publication 5869, Tax Gap Projections for Tax Year 2022, Figure 1 (showing the individual underpayment gap to be $57 billion compared to $396 billion for underreporting).

to pay the estimated tax payments fully and on time and may incur a penalty as a result….[I]f the penalty or amount owed is more than workers can afford, they are at risk of falling into a cycle of noncompliance."[29]

The Joint Committee on Taxation estimates that the change to the $600 threshold for Form 1099-K (from the old $20,000 and 200 transactions threshold) would result in an additional $8.4 billion of revenue from 2021 to 2031.[30] This estimate reflects the idea that more tax information (i.e., a lower threshold for 1099 issuance) should result in more taxable income being reported by taxpayers. However, it is unclear if this estimate relies on historic compliance rates with respect to information reporting (such as compliance rates for dividends and interest) and whether it accounts for potential nonpayment by independent contractors. Accordingly, the magnitude of the compliance benefit of expanding Form 1099 reporting is uncertain.

As discussed above, sending taxpayers quarterly 1099s may serve to notify and educate them about their obligations to pay estimated taxes, help them budget, and improve overall compliance rates for self-employment income. But the magnitude of this compliance benefit is also uncertain. Given that year-end information reporting has a decades-long track record of proven success, it is unclear what, if anything, quarterly returns would add. If the deterrent effect of receiving a year-end Form 1099 is high enough, quarterly 1099s may accelerate tax payments (thereby helping taxpayers avoid estimated tax penalties), but they may not have a significant impact on overall tax remittances. In other words, if most taxpayers report and pay tax on income reported on a Form 1099-K, even if they failed to pay estimated taxes, the government should not lose revenue, particularly if it can be compensated for the late payments through estimated tax penalties. On the other hand, if failing to budget for estimated taxes, and/or overall lack of knowledge about tax obligations, causes independent contractors to not remit their full tax liability by the end of the year, quarterly 1099s may prove to have substantial tax benefits. The TIGTA report discussed above, which found that 25 percent of taxpayers who received a Form 1099-K did not correctly report their income, suggests this is a realistic possibility.[31]

A recent IRS Report on Revenue Estimates for IRS Funding suggests improving compliance with respect to estimated taxes may have significant revenue effects.[32] The report describes the IRS's "Estimated Tax Payments Program," which "seeks to leverage behavioral science tactics like nudging and reminders." Relying on social science research that suggests timely reminders can be effective, the program proposes to "educate and prompt taxpayers well ahead of estimated tax deadlines" with notices. The IRS estimates the initiative will generate an additional $7.5 billion of revenue annually (beginning in 2028), with a total of $53 billion by 2034. The report does not detail how the $7.5 billion figure was determined and notes it is not an "official metric." However, this estimate suggests that improving compliance for quarterly estimated taxes may generate substantial tax revenue. And while quarterly 1099s would impose more costs on third parties than the cost of the generic reminders, a Form 1099-ES that provides the taxpayer's specific earnings for the quarter along with instructions for how to make estimated tax payments is likely to provide a larger compliance benefit than a generic notice.[33]

In sum, lowering the Form 1099-K threshold should generate more income reported and more tax revenue; the lower the threshold, the more revenue (in theory). Although current estimates project $8.4 billion over a decade collected from lowering the threshold from $20,000/200 transactions to $600, the benefit remains uncertain if a substantial portion of new Form 1099-K recipients don't properly budget for and/or pay taxes. Regardless of the annual Form 1099-K threshold, estimated taxes pose a compliance challenge for many independent contractors. Thus, instituting quarterly 1099s has the potential to generate tax revenue even without lowering the annual Form 1099-K threshold to $600. It is possible the largest revenue benefits could be derived from a combination of the two approaches: a lower 1099-K threshold combined with quarterly 1099s.

[29] Government Accountability Office, Taxpayer Compliance: More Income Reporting Needed for Taxpayers Working Through Online Platforms, GAO-20-366, 14 (May 2020).

[30] The Joint Committee on Taxation, Estimated Revenue Effects of H.R. 1319, The "American Rescue Plan Act Of 2021," As Amended by The Senate, Scheduled for Consideration by The House of Representatives, JCX 14-21 (Mar. 9, 2021), https://www.jct.gov/publications/2021/jcx-14-21/.

[31] TIGTA, Expansion of the Gig Economy Warrants Focus on Improving Self-Employment Tax Compliance (Feb. 14, 2019), at 8, https://www.tigta.gov/sites/default/files/reports/2022-02/201930016fr.pdf .

[32] IRS, Return on Investment: Re-Examining Revenue Estimates for IRS Funding, Publication 5901 (2-2024), https://www.irs.gov/pub/irs-pdf/p5901.pdf.

[33] For a more in depth discussion of the social science research supporting a quarterly Form 1099-ES, including why a Form 1099-ES may be more effective than a generic notice, see Kathleen DeLaney Thomas, Rethinking Tax Information: The Case for Quarterly 1099s, So. Cal L. Rev. (forthcoming), draft available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4569358#:~:text=It%20offers%20a%20detailed%20proposal,a%20more%20taxpayer%2Dfocused%20approach.

### *5.2 Unintentional Misreporting and Taxpayer Confusion*

Lowering the Form 1099-K threshold from $20,000 to $600 will subject many more taxpayers to information reporting. In casting a wider net, it is possible that it will be harder to differentiate between taxable and nontaxable (personal) transactions, particularly when it comes to payment apps (like PayPal or Venmo) that are used broadly for both personal and business purposes. For example, some have argued that if the payment platform Venmo sends a Form 1099 to every taxpayer who received more than $600 of payments during the year, taxpayers may receive the form for personal transactions like collecting rent from roommates or reimbursements for a group gift. This has been widely reported in the media.[34] Similarly, taxpayers who sell items casually may not understand that their gross proceeds are not reportable as income because they can offset the sale with their basis in the item sold.

Why would a lower threshold make it harder to differentiate business versus personal transactions? One reason might be that the percentage of personal transactions as a total of all transactions on payment apps becomes much higher when the Form 1099-K threshold is lower. (In other words, many people likely never reach $20,000 worth of payments on Venmo even for personal purposes, and far more likely reach $600.) But a second reason might be that widespread media coverage of the new rule resulting in 1099s that relate to nontaxable transactions may lead taxpayers to assume they can ignore their Form 1099, or that they have received it in error even if they haven't. In this respect, too many Forms 1099-K with respect to online payments may cause them to lose their deterrent effect. Even if taxpayers are not mistaken, the salience of the new $600 threshold may lead taxpayers to a better understanding of how to "game the system." For example, payment apps like Venmo now allow taxpayers to designate transactions as nontaxable (such as "gifts"); this may prompt taxpayers to designate their transactions as nontaxable even when that isn't the case.

Although third parties (like payment platforms) have additional time under the IRS's delayed enforcement of the new $600 rule to put safeguards in place, some taxpayers are likely to inadvertently receive Forms 1099-K for personal transactions. Confusion over these forms could result in overreporting of income by taxpayers or additional costs incurred in determining how to report the incorrect information or in hiring an advisor for assistance. However, these costs must be weighed against the costs of not lowering the Form 1099-K threshold, which has resulted in a significant segment of independent contractors not receiving year-end tax information because they do not earn enough to reach the $20,000 threshold. Not receiving tax information can also result in confusion and inadvertent errors among taxpayers who want to comply but fail to keep adequate records or are not aware of their obligations. The net effect of changing the Form 1099-K threshold on inadvertent errors and the other costs imposed by taxpayer confusion is thus uncertain.

In terms of how quarterly information returns would impact confusion and complexity, third parties may similarly have a hard time differentiating business versus personal transactions under a quarterly Form 1099-ES regime. There is no mechanism, from the payer's side, that would better differentiate quarterly transactions compared to year-end transactions. For example, if a payee incorrectly designates business transactions as personal on a payment app, this incorrectly triggers quarterly 1099s and a year end 1099-K. Thus, the costs of taxpayer confusion with respect to incorrect 1099s is largely the same with or without quarterly 1099s.

However, quarterly 1099s are designed to mitigate mistakes and confusion with respect to paying quarterly estimated taxes on business income. A well-designed Form 1099-ES would provide taxpayers with clear instructions about how to pay estimated taxes and how much. Such a form might also contain a simple statement explaining what taxpayers should do if they receive a Form 1099-ES for a nontaxable transaction. On balance, then, quarterly 1099s may reduce confusion and taxpayer errors.

### *5.3 Administrative Costs to IRS*

A lower year-end threshold for Form 1099-K means more information returns for the IRS to process. According to a 2023 GAO Report, if the $600 threshold had been implemented in 2023, the number of Forms 1099-K received by the

[34] See, e.g., Alicia Adamczyk, No, The IRS Isn't Taxing Your Venmo Transactions, CNBC (Jan. 12, 2022), https://www.cnbc.com/2022/01/12/irs-isnt-taxing-your-venmo-transactions.html; Michelle Singletary, Venmo, PayPal and other payment apps have to tell the IRS about your side hustle if you make more than $600 a year, Washington Post (Jan. 21, 2022), https://www.washingtonpost.com/business/2022/01/21/venmo-paypal-new-income-reporting-requirement/.

IRS would have increased from about 14 million to about 44 million.[35] Further, if there is public perception that the IRS cannot effectively process the influx of information returns resulting from a lower Form 1099-K threshold, the forms may lose some of their deterrent effect.

On the other hand, modernization of IRS technology should allow more information returns to be processed and matched to returns, enhancing compliance. Indeed, information reporting has always imposed administrative costs on the IRS to process those returns, along with enforcement costs to follow up on discrepancies, but substantial compliance benefit has presumably outweighed these costs historically. The question going forward is whether the IRS can effectively manage the number of new Form 1099-K returns once a lower $600 threshold takes effect or if a new compromise threshold of $5,000 were enacted. The lower the threshold, the higher the cost imposed upon the IRS.

Quarterly information returns will pose substantially fewer costs on the IRS because only taxpayers will receive these returns. From the IRS's perspective, they will continue to receive the taxpayer's year-end Form 1099-K but will not receive additional forms during the tax year. However, if a quarterly Form 1099-ES requirement were to take effect, the IRS would have to enforce it, presumably through audits and penalties of third parties. (This is also the case for year-end information returns.) This cost, though comparatively modest, would have to be weighed against the compliance benefit of quarterly 1099s. Although it is beyond the scope of the discussion here, this cost to the IRS would also have to be weighed against the cost of other alternatives such as soft reminder notices issued quarterly by the IRS.

### *5.4 Costs to Third Parties*

Information reporting imposes costs on third parties who must submit information returns to taxpayers and the IRS. By lowering the Form 1099-K threshold for TPSOs, those entities will incur increased compliance costs as the number of information returns they must submit will presumably increase. A $600 threshold will be more costly than a $5,000 threshold, which will be more costly than the older $20,000 threshold. These third-party costs may include internal costs like recordkeeping and employee time, as well as external costs such as payments to software companies or tax return preparers.[36]

Imposing administrative costs on third parties can be efficient and is foundational to tax collection in the United States. For example, withholding imposes administrative costs on employers, but it is more efficient to have taxes collected and paid by the employer than to impose payment obligations on each employee. Withholding also results in higher tax compliance and more revenue collected. Similarly, third-party information reporting in all areas, from investment income to broker transactions to independent contractor payments, can be justified by the compliance benefits. The question presented here is not whether these administrative costs are ever justified, but whether lowering the Form 1099-K threshold to $600 or $5000 is justified by the additional costs imposed on third parties.

As between individual independent contractors and TPSOs (and other third-party reporters), the third parties will often be in a better position to efficiently keep records of payments and report them to the IRS. With advances in technology in the past several decades, recordkeeping by third parties has become increasingly automated, which lowers costs. Economies of scale also allow TPSOs to record and report payments with low marginal costs per additional payee compared to compliance burdens imposed on individual taxpayers. A 2007 GAO study of costs imposed by third-party information returns found that "existing information return costs, both in-house and for external payments, were relatively low."[37] According to the report, one small business employing under 5 people would possibly spend 3-5 hours per year on information reporting, while a business with more than 10,000 employees estimated spending less than 0.005 percent of its yearly staff time. GAO interviews with businesses also revealed that, "[a]s expected, unit prices for services provided to taxpayers by selected software vendors, service bureaus, and return preparers decreased as the number of forms handled increased."[38]

[35] GAO Snapshot, Tax Enforcement: IRS Can Improve Use of Information Returns to Enhance Compliance (Nov. 2023), https://www.gao.gov/assets/d24107095.pdf.

[36] GAO, Tax Administration: Costs and Uses of Third-Party Information Returns, GAO-08-266 (Nov. 2007), https://www.gao.gov/products/gao-08-266.

[37] *Ibid.*

[38] *Id.* at 3

In sum, although lower 1099 thresholds impose more costs on third parties, these costs appear modest, particularly when the third party is sophisticated, large, and already has information reporting obligations. It should also be noted that the proposed $600 threshold is the same as the threshold already in place for Form 1099-MISC and is higher than the information reporting threshold for some other types of income (e.g., interest).

Requiring third parties to issue quarterly information returns is likely to impose higher costs, although the magnitude is uncertain and merits further investigation. Unlike year-end Forms 1099, TPSOs do not already have the infrastructure in place to prepare and remit quarterly information returns. However, there is reason to think these costs may be modest, as they are for annual 1099 reporting. As tax information is digitized and easily shared online, it is less costly to compile and distribute electronically. The information that would be shared with taxpayers each quarter (i.e., gross payments) is information that third parties would already keep records of. Similarly, third parties already collect taxpayer identification numbers at the start of payments in anticipation of issuing a Form 1099-K. Marginal costs should also decrease when many quarterly forms must be issued. In sum, the same technological advancements and economies of scale that justify the cost of year-end 1099 reporting may justify the cost of quarterly information returns. Although quarterly 1099s would undoubtedly impose additional costs beyond a Form 1099-K requirement, justifying the cost depends on the additional revenue generated by quarterly returns and the value of decreasing complexity and compliance costs for independent contractors. The tradeoff between these costs and benefits is uncertain and merits further study.

### *5.5 Perceptions of Fairness*

Commentators have suggested that increased information reporting may enhance perceptions that the tax system is fair.[39] More specifically, the higher compliance rates brought about by information reporting may lead to the perception that everyone is paying their fair share of taxes, level the playing field between honest and dishonest taxpayers, and generally increase taxpayer morale by reducing evasion and decreasing the deficit.[40] However, with respect to the recent changes to Form 1099-K reporting, politicians and interest groups have widely argued that the lower $600 threshold is unfair to businesses and individual taxpayers.[41] Some of these arguments have incorrectly framed the lower reporting threshold as a new "tax increase" on workers, which it is not.[42] Although platform workers who have never received a Form 1099-K may be surprised to receive one under the lower reporting threshold, particularly if they were not previously reporting their earnings, information returns do not change taxpayers' substantive obligations to pay income and self-employment taxes. Regardless, the argument that the new threshold is unfair has been widely publicized and may influence taxpayer perceptions of fairness. That relatively low-earning independent contractors may become subject to penalties and face tax bills they do not have funds to pay might be perceived as particularly unfair, even if those workers should have been paying taxes before the change in information reporting.

Interest groups representing TPSOs are likely to make similar arguments about unfair burdens if a quarterly 1099 requirement were enacted. However, unlike expanded year-end information reporting, there may be less resistance when it comes to individual taxpayers. This is because quarterly returns will not be sent to the IRS and therefore won't impact whether taxpayers face a risk of audit or penalties for not reporting their earnings. In other words, even if some independent contractors oppose expanded Form 1099-K reporting because they generally oppose having to report their earnings to the government, quarterly 1099s don't result in any new or additional tax information being shared with the government.

On the contrary, quarterly 1099s may enhance perceptions of fairness, particularly in connection with a lowered threshold (whether $600 or $5,000) for annual Form 1099-K reporting. Given that one major thread of opposition to the

[39] See, *e.g.,* Chuck Marr & Samantha Jacoby, Center on Budget and Policy Priorities, Reducing the Tax Gap: Five Key Points on Information Reporting (July 2021), https://www.cbpp.org/research/federal-tax/reducing-the-tax-gap-5-key-points-on-information-reporting; Galen Hendricks & Seth Hanlon, Better Tax Enforcement Can Enhance Fairness and Raise More Than $1 Trillion of Revenue (April 2021), https://www.americanprogress.org/article/better-tax-enforcement-can-advance-fairness-raise-1-trillion-revenue/.

[40] Marr & Jacoby.

[41] See, *e.g.,* The Coalition for 1099-K Fairness, https://1099kfairness.org/.

[42] See, *e.g.,* Rick Scott, Press Releases, Sen. Rick Scott's Legislation Recognized On National Taxpayer's Union "No Brainer" List (Sept. 15, 2022), https://www.rickscott.senate.gov/2022/9/sen-rick-scott-s-legislation-recognized-on-national-taxpayers-union-no-brainer-list ("Along with trillions in unnecessary and unrelated in spending in the American Rescue Plan, Biden inserted a tax increase on gig workers, like hardworking Americans that work as drivers for Uber, Lyft or DoorDash.").

lower 1099-K threshold is unfair surprise and burden on gig economy workers,[43] quarterly 1099s (along with a safe harbor option for paying estimated taxes) would prevent surprises at the end of the year and assist taxpayers in budgeting for and making timely remittances of estimated taxes. To the extent criticism of the new 1099-K threshold reflects concerns about taxpayers mistakenly reporting nontaxable income (e.g., gross proceeds that do not exceed tax basis), quarterly 1099s would not exacerbate that concern. If a quarterly Form 1099-ES could incorporate effective guidance as to what types of income do not need to be reported, then quarterly returns may alleviate that concern.

## 6. Conclusions and Issues for Further Study

There are multiple tradeoffs to consider in changing the annual Form 1099-K threshold for TPSOs. The current law (not yet being enforced), which imposes a $600 threshold, would very likely result in more income being reported based on historical compliance trends. Whether higher compliance offsets the costs imposed by the lower threshold remains uncertain. The IRS will have to process substantially more information returns, and more taxpayers are likely to receive a Form 1099 for transactions that are not taxable, which creates complexity and confusion. These factors may also negatively impact taxpayer perceptions of the IRS's enforcement capabilities and/or the fairness of the tax system. However, the costs imposed on third parties are likely modest due to digitization of data. Although a $600 threshold will impose more costs than a higher threshold, these third-party costs are likely outweighed by increased tax compliance.

On the other end of the spectrum, restoring the $20,000 threshold would likely eliminate many of the concerns about confusion and complexity, and would result in comparatively less processing for the IRS and costs for third parties. But given that most platform workers do not exceed $20,000 of payments from TPSOs, the continued opportunities for evasion and foregone tax revenue are likely to be significant if the reporting threshold remains that high. Low compliance rates among taxpayers who are paid by TPSOs may also contribute to perceptions that the tax system arbitrarily favors some taxpayers over others and may negatively impact morale.

A compromise threshold, such as $5,000, may balance the competing considerations of enhanced revenue collection and decreased opportunities for evasion versus creating an influx of returns for the IRS to process and creating confusion with respect to nontaxable transactions. One area that merits further study is to what degree a compromise threshold (e.g., $5,000) would bring a significant number of taxpayers into compliance who were formerly not reporting all or any of their income, and to what degree this threshold might eliminate the unnecessary complexity of small, nontaxable transactions showing up on information returns. It may be the case, however, that with sufficient taxpayer education, confusion and complexity can be reduced to such a level that a $600 threshold is optimal. Similarly, advances in IRS computer systems may mean that the agency can efficiently process the influx of new information returns even at a $600 threshold. The IRS's proposed enforcement plan, which adopts a $5,000 threshold for 2024, may offer opportunities to study the costs and revenue benefits of this compromise approach.

In assessing compliance benefits of a lower threshold for Form 1099-K (either $600 or $5,000), another area that merits study is compliance with respect to payment of estimated taxes and year-end tax obligations (including self-employment taxes) by independent contractors. If, as studies indicate, platform workers and other independent contractors are not able to sufficiently save for tax payments, there could be a gap between additional income reported and revenue collected on that income.

How does increased information reporting compare to more frequent information reporting? The magnitude of the benefit of quarterly information returns is uncertain and merits further study. IRS projections of the benefit from improving estimated tax compliance through soft notices indicate that the potential revenue benefit could be as great, if not greater, than the projected revenue from lowering the information reporting threshold to $600. However, quarterly returns would potentially impose more costs on balance (on third parties) than generic reminder notices sent by the IRS.

[43] See, *e.g.,* Demian Brady, National Taxpayers Union Foundation, Taxpayers Aren't Ready for the Coming 1099-K Deluge – And the IRS May Not Be Either (Jul. 2023), https://www.ntu.org/foundation/detail/taxpayers-arent-ready-for-the-coming-1099-k-deluge-and-the-irs-may-not-be-either ("Unless Congress acts to create a more permanent fix, millions of taxpayers casually selling goods online could be expected to report that income to the IRS — a deluge of information the IRS seems to be ill-prepared to handle….If the proposal is left in place, people who sell casually online or use services like Venmo could be in for a taxing surprise at the end of the year, even though the financial transaction data reported on the 1099-K is not necessarily taxable.")

These costs may be modest given that third parties have digital records of taxpayer earnings, but the logistics of preparing and issuing quarterly statements would need to be assessed.

Quarterly information returns serve a different purpose than year-end information returns, and thus may be best suited as a complement to expanded information reporting requirements rather than an alternative. The most substantial benefit of year-end information returns is they provide the IRS with crucial information about the taxpayer's earnings. While this aids enforcement and creates a deterrent effect, year-end reporting does not help taxpayers understand and manage estimated tax obligations and it does not help them effectively budget for taxes during the year. To the extent a lower Form 1099-K threshold might capture a larger swath of inexperienced taxpayers who struggle to manage self-employment tax obligations, an accompanying quarterly 1099 requirement might mitigate some of the concerns associated with a lower Form 1099-K threshold. For example, quarterly returns might improve overall revenue collected from a lower threshold (by helping taxpayers make timely estimated tax payments), it might reduce confusion and complexity associated with year-end "surprises," and it might improve overall perceptions of fairness.

Finally, imposing a quarterly Form 1099 requirement might provide a political compromise regarding the path forward for 1099-K reporting. For example, policymakers might consider enacting a higher threshold (e.g., $5,000 instead of $600), but in combination with a quarterly 1099 requirement. Such an approach might yield the same or more revenue than the $600 threshold would on its own, while responding to some, though not all, of the concerns regarding confusion, complexity, and fairness.

# Investigating the Impact of Free E-File Letter Intervention on Taxpayer's Tax Filing and Preparation Methods

*Pei-Hua Chen, Astin C. Cornwall, Anne D. Herlache, Scott P. Leary, Alexander E. Saak, Brenda Schafer, Melissa Vigil, and Rizwan U. Javaid (IRS, RAAS)*

## 1. Introduction

The Internal Revenue Service (IRS) has steadily encouraged the transition to electronic filing (e-filing) due to its numerous advantages, such as cost efficiency, expedited processing times, and reduced error rates. This shift is not only advantageous to the IRS, but it also benefits taxpayers through a more streamlined filing process and faster refunds. Despite these incentives, approximately 10% of taxpayers continue to submit returns via paper (IRS, 2021; 2022; 2023). Notably, there exists a segment of the taxpayer population earning less than $73,000 annually that is eligible for free e-filing but has not utilized that option. The IRS Taxpayer Experience Office (TXO) and Research, Applied Analytics, and Statistics (RAAS) collaborated to test ways to effectively encourage eligible paper filers to switch to e-filing, thereby streamlining their experience and supporting the IRS' modernization efforts. This paper focuses on one such effort, testing outreach in promoting the use of the IRS Free File program.

This study aims to measure the impact of behaviorally informed outreach strategies on the preparation and filing methods taxpayers choose when submitting their returns, specifically focusing on individuals who have historically filed their tax returns via paper. This study evaluates whether outreach about IRS Free File will shift eligible taxpayers from paper filing to using the Free File program or e-filing via a different preparation method. The study will also assess whether taxpayer characteristics like age, living in an urban versus rural environment, income, and how complex their taxes are might affect their choice. The findings will provide insights into communication options for encouraging taxpayers to switch to e-filing through the Free File program.

## 2. Background and Related Research

The adoption of e-filing of income tax returns has become increasingly prevalent, offering benefits to both taxpayer and tax authorities. E-filing streamlines the tax filing process, reduces transcription errors, and accelerates return processing and refund turnaround times. It also reduces math errors and provides guidance to help taxpayers claim the tax benefits they deserve. However, despite these advantages, a significant number of taxpayers continue to file paper returns. While this study focuses on the impact of outreach on filing behavior, we also consider taxpayer characteristics that may covary with an individual's choice to e-file (e.g., age, urbanicity, and income level). Including these characteristics in our analysis provides a better understanding of how to tailor outreach to various taxpayer segments.

### *2.1 IRS E-Filing History and Current Status*

The IRS e-filing system has undergone significant development since its inception. Initially introduced in the 1980s on a limited scale, electronic filing gained traction as technology advanced. By the 1990s, the IRS began promoting e-filing as a more efficient and convenient alternative to paper filing, aiming to streamline tax processing and reduce errors. The introduction of the IRS Modernized e-filing (MeF) system in 2004 marked a pivotal moment, providing a more robust platform capable of handling complex tax returns for both individuals and businesses.

According to the MITRE Advancing E-file study Phase 1 Report, the IRS aimed to achieve an 80% e-file rate as established by the IRS Restructuring and Reform Act of 1998 (MITRE Corporation, 2008). This comprehensive strategy addressed various technical and policy considerations to promote e-filing adoption. By 2015, IRS had reached a 50% e-filing milestone. Between 2021 to 2023, the paper filing rate was between 7 and 10% (IRS, 2021; 2022; 2023). As of the latest updates, e-filing remains a cornerstone of the IRS's efforts to modernize tax administration.

## *2.2 Previous E-Filing Studies*

Over the years, the IRS, along with tax researchers in the U.K., has implemented various outreach strategies to improve tax compliance, focusing on encouraging e-filing and increasing the uptake of benefits like the Earned Income Tax Credit (EITC). These studies have utilized different communication methods, including postcards, letters, and social norm messaging (John and Blume, 2018) to nudge taxpayers toward filing their returns and using specific tax preparation methods.

Bhargava and Manoli (2015) conducted a significant experiment targeting EITC-eligible taxpayers who had not claimed the credit. Their study involved sending letters that provided information about the EITC, resulting in a notable increase in the credit's uptake. This demonstrated that targeted communication could effectively prompt taxpayers to claim benefits they might otherwise overlook. However, the study did not differentiate whether the effectiveness was due to the content of the letter or merely the act of sending it.

Similarly, Orlett et al. (2017) examined the impact of informational communication on encouraging non-filers to submit their tax returns. Their study compared the effectiveness of postcards and letters sent to taxpayers who had previously resolved non-filer cases. Both types of communication were found to increase filing rates, with letters generally proving more effective than postcards. Despite these findings, the study did not explore whether the success of the letters in influencing tax return filing behavior was attributable to their content or simply the nudge effect of receiving mail.

Further research by Javaid et al. (2018) during the 2017 filing season investigated the effectiveness of outreach in encouraging paper filers to transition to free-assisted tax preparation methods. This study provided insights into the impact of informational postcards on taxpayers' choices between the IRS Volunteer Income Tax Assistance (VITA), Free File, and paid preparers. While it showed a positive response to outreach, the study compared only a limited set of tax preparation options and the demographic analysis considered age and income, with each broken into two groups.

Moreover, the study by Javaid et al. (2020) and the IRS's 2019 outreach experiments highlighted the importance of providing clear options for free-assisted tax preparation methods, such as VITA and Free File. However, these studies did not explore the effects of these interventions across a broader range of demographic variables.

Herlache et al. (2020) also explored the comparative effectiveness of enforcement versus outreach strategies. Their findings suggested that softer, supportive communication might be more effective for those who are already compliant, while more direct enforcement actions might be necessary for non-filers. However, the study did not assess the impact of the outreach on various tax preparation methods.

The current study adds to the literature by considering IRS outreach in the context of mode of filing and preparation method. It also includes a broader exploration of how demographic factors relate to those choices.

## *2.3 Factors Related to E-filing Behavior: Understanding Individual Differences and Appeal Factors*

When exploring the factors influencing taxpayer filing behavior, it is crucial to consider both the appeal of e-filing compared to paper filing and individual differences. The attractiveness of e-filing versus traditional paper filing methods is a critical factor. Factors such as convenience, ease of use, perceived security and the incentives can influence taxpayers' decisions to adopt e-filing. Understanding these appeal factors is essential for policy makers and tax authorities seeking to promote higher e-filing rates and improve overall tax compliance.

Additionally, demographic characteristics such as age, income level, and education background often play a significant role in shaping taxpayers' filing preferences. These individual differences can impact the adoption and acceptance of e-filing initiatives.

### 2.3.1 Appeal Factors: E-filing vs. Paper Filing

A survey of taxpayer experiences (IRS, 2023b) found that that cost and privacy were key factors in taxpayers' decisions to use an online filing platform. MITRE and YouGov (2023) conducted a survey involving 2,000 taxpayers and provided insights into the factors influencing e-file adoption. The study highlighted the growing preference for electronic filing

methods due to their convenience and efficiency. In terms of using Free File provided by IRS, trust in the IRS and the government plays a large role when selecting software (MITRE and YouGov, 2023).

LoopMe (2024) conducted a survey of 14,771 U.S. consumers between January 18th and January 21st, 2024, and found that 67% of Americans were planning to file their taxes using a self e-file service or an assisted service provider. The three most important factors influencing their choice of tax filing service were 'Efficiency' (21%), 'Convenience' (21%), and 'Cost' (21%). Additionally, relationships with accountants remained important, as many respondents were reluctant to change providers due to established connections. These findings align with MITRE and YouGov's (2023) research, emphasizing the importance of simplicity and efficiency in e-filing adoption.

### 2.3.2 Demographic Influences on Taxpayer Behavior

In a taxpayer experience survey (IRS, 2023b), the IRS found that taxpayers who are younger, self-prepare their returns, or have limited English proficiency were all more likely to be interested in e-filing. In addition, Wang (2003) studied the factors affecting the adoption of e-filing and identified that computer self-efficacy had significant effect on adoption intention. Since direct assessment of computer self-efficacy for each participant was not feasible in our study, we leveraged generational or age differences as a proxy.

Generational cohorts reflect varying levels of exposure and adaptability to digital technology. Younger generations, like Millennials and Gen Z, who were raised during the digital revolution, are likely to be more adept and comfortable using technology, including e-filing systems. According to Parsad, Jones and Greene (2005), the percentage of public schools with internet access increased from 35% in 1994 to 95% in 2005. A survey study conducted by Perrin and Duggan (2015), identified the number of Americans with internet access at home was 67% in 2001. Based on computer and internet accessibility for general households after the early 2000s, we expect that people who are 43 years or younger (Millennials and Gen Z) have higher digital literacy whereas people who are 77 years or older (Silent) have low digital literacy. For those between 44 and 78 years of age (Gen-X and Baby Boomers), there is likely greater variability in digital literacy, stemming from education and work experiences. According to the MITRE and YouGov (2023) taxpayer filing preference survey, younger population shows higher interest in IRS Return Free File since there is little effort required on their part whereas older generations are sticking with what they know and trust and opt to maintain their established filing method.

While Parker (2023) suggests that comparisons between generations should ideally be made using historical data at similar ages, to account for life stage effects, this approach is not directly applicable to the objectives of our study. Our research focuses on capturing the influence of the technological environment prevalent during different generational periods, rather than comparing life stages across generations. In our study, we use age groups to illustrate variations in exposure to the technological landscape, which has fundamentally shaped individuals' interactions with and adoption of digital technologies.

We also evaluate urbanicity, which may covary with access to broadband service. Pippin and Tosun (2014) examine the determinants of e-filing by different population segments and regions, and they found that e-filing rates are lower in rural counties and counties with low population size. Individuals in urban areas may have more reliable access to internet services, which can lead to greater familiarity and comfort with technological resources. Likewise, higher-income individuals potentially have more resources and opportunities to develop and utilize computer skills.

The degree of income tax complexity is also an important factor affecting taxpayers' tax filing and preparation method. The MITRE and YouGov taxpayer filling preference survey (2023) showed that many of the taxpayers with simple tax returns opt to continue using their current software, preferring to maintain the previously adopted filing method. Most taxpayers with complex tax returns are open to the use of free IRS Direct File software instead of paying for commercial software. However, there are some who are willing to pay for commercial software if it provides better data security, audit protections, and customer service.

By understanding the correlations and trends associated with these covariates, we can gain meaningful insights into the characteristics influencing e-filing behavior and potentially identify tailored strategies to encourage the adoption of electronic filing. The current study will add to these insights by evaluating the influence of outreach on filing mode and

preparation method among individuals likely to be eligible for IRS Free File and by layering on how demographic factors may relate to those choices.

### *2.4 Addressing the Research Gap*

The existing body of research has provided valuable insights into the effectiveness of different outreach methods in improving tax compliance. However, there remains a significant gap in understanding whether the content of the communication (e.g., detailed information about tax preparation options) or the mere act of sending a communication (e.g., a nudge effect) is the primary driver of increased compliance. Additionally, previous studies have largely limited their demographic analysis to straightforward categorizations, which may overlook important variations in taxpayer behavior.

This study seeks to address these gaps by conducting a field experiment that compares the effectiveness of sending a detailed letter, a simple tax checklist, and no communication (control group) in influencing taxpayers' filing behavior. Furthermore, this study employs a more nuanced approach to demographic analysis, using income as a continuous variable and considering a broader range of age groups.

## 3. Method

### *3.1 Research Design*

This study aims to measure the impact of behaviorally informed outreach strategies on the methods chosen by taxpayers for preparing and filing their tax returns, specifically focusing on individuals who have historically filed their taxes via paper and appear to be eligible for the IRS Free File program. A stratified random sampling design was employed to evaluate the effectiveness of different intervention methods on filing behavior. Participants were divided into two distinct strata based on their filing history: frequent filers and new or infrequent filers. Within each stratum, participants were randomly assigned to one of three groups—receiving either a letter, a checklist, or no communication (control).

### *3.2 Sample*

The study's sample consists of two distinct strata of taxpayers, each representing a different segment of paper filers who are eligible for free e-filing.

#### 3.2.1 Strata 1: Frequent Filers

This group consists of individuals who have been consistent in their tax filing habits, demonstrating an ongoing engagement with the tax system. The criteria for inclusion in this stratum are the following:

- Taxpayers who self-prepared and paper-filed their taxes in the Tax Year 2021.
- Those who were eligible for the IRS Free File program in Tax Year 2021, which is generally determined by an income threshold—in this case, less than $73,000.
- Individuals who have filed at least one tax return between the Tax Years 2018 and 2020.
- Importantly, these taxpayers should not have paper-filed every year from 2018 to 2021, indicating that while they are frequent filers, they are not exclusively committed to paper filing and may, therefore, be more open to changing their filing method.

#### 3.2.2 Strata 2: New or Infrequent Filers

The second stratum focuses on taxpayers with a less consistent filing history. Its criteria are the following:

- This includes individuals who did not file a tax return from 2018 to 2020.
- They self-prepared and paper-filed their taxes in Tax Year 2021.

- These taxpayers were also eligible for the Free File program in Tax Year 2021.

The reason why we choose to focus on taxpayers that fit the criteria for these strata is twofold. By selecting taxpayers who are eligible for free e-filing, the study effectively isolates the monetary factor as a barrier to e-filing adaptation. This approach allows us to examine whether the e-filing method is inherently appealing to taxpayers once the obstacle of cost is removed. Moreover, non-habitual paper filers are not consistently tied to a single method of filing and may be more open to change compared to habitual paper filers. This flexibility presents a unique opportunity to influence their behavior with outreach interventions. Additionally, new or infrequent filers represent a demographic that might either be new to tax responsibilities or exhibit sporadic engagement with tax filing, which could stem from various factors such as variable income or changes in filing requirements. This group might respond differently to outreach efforts, especially if those efforts alleviate confusion or some of the perceived burden of the filing process. In employing stratified sampling, our aim is to ensure representation from each of these segments. That way our findings can be generalized to the broader population, while also providing insights into the specific needs and barriers faced by each subgroup. In essence, segmenting the population in this manner will produce insights that will allow for a more tailored approach in outreach and intervention strategies.

#### 3.2.3 Sample Population Description

In Table 1 we report the median age, income, and size for the two distinct strata within our study. As can be seen, the Frequent Filers represent an older group of individuals with more income, as compared to the New/Infrequent Filers stratum.

**TABLE 1. Descriptions of the two strata in our sample population**

| Stratum | Median Age | Median Income | Total Number |
|---|---|---|---|
| Frequent Filers | 54 | $23,560 | 107,254 |
| New/Infrequent Filers | 22 | $10,741 | 53,389 |

Given the significant disparities in median age and income between the two populations, we conducted the analysis separately for each stratum rather than combining them into one model with strata as a variable. This approach allows us to tailor our interpretation of the treatment effects to the unique characteristics of each stratum.

### *3.3 Treatments*

The treatments used in this study include:

1. **Free File Letter:** Letter 6171 (See Appendix A) was designed to inform paper filers about the option to e-file their tax returns at no cost, given that their income is below a certain threshold ($73,000 in TY 2022). This letter highlights the benefits of free e-filing, such as speed, security, and accuracy. The Free File letter, acting as the first treatment, is crafted to inform taxpayers of the no-cost electronic filing option available to them, aiming to enhance the convenience and appeal of e-filing. Unlike the checklist, this letter explicitly promotes the use of e-filing, and thus, it is expected to lead to a higher uptake of e-filing amongst the recipients when compared to those who do not receive this information.
2. **Filing Checklist:** Publication 5732 (See Appendix B) is a structured checklist which was provided to assist taxpayers in the filing process, ensuring they have all the necessary information and documentation ready to file their taxes. The checklist, serving as the second treatment of the study, is intended to prompt taxpayers to file their taxes, potentially elevating the overall filing rate compared to the control group. However, since the checklist is not specifically for e-filing, it is not anticipated to directly increase the rate of electronic submissions.
3. **Control Condition:** A segment of the population received no mailing to serve as a baseline against which the impact of the above treatments can be measured. The behavior of this group will help to isolate the effect of the outreach efforts from other variables that might influence the decision to e-file or paper file.

By contrasting the tax return preparation and filing behaviors of those who receive the outreach (either the Free File Letter or the Filing Checklist) with those who do not (the control group), the study seeks to determine whether behaviorally informed materials are effective tools in influencing taxpayer behavior.

**TABLE 2. Treatment Content for Each Group**

| Group No. | Group Type | Letter Content |
|---|---|---|
| 1 | No-Contact Control | None |
| 2 | Treatment Group 1 | Free File Letter (Letter 6171): You may be qualified for free e-file: fast refund, fewer errors and free |
| 3 | Treatment Group 2 | Checklist to file tax (Publication 5732) |

**Mailing Schedule.** A total of 125,000 taxpayers, identified as previous paper filers, were selected to receive behaviorally informed outreach over the course of five mailings. The mailings were split to reduce burden on the print sites. Each mailing consisted of 25,000 letters and was randomized evenly into two groups:

1. **Free File Letter Group:** 12,500 taxpayers were allocated to receive a letter detailing the Free File option, which allows for free electronic filing of tax returns if certain criteria, such as income threshold, are met.
2. **Tax Filing Checklist Group:** Another set of 12,500 taxpayers were chosen to receive a comprehensive checklist intended to guide them through the e-filing process.

Working within the logistic constraints of our print site, we chose to align the distribution with each participant's TY 2021 tax filing dates, with earlier TY 2021 filing dates corresponding to an earlier mailing segment. These five separate mailing dates are: January 17, January 24, January 27, February 3, and February 10, 2023. This approach is based on the effect of temporal proximity on behavioral cues in behavioral science, which suggests that individuals are more receptive to taking related actions when cued at a time close to when they would normally engage in the behavior. By aligning our interventions with each taxpayer's prior filing timeline, we aim to capitalize on the timing of their decision-making process.

### *3.4 Research Questions*

To analyze the impact of our interventions on taxpayers' filing behaviors, our study is driven by a series of focused research questions. These questions aim to dissect the effectiveness of each outreach approach and identify potential differences in response across various taxpayer characteristics. The following research questions have been developed to guide our evaluation and to provide a structural framework for our subsequent analyses and interpretations. Following these questions, we present corresponding hypotheses.

- **Research Question 1:** How does the provision of a Free File letter influence taxpayers' filing choice between e-filing and paper filing?
  - ► **Hypothesis 1:** Individuals who receive a Free File letter will be more likely to e-file compared to individuals in both the control group and the filing checklist group.
- **Research Question 2:** Are there any demographic differences in how the treatments influence the decision to e-file or the overall tax filing rate?
  - ► **Hypothesis 2:** There will be demographic differences in the effectiveness of the outreach treatments. Specifically, we identify three covariates: Age, Urbanicity, Income Tax Complexity.
    - » **Age:** Younger generations, often referred to as digital natives (e.g., Bennett, Maton, and Kervin (2008), have grown up in an era of pervasive technology. Their familiarity with digital platforms, smartphones, and online services positions them as natural candidates for e-filing adoption. On the other hand, older taxpayers exhibit greater variability in their responses to e-filing interventions based on their technology comfort, habitual tax filing behavior, trust in data security, and the need for assistance and support.

Therefore, we expect that:

- **Hypothesis 2.1:** Younger taxpayers will be more likely to e-file in response to the Free File letter, whereas older taxpayers will show greater variability in their responses.
  - **Urbanicity:** Urban areas have historically had better digital infrastructure, including high speed internet access (Molnar, Savage, and Sicker, 2019). Urban residents are more likely to have the technological prerequisites for e-filing. Rural areas may continue to face challenges related to digital access. We propose that:
- **Hypothesis 2.2**: Urban residents will be more responsive to the e-filing intervention than non-urban taxpayers.
  - **Income Tax Complexity:** Taxpayers with straightforward tax situations often involve fewer variables and calculations. These taxpayers are more likely to embrace e-filing for its core advantages of speediness, ease, and reduced paperwork. Individuals with more complex tax situations may have different considerations about and greater variability with e-filing. Some may readily adopt e-filing, while others may prefer to use tax professionals due to familiarity or perceived ability to handle complexity. We propose that:
- **Hypothesis 2.3**: Taxpayers with the least complex tax situations will be more responsive to e-filing interventions than those with more complex tax situations.

- **Research Question 3:** Does receiving a Free File letter or checklist increase the rate of filing taxes compared to not receiving any intervention?
  - **Hypothesis 3:** Individuals who receive a Free File letter or a checklist will be more likely to file their tax returns compared with those who do not receive any intervention.

#### 3.4.1 Exploring Potential Interactions in E-filing Adoption

While we have clear hypotheses for the main effects of factors like treatment group, age, and population density on e-filing adoption, the interaction effects we're considering (Treatment Group ×Urban/Rural, etc.) are exploratory in nature. Here's why:

**Limited Prior Research**: Extensive research on the specific interaction effects we're considering might not be readily available. This makes it challenging to formulate precise predictions about the direction or strength of these interactions.

**Data-Driven Discovery**: By treating these interactions as exploratory, we allow the data themselves to guide us in uncovering potential influences that might not have been anticipated based on existing knowledge.

#### 3.4.2 Understanding Baseline Differences and Treatment Impact

One of our primary goals is to understand both the baseline differences in e-filing adoption between repeat and new filers and how the intervention's effectiveness might vary between these groups. For example: the effectiveness of the letter might be stronger for younger adults who are already comfortable with technology compared to older adults; the Free File letter might be more effective in rural areas where individuals may have more limited access to e-filing options; and the positive effect of receiving a Free File letter (T1) may be stronger for individuals with lower income tax complexity compared to those with higher complexity.

- **H3.1: Treatment Group × Age:** The effect of T1 on e-filing will differ across age groups.
- **H3.2: Treatment group × Urban/rural:** The effect of T1 on e-filing will differ between urban and rural areas.
- **H3.3: Treatment Group × Income Tax Complexity:** The effect of T1 on e-filing will differ depending on the complexity of individuals' tax situations.

## *3.5 Data Handling*

### 3.5.1 Undeliverable Cases

As the mailings were distributed, a quality control measure (postal non-deliverable, PND) was in place to track the deliverability of the materials. In our study, the treatment of undeliverable mailings poses a challenge. These instances may not be random and could indicate underlying differences in the characteristics of these groups.

To handle the PNDs, we will compare the results of "Treatment on the Treated" (TOT) and "Intent to Treat" (ITT) approaches, which are commonly used in clinical trials and other interventional studies. The TOT approach analyzes the effect of the outreach on only those participants who received the letter or checklist. This can provide a more accurate estimate of the effectiveness for those who receive the mailing, but it may introduce selection bias since it includes only those whose mailings were successfully delivered. In the ITT analysis, participants are analyzed in the groups to which they were originally assigned, regardless of whether they received the mailing. This approach is used to preserve the initial random assignment and to provide an unbiased estimate of the outreach's effect within current operational constraints (i.e., not introducing bias regarding differential handling of the treatment and control conditions, e.g., operational constraints of imperfect address information). Comparing the ITT and TOT can tell us whether there is a significant discrepancy between the number of participants assigned to the intervention and those who received it. Due to time constraints, only the more conservative ITT is presented here—we will revisit TOT in future iterations.

### 3.5.2 Early Filers

In our study, individuals who filed their income taxes before the mailing of these letters or checklists are defined as early filers. Strata 1 contains 290 early filers and Strata 2 contains 248. Although they represent a small fraction of the sample (less than 3%), it is important to consider the implication of their actions on this study's findings. Given their minimal percentage, one approach is to exclude these early filers from the analysis on the impact of the mailing. This exclusion is justified as their behavior does not reflect the intervention's influence.

### 3.5.3 Demographic and Social-Economic Characteristics

To conduct a comprehensive and insightful analysis, we will look at our population based on several key demographic and socio-economic characteristics.

**Urbanicity:** We used a population density variable from the Census 2020 dataset. The 2020 variable differs from the 2010 variable in several important ways:

- The minimum population to be classified as an urban area increased from 2,500 to 5,000.
- There is a new alternative classification based on a minimum housing unit threshold.
  - » 425 housing units per square mile define the initial urban core.
  - » Then 200 units per square mile fill in the remainder of the urban area, which is similar to the 2000 and 2010 censuses.
  - » 1,275 housing units per square mile ensures each qualifying urban area contains at least one high density nucleus.
- The category "urbanized area" with a population of at least 50,000 was obsoleted.
- Note: After matching the Census data to the Tax Year 2023 analysis file, we discovered that 2,599 addresses in the treatment groups and 1,977 in the control group were not assigned a population. After reviewing the geographical areas with no population data, it was determined that most are commercial areas or locations where people do not actually reside (i.e., post office boxes). It was also noted that ZIP codes that are close in value numerically are also close geographically. Using this information, we replaced missing values with a population that was determined as follows:

For each ZIP code with a missing value:

- Pull a range of ZIP codes around the ZIP code with a missing value.
- Calculate the absolute value of the numerical distance between the ZIP code with the missing value and each of the six closest ZIP codes.
- Randomly assign the population of one of the six ZIP codes, weighted by population count.
- In our study, the urbanicity variable is coded as "1" for urban and "0" for rural areas.
- Urban: Individuals residing in city settings with more centralized resources and facilities.
- Rural: Individuals living in less densely populated areas with potentially fewer resources or reduced access to high-speed internet.

**Age Groups:** In this study, age is categorized into distinct groups. The age groups are defined as:

**TABLE 3. Age groupings**

| Group | Age Range |
|---|---|
| 1 | Under 30 |
| 2 | 30–44 |
| 3 | 45–59 |
| 4 | 60–74 |
| 5 | 75 and over |

**Income:** Adjusted Gross Income (AGI) is included in the model as a control variable. It was centered and scaled using the formula: AGI minus the mean AGI of its respective strata divided by 10,000, with mean AGI values of 29,241.37 for Strata 1 and 16,781.12 for Strata 2. Missing values were imputed using the median AGI of each combination of strata, treatment group, age group, and urbanicity.

**Income Tax Complexity:** Tax returns are assigned a complexity score as described in Table 4. These scores are based on the types of income, deductions, and credits reported.

**TABLE 4. Income tax complexity**

| Complexity | Definition |
|---|---|
| Low | Wage income; Interest income; Unemployment income; Withholding; Earned income tax credit (with no qualifying children) or advanced EIC; Does not meet any of the conditions for higher levels of differential burden |
| Low-Medium | Capital gain income (includes capital gains distributions and undistributed capital gains); Dividend income; Earned income tax credit (with qualifying children); Estimated tax payments; Retirement income (includes SS benefits, IRA distributions, or pensions and annuities); Any non-refundable credit (includes child and dependent care expenses, education credits, child tax credit, elderly or disabled credit); Household employees; Non-business adjustments; Does not meet any of the conditions for higher levels of differential burden |
| Medium | Itemized deductions (includes mortgage interest, interest paid to financial institution; charitable contributions, and medical expenses); Foreign income, expense, tax, credit, or payment; Moving expenses; Simple Schedule C or C-EZ; General business credit; Does not meet any of the conditions for higher levels of differential burden |

| Complexity | Definition |
|---|---|
| Medium-High | Farm income as reported on Schedule F; Owns rental property as reported on Schedule E, including farm rental and low-income housing; Estate or trust income as reported on Schedule E; Employee business expense deductions; Files AMT without AMT preference items; Prior year alternative minimum tax credit; Investment interest expense deduction; Net loss as reported on Schedule C; Depreciation or amortization as reported on Schedule C; Expenses for business use of home as reported on Schedule C; Does not meet any of the conditions for higher levels of differential burden |
| High | Cost of goods sold as reported on Schedule C; Partnership or S-Corp income as reported on Schedule E; Files AMT with AMT preference items |

Source: RAAS Taxpayer Burden Lab (5/7/2024)

Each of these characteristics was chosen based on its potential to influence e-filing decisions. For instance, younger individuals might be more inclined to e-file due to familiarity with technology, while those in rural areas might face barriers like lack of high-speed internet. Similarly, income tax complexity could affect the perceived ease or difficulty of e-filing.

By addressing these specificities, the insights from our interventions can be used for further tailoring for effective outreach.

#### 3.5.4 Outcome Variables To Be Analyzed

Three outcome variables will be used in this study. The first outcome variable being analyzed is the income tax submission method for the sample of this study. This can be measured as a binary outcome (e.g., e-filed vs. paper-filed) for each individual or as a continuous variable (e.g., percentage increase in e-file adaptation) for each subgroup. The analysis aims to understand how different messaging themes influence this outcome and how demographic factors play a role in a taxpayer's decision to e-file. The second outcome variable is the filing rate for the sample of this study. This can be measured as a binary outcome (e.g., filed vs. non-filed) for each individual or as a continuous variable (e.g., percentage increase in filing individual income taxes) for each subgroup. The third outcome is a categorical tax preparation method variable. It includes Free File, VITA, paid preparer, self-on-paper, and software–prepared-paper-filed (v-coded) returns. The analysis aims to understand how the outreach influences the taxpayer's filing behavior.

#### 3.5.5 Imported Variables

Our comprehensive dataset includes IRS tax administration data that allows for a multi-faceted analysis of taxpayer behavior and response to interventions. Alongside these data, our data collection process incorporates three key input files. The first file consists of postal data, including information on undeliverable mail, which is essential for understanding the reach and effectiveness of our mail-based interventions. The second file is the mailing list data, which provides the SURVEYID list of the study's participants. Participants' ZIP codes were matched with the ZIP code tabulation area (ZCTA) population density data from the 2020 Census to create the urbanicity variable (see Demographic and Social-Economic Characteristics under Data Handling). Additionally, we incorporate a file dedicated to tax complexity, which helps in categorizing taxpayers based on the burden of filing their income taxes.

### *3.6 Regression Approaches*

In this study, we aim to evaluate the impact of sending a letter encouraging e-filing to individuals who are ostensibly eligible for IRS Free File. Our primary challenge is that within our sample, a portion of individuals who did not file their taxes might not be required to file for various reasons, such as being below the income threshold. Although a single model that incorporates the three outcomes (non-file, e-file, and paper file) can capture the decision-making process comprehensively and reduce bias, using a two-step approach might be more appropriate when dealing with uncertainty regarding filing requirements. This approach allows us to separately analyze the decision to file, and, conditional on filing, the method of filing (e-file vs. paper file).

First, we use a logistic regression model to test whether the treatment groups (those who received a letter or a check-

list) have a higher overall tax filing rate compared to the control group. Second, among those who did file, we use another logistic regression model to determine whether the treatment letter increased the likelihood of choosing e-filing over paper filing. By focusing on conditional probabilities, we can accurately assess the treatment effect on e-filing behavior without the confounding influence of individuals who do not need to file taxes.

### 3.6.1 Treatment on E-filed vs. Paper-filed—Binomial Logistic Regression Analysis

To test the significance of our behaviorally informed outreach interventions on choice of filing method, we evaluate multiple predictors, including interactions between demographic characteristics and treatments. This will provide a more nuanced understanding of which factors most influence filers' decisions.

Our main interest is understanding letter treatment effects on the e-filing adoption and what drives e-file adoption. We can use logistic regression to analyze the impact of the treatment intervention (e-file letter and checklist) along with other relevant variables on the likelihood of adopting e-filing. A logistic regression model to predict the likelihood of e-file adoption (coded as 1 for e-filed and 0 for paper filed) can be represented as:

$$\log\left(\frac{P(Y=j)}{P(Y=m)}\right) = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2 + \cdots + \beta_{kj}X_k, \qquad \text{for } j \in \{1,2,\ldots,m-1\}$$

where $Y$ is the categorical outcome variable with $m$ categories ($m$ for binary outcomes), $X_1, X_2, \ldots, X_K$ are predictor variables, $\beta_{0J}, \beta_{1j}, \ldots, \beta_{kJ}$ are coefficients for category $j$, and $P(Y=j)$ is the likelihood of choosing category j.

### 3.6.2 Treatment on Tax Preparation Method—Multinominal Logistic Regression

Multinomial logistic regression provides a robust statistical framework for analyzing categorical outcomes with more than two categories. This approach is well-suited for our study, where taxpayers select their tax preparation method from several options (free e-file, software, self-preparation on paper, etc.).[1]

We model the probability of choosing category (e.g., free e-file) for outcome variable (e.g., tax preparation method), as

$$\mathrm{P}(Y=j|X) = \frac{\exp\left(X'\beta_j\right)}{\sum_{k=1}^{J}\exp(X'\beta_k)}$$

where $X$ represents the vector of independent variables and $\beta_J$ is the vector of coefficients associated with each independent variable. These coefficients indicate the magnitude and direction of the relationship between each variable and the odds of choosing category $J$ compared to the reference category.

By applying this multinomial logistic regression model, we can estimate the odds of a taxpayer choosing each tax preparation method based on the intervention (receiving the Free File letter) and other relevant factors included in the independent variables ($X$). This allows us to assess the independent effect of the Free File letter program on taxpayer behavior and their tax preparation method selection.

This study investigated the factors influencing tax preparation method choice among taxpayers. Initially, a single multinomial logistic regression model was employed to analyze data from both frequent filers and new/infrequent filers. However, upon closer examination, the results proved challenging to interpret effectively.

There were two key reasons why the single model presented difficulties:

- **Heterogeneity**: The data encompassed two distinct populations: frequent filers and new/infrequent filers. These groups likely exhibit significant differences in characteristics such as age, income (mean and standard deviation of Adjusted Gross Income—AGI), and overall tax filing experience. Combining them in a single model could obscure these underlying differences.

[1] Note that the software-prepared, paper-filed returns that were dropped from the analysis due to insufficient numbers for testing and disclosure concerns in reporting the results.

- **Limited Generalizability**: The single model's results might not accurately reflect the behavior of either frequent or new/infrequent filers. The combined effects might not represent the unique factors influencing each group's tax preparation method choices.

To address these limitations, we opted to analyze the data using separate multinomial logistic regression models for each group. By analyzing the data separately, we can gain deeper insights into the distinct tax preparation behaviors of frequent filers and new/infrequent filers.

# 4. Results

## *4.1 Descriptives*

To provide context for the regression analyses, we present the distribution of tax preparation methods among the three groups (control, treatment 1, and treatment 2). We also provide a table that shows the count of individuals using each tax preparation method within each group (See Table C.1 and Table C.2 in the Appendix).

### 4.1.1 Filing Rate

**TABLE 5. Filing by treatment**

| | N | Filed | % |
|---|---|---|---|
| Letter | 53,473 | 36,624 | 68.5 |
| Checklist | 53,370 | 36,944 | 69.2 |
| Control | 53,600 | 32,156 | 60.0 |

The Free File letter and checklist may act as a nudge—an intervention designed to influence behavior without forcing it—to promote tax filing. They might increase awareness and potentially lead to a higher filing rate. It appears that the interventions did have a positive impact on filing. We expand on this in Section 4.2.

### 4.1.2 Electronic Filing Rate

**TABLE 6. Electronic filing by treatment**

| | N | E-filed | % |
|---|---|---|---|
| Letter | 36,624 | 14,535 | 39.7 |
| Checklist | 36,944 | 14,107 | 38.2 |
| Control | 32,156 | 12,449 | 38.7 |

Conditional on filing, we do not see a large difference in electronic filing among conditions within our grouped sample. Breaking this outcome out by strata provides a more nuanced view (Table 7).

**TABLE 7. Electronic filing by frequency of filing (strata)**

| Type of Filer | New/ Infrequent | Repeat |
|---|---|---|
| Letter | 4,727 | 9,808 |
| Checklist | 4,480 | 9,627 |
| Control | 3,701 | 8,748 |
| | **% E-filed** | |
| Letter | 44.6 | 37.7 |
| Checklist | 42.2 | 36.6 |
| Control | 43.0 | 37.1 |

As expected, the group that received the Free File letter that included information about the benefits of electronically filing had more individuals file electronically. Similarly, individuals with a less established filing preference had more individuals file electronically.

**TABLE 8. Electronic filing by age group**

| Age Group | <30 | 30 44 | 45 59 | 60 74 | > 75 |
|---|---|---|---|---|---|
| Letter | 7,012 | 1,774 | 1,851 | 2,624 | 1,277 |
| Checklist | 6,721 | 1,563 | 1,844 | 2,659 | 1,320 |
| Control | 5,178 | 1,863 | 1,862 | 2,364 | 1,182 |
| | | | % E-filed | | |
| Letter | 48.2 | 46.4 | 37.0 | 31.5 | 25.8 |
| Checklist | 46.2 | 42.2 | 36.4 | 30.6 | 26.8 |
| Control | 48.2 | 48.0 | 37.3 | 29.8 | 25.6 |

Individuals under the age of 30 who received the Free File letter had had greater numbers filing electronically. Interestingly, age group 2 (30-44 years old) had fewer individuals filing electronically in the treatment groups than the control group.

**TABLE 9. Electronic filing by urbanicity**

| Urban | Yes | No |
|---|---|---|
| Letter | 4,700 | 9,835 |
| Checklist | 2,639 | 9,468 |
| Control | 4,121 | 8,328 |
| | % Electronically Filed | |
| Letter | 38.7 | 40.2 |
| Checklist | 37.3 | 38.6 |
| Control | 38.1 | 39.0 |

Perhaps the most interesting descriptive finding is that rural filers showed a slightly greater number filing electronically than urban filers. This is something we will explore further in future analyses.

**TABLE 10. Electronic filing by return complexity**

| Complexity Category | Low | Low Medium | Medium | Medium High | High |
|---|---|---|---|---|---|
| Letter | 7,297 | 5,068 | 1,447 | 566 | 157 |
| Checklist | 6,962 | 4,997 | 1,400 | 583 | 165 |
| Control | 5,731 | 4,562 | 1,374 | 595 | 187 |
| | | | % E-filed | | |
| Letter | 44.7 | 35.7 | 39.6 | 31.0 | 25.2 |
| Checklist | 42.5 | 34.6 | 38.1 | 32.7 | 25.6 |
| Control | 44.6 | 35.0 | 37.5 | 31.1 | 26.6 |

Across the conditions, those with low and low-medium complexity returns showed higher numbers of electronic filers than those among the med-high and high complexity categories. This is likely related to age (Table 8); younger taxpayers are less likely to have the types of income, deductions, and credits that are included in the higher complexity categories.

### *4.2 Results of Treatment on Filed vs. Not Filed*

Due to the absence of administrative data for non-filers, we estimated the average effect of treatments on the filing rate under the assumption that the filing decisions follows a binomial distribution, with covariates randomized across both treatment and control groups. Chi-squared tests for homogeneity were used to evaluate whether the treatments (letter and checklist) had a significant effect on the filing rates compared to the control group.

For Stratum 1 (frequent filers), the chi-squared test indicated a significant effect of the treatments on filing rates ( $\chi_2^2 = 659.613, p < 0.01$). The Cramer's V for this test was 0.078, indicating a small effect size, suggesting that while the treatments significantly influence filing decisions, the overall strength of this association was modest. For Stratum 2 (New/Infrequent filers), the chi-squared test also indicated a significant effect of the treatments on filing rates ( $\chi_2^2 = 659.613, p < 0.01$). ($\chi_2^2 = 642.596, p < 0.01$) . The Cramer's V for this test was 0.114, indicating a small to moderate effect size. This suggests that the treatments have a small to moderate, yet statistically significant influence on filing decisions.

Further analysis compared the effectiveness of the letter and checklist treatments within each stratum. The chi-squared test statistic for the comparison between the letter and checklist group was significant $(\chi_1^2 = 8.001, p < 0.01)$ for Stratum 1 with Cramer's V equal to 0.07, reflecting a small effect size, suggesting that the checklist treatment had a slightly greater impact on filing behavior compared to the letter. Conversely, the comparison between the letter and checklist groups for Stratum 2 was not significant $(\chi_1^2 = 0.35, p > 0.01)$.

Overall, these results suggest that both the letter and checklist treatments significantly increased filing rates compared with the control group, with an 8% increase among frequent filers and a 12% increase among new or infrequent filers. While the effects of treatments on the filing rates of infrequent filers are similar, the checklist has a somewhat bigger effect on frequent filers than the letter treatment.

### *4.3 Results of Treatment on E-filed vs. Paper-filed*

The binary logistic regression examined the effects of receiving a Free File letter or a checklist on the likelihood e-filing versus paper filing among both frequent (Table 11) and new or infrequent filers (Table 12). The main effects of the treatments revealed that the checklist has a negative effect on e-filing behavior for participants in both strata. For frequent filers, the checklist treatment showed a negative effect on e-filing with an odds ratio of *0.913 ($p<0.1$)*. This indicates that frequent filers who received a checklist were 9% less likely to e-file compared with the control group, suggesting that the checklist may discourage frequent filers from choosing to e-file. For new or infrequent filers, the negative effect of the checklist was even stronger. Participants in this group who received a checklist were approximately 21% less likely to e-file their taxes compared to the control group, reflecting a significant deterrent effect. The letters, in contrast, did not show a significant effect on e-filing behavior for either stratum. Here it is important to note that e-filing refers to any electronically submitted return. This encompasses more options than the Free File Program, which was the focus of the treatment letter. We present the treatment impact on Free File usage later in the results.

**TABLE 11. Filing by treatment by Stratum**

| | N | Filed | % |
|---|---|---|---|
| **Strata 1, Frequent filers** | | | |
| Letter | 35,677 | 26,020 | 72.93% |
| Checklist | 35,627 | 26,317 | 73.87% |
| Control | 35,750 | 23,553 | 65.88% |
| **Strata 2, Infrequent filers** | | | |
| Letter | 17,796 | 10,604 | 59.59% |
| Checklist | 17,743 | 10,627 | 59.89% |
| Control | 17,850 | 8,603 | 48.20% |

AGI has a statistically significant positive effect on the e-filing in both strata. The odds ratio is 1.08 for frequent filers and 1.069 for new or infrequent filers, indicating that for every $10,000 increase in AGI, the odds of e-filing increase by approximately 8% and 6.9% respectively. All age groups showed significant effects on e-filing behavior *($p < 0.05$)*. As shown in Figure 1 and 2, both groups (repeat filers and new or infrequent filers) exhibit a similar trend: as age increases, participants were consistently less likely to e-file their taxes, and this negative effect becomes stronger with each successive age group. It shows that older individuals are progressively less inclined to choose e-filing over paper filing.

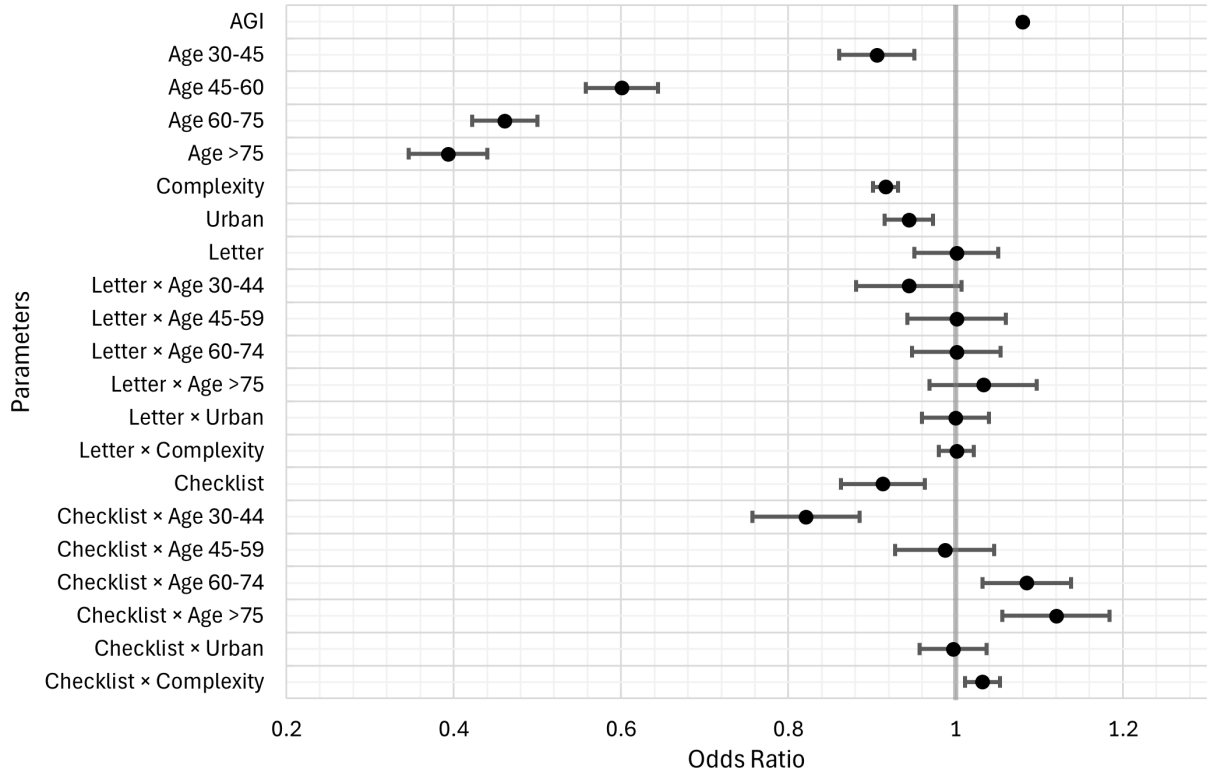
Income tax complexity negatively affected e-filing behavior in both strata. For each 1-unit increase in the income tax complexity score, taxpayers were 8 to 9% less likely to e-file.

Urbanicity had a negative effect on e-filing only among frequent filers, with those participants being 6% less likely to e-file.

For frequent filers, the interaction effects between the checklist treatment and the age group 30 to 44 showed a negative effect on e-filing behavior compared with the under 30 group. Conversely, the interaction between the checklist and those over 75 years of age had a positive effect on e-filing, suggesting that the checklist was effective in encouraging e-filing among the oldest age group. The other groups were not as responsive to the mailings.

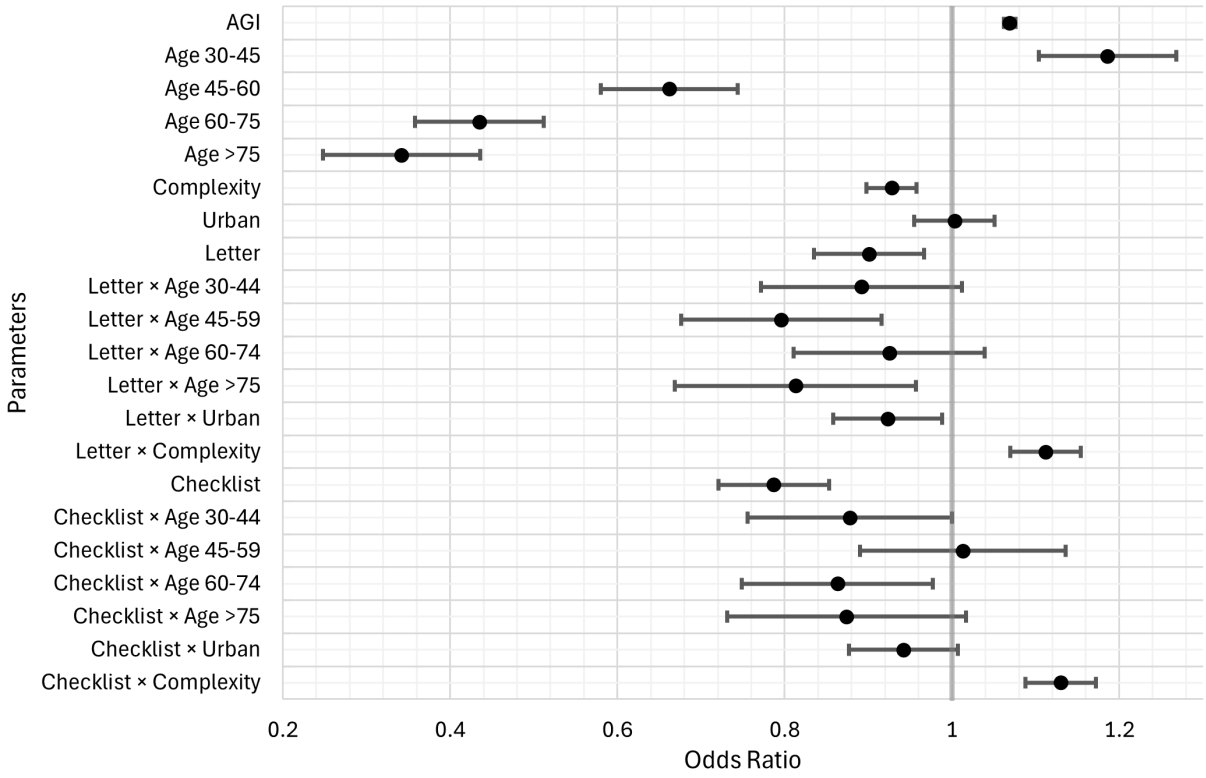
In the new or infrequent filer group, interaction effects between the treatments (letters and checklists) and income tax complexity have a significant positive effect on e-filing behavior *($p<0.05$)*. As the complexity score increases by one unit, participants receiving a letter are 11.2% more likely to e-file, whereas participants receiving a checklist are 13% more likely to e-file.

**FIGURE 1**




NOTE: see the Appendix C for results in tabular form.

**FIGURE 2**




NOTE: see the Appendix C for results in tabular form.

## *4.4 Results of Treatment on Tax Preparation Method*

This analysis investigated the factors influencing tax preparation method choice among taxpayers, with a particular focus on potential differences between frequent filers and new/infrequent filers. We employed multinomial logistic regression to examine the effects of mailing a Free File letter or a checklist on the likelihood of choosing software filing, self-preparing on paper forms, using VITA, or filing with paid preparer assistance. Demographic factors such as age, location (urban vs. rural), and income (measured by AGI) can influence tax preparation behavior. Including these factors as control variables in the models helps account for these potential differences and strengthens the analysis.

### 4.4.1 Demographic Controls and Heterogeneity

Initially, we considered a single model encompassing all taxpayers. However, the data exhibited significant heterogeneity. Taxpayers can be categorized into distinct groups based on filing frequency (frequent vs. new/infrequent), and these two groups possess varying characteristics in terms of their age and income. To gain a more nuanced understanding, we conducted separate analyses for the two strata.

### 4.4.2 Free File Letter Effect on Tax Preparation Method

The analysis of tax preparation methods for frequent filers (See Table 14) reveals a positive treatment effect for the Free File letter. When compared to the control group who didn't receive the letter, frequent filers who received the Free File letter

were 1.438 times more likely to choose the Free File service over self-filing on paper forms. This suggests that the Free File letter had a significant impact on encouraging filers to utilize the free filing options available. Similar results can be found on new/infrequent filers (See Table 15). New/infrequent filers who received the Free File letter were 1.73 times more likely to choose the Free File over self-prepared with paper forms.

### 4.4.3 Demographic Variables

*Income*

The odds ratios for choosing software filing over paper filing are 1.067 (frequent filers) and 1.048 (infrequent filers) when AGI increases by $10,000. This is statistically significant (positive coefficient) and indicates a positive association. The odds ratios for choosing paid preparation over self-preparation on paper are 1.117 (frequent filers) and 1.073 (new/infrequent filers) when AGI increases by $10,000. This is statistically significant (positive coefficient) and indicates a positive association. It implies that when taxpayers with a higher AGI are more likely to choose software and paid preparer to prepare their income tax return.

*Income Tax Complexity*

The odds ratios for choosing software filing over self-prepared via paper forms for filing are 1.191 (frequent filers) and 1.117 (new/infrequent filers) when the income tax complexity score increases by one unit. This suggests a weak positive association (positive coefficient). The odds of new/infrequent filers choosing a paid preparer over self-prepared by paper increase by 1.635 times with each one-unit increase in the income-tax-complexity score, compared to a 1.378 time increase in odds for frequent filers. The new or infrequent filers are 1.153 times more likely to choose Free File over self-prepared by paper and 1.156 times more likely to choose VITA over self-prepared by paper when the income tax complexity score increases by one unit.

*Age Group*

Both results from the frequent and new/infrequent filers, as shown in Figures 3 and 4, indicate that as age increases, individuals are progressively less likely to use Free File over self-prepared by paper. Figure 3 illustrates the odds ratio for the four different tax preparation methods for frequent filers, while Figure 4 shows the same for new or infrequent filers. In both strata, each subsequent age group shows a 10-to-20 percent decrease in the odds of using Free File. Each subsequent age group shows a 10 to 20% decrease in the odds of using Free File. Furthermore, frequent filers in older age groups are increasingly likely to use VITA over self-prepared by paper. New/infrequent filers also show a preference for using VITA for age groups 30 and older. For individuals under 45 years old, there is no significant difference in the likelihood of using software or paid preparers compared to self-prepared paper filing. However, the likelihood of using software or paid preparers, compared to self-prepared by paper, decreases for age groups 45 and older and continues to decline as the age group increases. A similar pattern was also found with paid preparer versus self-prepared paper filing, where frequent filers under 45 show no significant preference between the two methods, while those 45 and older are less likely to use paid preparer. Taken together, younger filers are more likely to use Free File, and older filers are more likely to use VITA sites to prepare their returns or self-prepare on paper.

**FIGURE 3: The Odds Ratios Plots of Tax Preparation Method for Repeat Filers**

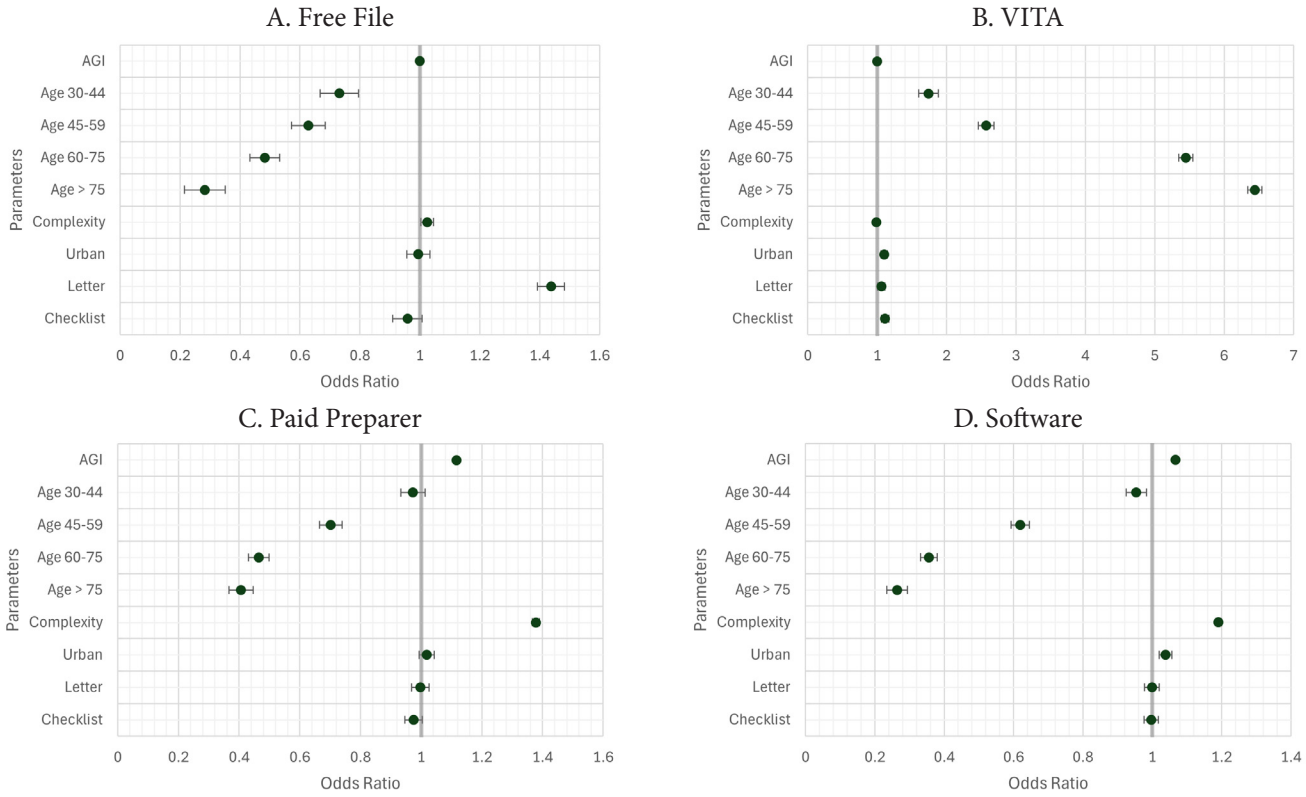
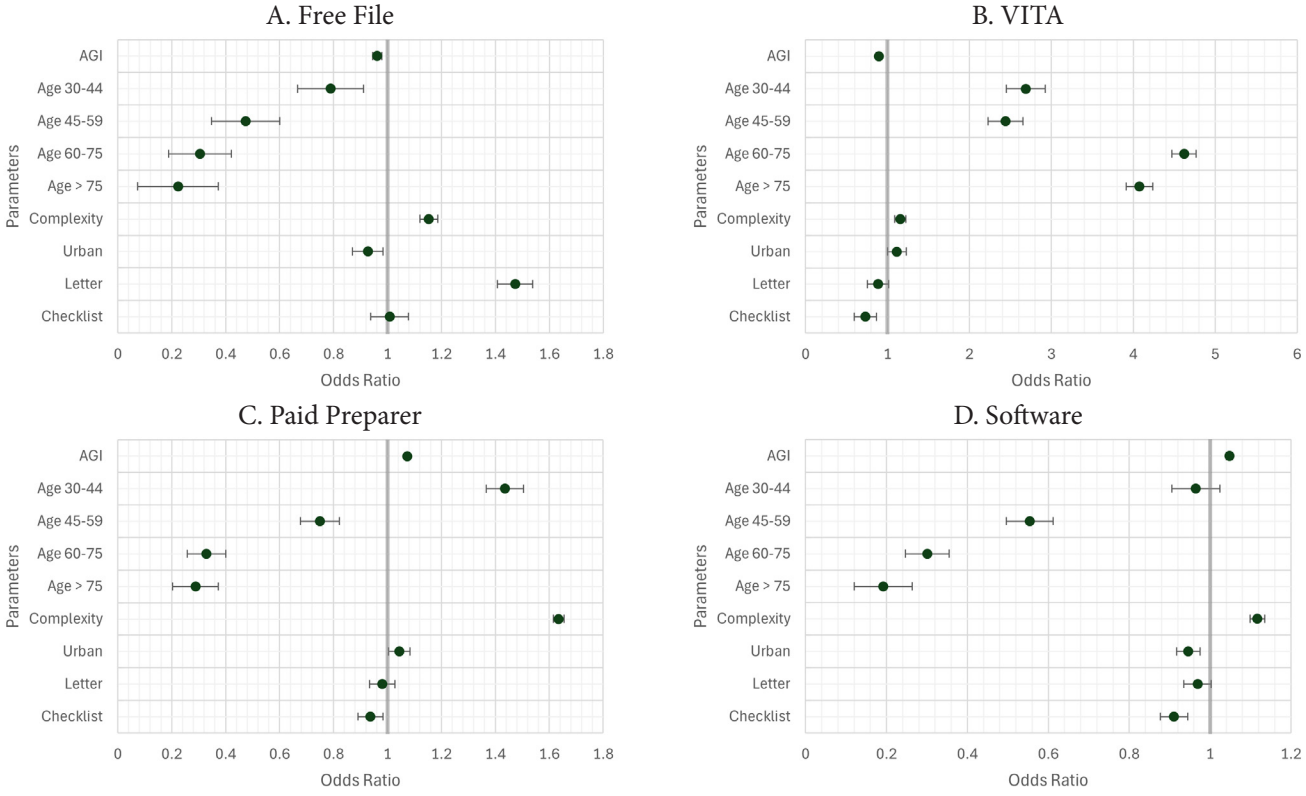



**FIGURE 4: The Odds Ratios Plots of Tax Preparation Method for New/Infrequent Filers**



## 5. Discussion and Conclusion

This study utilized binomial and multinomial logistic regression models to investigate the intervention effect of Free File letter and checklist, as well as other demographic factors influencing taxpayer behavior regarding filing and e-filing behaviors and tax preparation methods among frequent filers and new/infrequent filers.

### *5.1 Filing Versus Not Filing*

The chi-squared tests demonstrate that both the letter and checklist interventions significantly increased filing rates compared to the control group in both frequent and new or infrequent filers. Specifically, the filing rates increased by 8% among frequent filers and 12% among new or infrequent filers. However, the effect sizes of Cramer's V were small to moderate, suggesting that while the interventions were statistically significant, their practical impact was modest.

### *5.2 E-File vs. Paper File*

The logistic regression revealed that demographic variables, particularly income and age, play a significant role in influencing e-filing behavior among taxpayers. Higher income was positively associated with an increased likelihood of e-filing, suggesting that individuals with higher income have more financial resources to utilize digital filing methods. It is important to note that this study focuses on individuals with income capped at $73,000 in the prior tax year. This income threshold suggests that individuals with higher incomes, even below $73,000, may be more able to engage with e-filing, beyond the use of the Free File Program.

All age groups showed a significant negative impact on e-filing likelihood compared to the youngest group. As age increased, taxpayers were progressively less likely to e-file. The results are aligned with our generational difference assumption, and it reflects a broader trend where older individuals may not have been exposed to digital technology and may be less inclined to adopt e-filing. Younger generations, having grown up with the internet and technology, might feel more comfortable navigating websites and software interfaces as well as utilizing electronic filing methods compared to older individuals who might be less familiar with the technology. Trust in technology as a potential obstacle also warrants consideration, particularly for older adults. While the current study doesn't directly measure trust in technology, literature has shown that it plays an important role in people's adaption of e-banking (Suh and Han, 2002). We have a follow-up study underway, which has incorporated the issue of security more directly. The upcoming research modifies the e-file treatment letter to explicitly address taxpayer concerns about data security when using electronic filing methods.

In addition to demographic factors, income tax complexity had a significant negative effect on e-filing behavior across both strata. It indicates that as tax returns become more difficult to navigate, taxpayers may feel less confident in handling the process electronically. This shows that while e-filing is promoted for its ease and efficiency, taxpayers facing complex tax situations may not perceive e-filing as straightforward or user-friendly. A further investigation of the relationship of income tax complexity and how taxpayers prepare their tax returns is needed.

Given the strong effects of demographic and complexity covariates, it is perhaps not surprising that the letter did not have a significant main effect on e-filing rates. The results suggest that the influence of age, income, and complexity may be more powerful than the persuasive attempts of the intervention, overshadowing the effect of the letter on e-filing beyond the Free File Program. Two possible reasons may explain the letter's limited effectiveness outside of the Free File Program:

First, the letter may not have been persuasive enough to influence filing behavior, particularly when compared with the strong behavioral drivers associated with income, age, and complexity. If our goal is to drive greater e-filing more broadly, then these results suggest that redesigning the letter may be beneficial. For example, a redesigned letter could better address the concerns and motivations of different demographic groups by including more tailored content or clearer benefits of e-filing.

Second, the lack of significant impact could also have been attributed to the way the outcome variable was defined, including Free file, software, and e-filing submitted by a paid preparer. This broad definition may have diluted the specific impact of the letter on promoting self-e-filing through software or other electronic options like Free File.

## *5.3 Tax Preparation Methods*

### 5.3.1 Free File Letter Effect

The results provided compelling evidence for the effectiveness of the Free File letter in encouraging use of the program. Receiving the Free File letter yielded a statistically significant increase in the likelihood of choosing the Free File service over self-preparation on paper forms for both the frequent (odds ratio = 1.438) and new/infrequent filers (odds ratio = 1.473). This suggests that the Free File letter effectively encouraged these filers to utilize free electronic options. The Free File program may not be widely known among taxpayers, and the letter might have raised awareness of the program. By informing them about the availability and highlighting the potential benefits of free electronic filing services, such as convenience and claiming relevant tax credits, the program could have steered filers away from traditional paper-based methods.

### 5.3.2 Demographics

For both the frequent and new/infrequent filers, AGI had a positive association with choosing software filing over self-preparing on paper forms (odds ratio = 1.067 and 1.048) and paid preparer assistance over self-preparing on paper forms (odds ratio = 1.114 and 1.073). This suggests that taxpayers with higher income might be willing to pay extra for the convenience and professional tax support offered by software or paid preparer service.

The results suggest that tax complexity plays a role in increasing the likelihood of choosing software filing over self-preparing on paper forms for both strata. However, the odds of choosing paid preparer service over self-preparation on paper increased notably with rising complexity, with new or infrequent filers showing 1.635 times increase and frequent filers showing 1.378 times increase in odds. This implies that when their tax situation becomes more complex, taxpayers might feel less comfortable tackling it themselves. They might be more likely to seek professional assistance from paid preparers who can navigate the complexities and ensure accurate filing.

Age significantly affected tax preparation decisions, with both frequent and new or infrequent filers showing a clear trend: as age increased, individuals were progressively less likely to use Free File over self-preparing on paper. Older taxpayers were increasingly likely to use VITA over self-prepared on paper, perhaps reflecting a preference for in-person assistance. For individuals over 45, the likelihood of using software or a paid preparer declined, perhaps reflecting a reluctance to adopt newer filing options or to use a paid preparer when they can self-prepare on paper.

## *5.4 Implications and Limitations*

This study shows the importance of tailoring messaging on different preparation methods to address the diverse needs of different age groups, income level, and income tax complexities. The effectiveness of the Free File letter demonstrates that targeted communication can raise awareness and encourage taxpayers to utilize a program available to them. The IRS is also piloting the IRS Direct File platform, which has the potential to encourage more taxpayers to move to e-filing—this would be mutually beneficial for the taxpayer and the IRS. By building on this momentum, the IRS can further increase e-filing adoption rates and support modernization efforts. In addition, the IRS could consider developing a focused awareness campaign that emphasizes the benefits of electronic filing for taxpayers.

The study revealed interesting insights into how taxpayer demographics influence filing method choice. For older filers, outreach strategies could encourage the use of VITA centers, where they can receive personalized assistance in-person. Conversely, younger filers could be encouraged to use Free File services, emphasizing the convenience, cost, and ease of use. If frequent filers facing complex tax situations are more likely to feel comfortable with using tax preparation software, then emphasizing the convenience and efficiency of e-filing software automation might be more effective. By understanding the underlying reasons behind filing method choice, the IRS can design more tailored messaging that resonates with specific taxpayer groups and their needs.

### 5.4.1 Limitations

1. Sample Selection and Generalizability: The study focuses on taxpayers who were likely eligible to use Free File (e.g., had reported income under $73,000 in FY2021) and who are not habitual paper filers (i.e., had not paper-filed in each of the previous three years). Consequently, we do not know if a similar intervention would work on those with higher incomes. Likewise, we do not know if the findings would generalize to the broader paper filer population (i.e., those who are consistent paper filers). We are currently conducting a follow-up study that removes the income restriction to better understand a broader range of taxpayers and tests additional letter versions. This will allow for a more comprehensive understanding of how factors like income and filing complexity influence e-filing behavior.
2. Behavioral Response to Intervention: The interventions (Free File letter and tax filing checklist) might influence taxpayer behavior in ways that are not fully captured by the study. For example, some taxpayers might respond to the outreach efforts differently than others due to personal beliefs or attitudes toward government communications. These individual differences are addressed via randomization but may differentially impact results with any future letter modifications or changes to external factors.
3. Limited Control Over External Factors: There may be external factors influencing taxpayers' decisions to e-file or paper file that are not accounted for in the study, such as changes in tax laws, economic conditions, or personal life events. Again, randomization addresses those concerns within the context of this study, but this may impact generalizability under changing circumstances.
4. Temporal Limitations: As was previously implied, the study's timing and the specific years of data collection may limit its applicability in different tax years or under different economic conditions.
5. Free File Eligibility: We selected our sample based on prior year return information. Some members of our treated group likely became ineligible by the time of the study. We will explore treatment effects within this group in future analyses.

By considering these findings and limitations, the IRS and the broader tax administration community can gain valuable insights into how to encourage electronic filing adoption among different taxpayer groups. Future research can build upon this study by refining interventions and exploring long-term effects to increase overall e-filing rates.

## References

Bennett, S., Maton, K. and Kervin, L. (2008). The 'digital natives' debate: A critical review of the evidence. British Journal of Educational Technology, 39: 775-786. https://doi.org/10.1111/j.1467-8535.2007.00793.x.

Bhargava, S. and Manoli, D. (2015). Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. American Economic Review, 105(11), 3489–3529.

Herlache, A., Javaid, R., Roy, I., Turk, A., and Orlett, S. (2020). Enforcement vs. outreach: Impacts on tax filing compliance. Presented at the IRS-TPC Research Conference. Internal Revenue Service. Retrieved from https://www.irs.gov/pub/irs-prior/p1500--2020.pdf .

Internal Revenue Service. (2021). IRS Data Book 2020. Retrieved from https://www.irs.gov/statistics/p55b--2021.pdf.

Internal Revenue Service. (2022). IRS Data Book 2021. Retrieved from https://www.irs.gov/statistics/p55b--2022.pdf.

Internal Revenue Service. (2023a). IRS Data Book 2022. Retrieved from https://www.irs.gov/statistics/p55b--2023.pdf.

Internal Revenue Service. (2023b). IRS Report to Congress: Inflation Reduction Act §10301(1)(B) IRS-run Direct e-File Tax Return System (Publication 5788 (5-2023)) [Report]. Department of the Treasury. Retrieved from https://www.irs.gov/pub/irs-pdf/p5788.pdf .

Javaid, R., Schafer, B., Goldin, J., Homonoff, T., and Isen A. (2018). Can IRS move paper filers to free assisted tax preparation? 2018 IRS Research Bulletin. Retrieved from: https://www.irs.gov/pub/irs-soi/18resconjavaid.pdf.

Javaid, R., Schafer, B., Goldin, J., and Homonoff, T. (2020). Filing Season 2019 Outreach Experiment on Paper Filers and Nonfilers. Proceedings. Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association, 113, 1–30. https://www.jstor.org/stable/27143900.

John, P., and Blume, T. (2018). How best to nudge taxpayers? The impact of message simplification and descriptive social norms on payment rates in a central London local authority. Journal of Behavioral Public Administration, 1(1).

LoopMe. (2024). 2024 tax season consumer market research. LoopMe. Retrieved from: https://loopme.com/insights/loopme-consumer-snapshot-tax-season-insights/.

MITRE Corporation. (2008). Advancing E-file Study Phase 1 Report. MITRE. https://www.mitre.org/news-insights/publication/advancing-e-file-study-phase-1-report.

MITRE and YouGov. (2023). Taxpayer Filing Preference Survey. MITRE. Retrieved from https://www.mitre.org/sites/default/files/2023-05/PR-23-1221-MITRE-Taxpayer-Filing-Preference-Surveys.pdf.

Molnar, G., Savage, S. J., and Sicker, D. C. (2019). High-speed Internet access and housing values. Applied Economics, 51(55), 5923–5936. https://doi.org/10.1080/00036846.2019.1631443.

Orlett, S., Javaid, R., Koranda, V., Muzikir, M., and Turk, A. (2017). Impact of Filing Reminder Outreach on Voluntary Filing Compliance for Taxpayers with a Prior Filing Delinquency. IRS Research Bulletin. Retrieved from: https://www.irs.gov/statistics/soi-tax-stats-irs-research-conference on 07/01/2022.

Parker, K. (2023, May 22). How Pew Research Center will report on generations moving forward. Pew Research Center. https://www.pewresearch.org/short-reads/2023/05/22/how-pew-research-center-will-report-on-generations-moving-forward/.

Parsad, B., Jones, J., and Greene, B. (2005). Internet access in U.S. Public Schools and Classrooms: 1994-2003. E.D. Tab. NCES 2005-015. US Department of Education.

Perrin, A. and Duggan, M. (2015, June 26). Americans' Internet Access: 2000–2015. Pew Research Center. https://www.pewresearch.org/internet/2015/06/26/americans-internet-access-2000-2015/.

Pippin, S. E., and Tosun, M. S. (2014). Electronic tax filing in the United States: An analysis of possible success factors. Electronic Journal of e-Government, 12(1), 20–36.

Suh, B., and Han, I. (2002). Effect of trust on customer acceptance of Internet banking. Electronic Commerce research and applications, 1(3-4), 247–263.

Wang, Y. S. (2003). The adoption of electronic tax filing systems: an empirical study. Government Information Quarterly, 20(4), 333–352.

# Appendix A: Free File Letter (LTR 6171 – 01-2023 Revision)

Department of the Treasury
Internal Revenue Service
P.O. Box 75426
Washington, DC 20002

Date:

[Taxpayer Name and Address]

**Faster refund? ✓ Fewer errors? ✓ Free? ✓**
**Check your eligibility for IRS Free File today!**

**What you need to know**
There are many potential advantages to free online tax preparation:

- Free electronic filing of your federal tax return.
- Getting your refund faster.
- Access to free commercial software for federal and state returns.
- Less chance of making a mistake on your tax return or missing a tax benefit, like the Earned Income Tax Credit (EITC).

Read below for information about free IRS-sponsored programs.

**Free File program**
What is the Free File Program?

- Free File provides free commercial software to help prepare your return online.
- Most taxpayers qualify if they earned $73,000 or less in 2022.
- You will need only your 2021 tax return, 2022 tax documents, and a valid email address to begin.
- **For more information, visit www.irs.gov/FreeFile.**

**Other information**

- If you have questions about this letter, you can call 888-525-6797 (toll-free).
- You don't need to respond to this letter.

**Letter 6171 (Rev. 1-2023)**
Catalog Number 72138R

## Appendix B. Tax Filing Checklist (Pub 5732 – 12-2022 revision)

# Tax Filing Checklist

The checklist below will assist you in properly filing your federal income tax return and help you avoid costly penalties for filing incorrectly.

| # | Action | ✓ |
|---|---|---|
| 1. | **I used the correct filing status.**<br>• If you are married and living with your spouse, neither of you may file a Head of Household return.<br>• For help selecting the correct filing status, visit irs.gov/help/ita/what-is-my-filing-status. | |
| 2. | **I used my correct address.**<br>• The IRS must be able to contact you by mail if there is a question about your return. This would be the address where you live or regularly receive your mail. | |
| 3. | **I reported all of my income.**<br>• You must report all taxable income as well as tax-exempt interest.<br>*Note:* Generally, all income you receive is taxable, including income from bartering. Money and assets that you receive as a gift or inheritance are not taxable to you. | |
| 4. | **I claimed only the deductions to which I am entitled.**<br>• Be sure to claim all allowable expenses. Maintain records of those expenses for at least three years.<br>• If you are self-employed, see Publication 535 for information on expenses you may claim for your business. To view the publication, go to irs.gov/publications/p535. | |
| 5. | **I claimed only the credits to which I am entitled.**<br>• For more information about the Earned Income Tax Credit (EITC), visit irs.gov/eitc.<br>• For more information about the Child Tax Credit, visit irs.gov/ctc. | |
| 6. | **I signed my return.**<br>*Important:* If you are married filing a joint return, both spouses must sign the return. | |
| 7. | **I saved a copy of my return.**<br>You should keep a copy of your tax return for at least three years. | |
| 8. | I filed my return on _____ (enter date) | |

**Tips to remember when selecting a preparer:**

- **Ask about Service Fees.** Avoid preparers who base fees on a percentage of the refund or who boast bigger refunds than their competition. When asking about a preparer's services and fees, don't give them tax documents, Social Security numbers or other information before you decide to hire the preparer.
- **Make Sure the Preparer is Available.** Make sure your preparer will be available after your return is filed to ensure he or she will be available for follow-up should you need additional assistance.
- **Provide Records and Receipts.** Good preparers will ask to see a taxpayer's records and receipts. They'll ask questions to figure things like the total income, tax deductions, and credits.
- **Make sure that your refund goes directly to you** – not to the preparer's bank account. Review the routing and bank account number on the completed return.
- **Never Sign a Blank Return.** Don't use a tax preparer who asks you to sign a blank tax form.
- **Ensure the Preparer Signs and Includes Their PTIN.** By law, paid preparers must sign returns and include their Preparer Tax Identification Number (PTIN). You are also able to look a preparer up online by their PTIN to ensure that it is legitimate, https://irs.treasury.gov/rpo/rpo.jsf.

Publication 5732 (12-2022) Catalog Number 93688E Department of the Treasury Internal Revenue Service publish.no.irs.gov

# Appendix C. Additional Tables

**TABLE C.1. Summary of logistic regression analysis for variables predicting decision to electronically file for frequent filers, controlling for taxpayer characteristics.**

| Parameter | Estimate | SE | P-value | Odds Ratio |
|---|---|---|---|---|
| Intercept | 0.123 | 0.037 | 0.001 | |
| AGI | 0.077 | 0.003 | <.0001 | 1.08 |
| Age 30-45 | -0.099 | 0.045 | 0.030 | 0.906 |
| Age 45-60 | -0.510 | 0.043 | <.0001 | 0.601 |
| Age 60-75 | -0.775 | 0.039 | <.0001 | 0.461 |
| Age >75 | -0.933 | 0.047 | <.0001 | 0.393 |
| Complex | -0.088 | 0.015 | <.0001 | 0.916 |
| Urban | -0.058 | 0.029 | 0.048 | 0.944 |
| Letter | 0.001 | 0.050 | 0.990 | 1.001 |
| Letter × Age 30-44 | -0.060 | 0.063 | 0.339 | 0.944 |
| Letter × Age 45-59 | -0.004 | 0.059 | 0.940 | 1.001 |
| Letter × Age 60-74 | 0.089 | 0.053 | 0.097 | 1.001 |
| Letter × Age >75 | 0.033 | 0.064 | 0.610 | 1.033 |
| Letter × Urban | 0.000 | 0.040 | 0.997 | 1.000 |
| Letter × Complexity | 0.001 | 0.021 | 0.967 | 1.001 |
| Checklist | -0.091 | 0.050 | 0.070 | 0.913 |
| Checklist × Age 30-44 | -0.197 | 0.064 | 0.002 | 0.821 |
| Checklist × Age 45-59 | -0.013 | 0.059 | 0.828 | 0.987 |
| Checklist × Age 60-74 | 0.082 | 0.053 | 0.124 | 1.085 |
| Checklist × Age >75 | 0.113 | 0.064 | 0.078 | 1.120 |
| Checklist × Urban | -0.003 | 0.040 | 0.944 | 0.997 |
| Checklist × Complexity | 0.031 | 0.021 | 0.129 | 1.032 |
| **McFadden** | **0.029** | | | |
| **AIC** | **97,287** | | | |
| **Likelihood Ratio Test** | **$X^2$ (21) = 2,883.3, p < 0.0001** | | | |
| **N** | **75,890** | | | |

**TABLE C.2. Summary of logistic regression analysis for variables predicting decision to electronically file for new/infrequent filers, controlling for taxpayer characteristics**

| Parameter | Estimate | SE | P-value | Odds Ratio |
|---|---|---|---|---|
| Intercept | 0.032 | 0.049 | 0.518 | |
| AGI | 0.067 | 0.007 | <.0001 | 1.069 |
| Age 30-45 | 0.171 | 0.082 | 0.038 | 1.186 |
| Age 45-60 | -0.412 | 0.082 | <.0001 | 0.662 |
| Age 60-75 | -0.833 | 0.077 | <.0001 | 0.435 |
| Age >75 | -1.073 | 0.094 | <.0001 | 0.342 |
| Complex | -0.075 | 0.030 | 0.011 | 0.928 |
| Urban | 0.003 | 0.048 | 0.956 | 1.003 |
| Letter | -0.104 | 0.066 | 0.116 | 0.901 |
| Letter × Age 30-44 | -0.114 | 0.120 | 0.341 | 0.892 |
| Letter × Age 45-59 | -0.229 | 0.120 | 0.057 | 0.796 |
| Letter × Age 60-74 | -0.078 | 0.114 | 0.491 | 0.925 |
| Letter × Age >75 | -0.206 | 0.144 | 0.153 | 0.813 |
| Letter × Urban | -0.080 | 0.065 | 0.216 | 0.923 |
| Letter × Complexity | 0.106 | 0.042 | 0.011 | 1.112 |
| Checklist | -0.240 | 0.066 | 0.000 | 0.787 |
| Checklist × Age 30-44 | -0.130 | 0.122 | 0.288 | 0.878 |
| Checklist × Age 45-59 | 0.013 | 0.123 | 0.918 | 1.013 |
| Checklist × Age 60-74 | -0.148 | 0.114 | 0.194 | 0.863 |
| Checklist × Age >75 | -0.134 | 0.143 | 0.346 | 0.874 |
| Checklist × Urban | -0.060 | 0.065 | 0.353 | 0.942 |
| Checklist × Complexity | 0.122 | 0.042 | 0.004 | 1.130 |
| **McFadden** | **0.024** | | | |
| **AIC** | **39,862** | | | |
| **Likelihood Ratio Test** | **$X^2$ (21) = 998.4, p < 0.0001** | | | |
| N | **29,834** | | | |

**TABLE C.3. Multinominal logistic regression results for the tax preparation method: Frequent filers**

| Parameter | Estimate | Standard Error | p value | Odds Ratio |
|---|---|---|---|---|
| **Free vs. Self on Paper** | | | | |
| Intercept | -2.153 | 0.057 | <.0001 | |
| AGI | -0.002 | 0.008 | 0.813 | 1.000 |
| Age 30-44 | -0.314 | 0.064 | <.0001 | 0.731 |
| Age 45-59 | -0.465 | 0.056 | <.0001 | 0.628 |
| Age 60-75 | -0.729 | 0.050 | <.0001 | 0.483 |
| Age > 75 | -1.263 | 0.068 | <.0001 | 0.283 |
| Complexity | 0.025 | 0.021 | 0.233 | 1.025 |
| Urban | 0.005 | 0.039 | 0.896 | 0.995 |
| Letter | 0.363 | 0.045 | <.0001 | 1.438 |
| Checklist | -0.042 | 0.049 | 0.388 | 0.959 |
| **VITA vs. Self on Paper** | | | | |
| Intercept | -4.277 | 0.109 | <.0001 | |
| AGI | -0.020 | 0.009 | 0.032 | 0.998 |
| Age 30-44 | 0.553 | 0.139 | <.0001 | 1.739 |
| Age 45-59 | 0.943 | 0.115 | <.0001 | 2.569 |
| Age 60-75 | 1.695 | 0.100 | <.0001 | 5.447 |
| Age > 75 | 1.863 | 0.102 | <.0001 | 6.444 |
| Complexity | -0.012 | 0.027 | 0.657 | 0.988 |
| Urban | 0.095 | 0.045 | 0.037 | 1.099 |
| Letter | 0.056 | 0.055 | 0.305 | 1.058 |
| Checklist | 0.106 | 0.054 | 0.048 | 1.112 |
| **Paid Preparer vs. Self on Paper** | | | | |
| Intercept | -1.698 | 0.037 | <.0001 | |
| AGI | 0.111 | 0.004 | <.0001 | 1.117 |
| Age 30-44 | -0.027 | 0.040 | 0.499 | 0.973 |
| Age 45-59 | -0.353 | 0.037 | <.0001 | 0.702 |
| Age 60-75 | -0.767 | 0.034 | <.0001 | 0.465 |
| Age > 75 | -0.901 | 0.040 | <.0001 | 0.406 |
| Complexity | 0.321 | 0.012 | <.0001 | 1.378 |
| Urban | 0.017 | 0.025 | 0.485 | 1.018 |
| Letter | 0.003 | 0.029 | 0.919 | 0.997 |
| Checklist | -0.025 | 0.029 | 0.394 | 0.975 |
| **Software vs. Self on Paper** | | | | |
| Intercept | -0.332 | 0.027 | <.0001 | |
| AGI | 0.065 | 0.003 | <.0001 | 1.067 |
| Age 30-44 | -0.047 | 0.029 | 0.098 | 0.954 |
| Age 45-59 | -0.480 | 0.026 | <.0001 | 0.619 |
| Age 60-75 | -1.033 | 0.024 | <.0001 | 0.356 |
| Age > 75 | -1.330 | 0.030 | <.0001 | 0.264 |
| Complexity | 0.175 | 0.009 | <.0001 | 1.191 |
| Urban | 0.038 | 0.018 | 0.037 | 1.038 |
| Letter | 0.001 | 0.021 | 0.969 | 0.999 |
| Checklist | 0.003 | 0.021 | 0.877 | 0.997 |

**TABLE C.4. Multinominal logistic regression results for the tax preparation method: New or infrequent filers**

| Parameter | Estimate | Standard Error | p value | Odds Ratio |
|---|---|---|---|---|
| **Free vs. Self on Paper** | | | | |
| Intercept | -2.253 | 0.075 | <.0001 | |
| AGI | -0.038 | 0.017 | 0.027 | 0.962 |
| Age 30-44 | -0.237 | 0.122 | 0.053 | 0.789 |
| Age 45-59 | -0.747 | 0.127 | <.0001 | 0.474 |
| Age 60-75 | -1.188 | 0.116 | <.0001 | 0.305 |
| Age > 75 | -1.502 | 0.150 | <.0001 | 0.223 |
| Complexity | 0.142 | 0.033 | <.0001 | 1.153 |
| Urban | -0.076 | 0.057 | 0.184 | 0.927 |
| Letter | 0.387 | 0.066 | <.0001 | 1.473 |
| Checklist | 0.008 | 0.070 | 0.906 | 1.008 |
| **VITA vs. Self on Paper** | | | | |
| Intercept | -4.523 | 0.160 | <.0001 | |
| AGI | -0.115 | 0.032 | 0.0003 | 0.892 |
| Age 30-44 | 0.988 | 0.238 | <.0001 | 2.686 |
| Age 45-59 | 0.892 | 0.213 | <.0001 | 2.440 |
| Age 60-75 | 1.531 | 0.148 | <.0001 | 4.620 |
| Age > 75 | 1.405 | 0.164 | <.0001 | 4.075 |
| Complexity | 0.145 | 0.066 | 0.029 | 1.156 |
| Urban | 0.108 | 0.114 | 0.346 | 1.114 |
| Letter | -0.121 | 0.130 | 0.351 | 0.886 |
| Checklist | -0.315 | 0.135 | 0.019 | 0.730 |
| **Paid Preparer vs. Self on Paper** | | | | |
| Intercept | -1.934 | 0.051 | <.0001 | |
| AGI | 0.071 | 0.009 | <.0001 | 1.073 |
| Age 30-44 | 0.362 | 0.069 | <.0001 | 1.436 |
| Age 45-59 | -0.290 | 0.072 | <.0001 | 0.749 |
| Age 60-75 | -1.111 | 0.072 | <.0001 | 0.329 |
| Age > 75 | -1.244 | 0.085 | <.0001 | 0.288 |
| Complexity | 0.492 | 0.020 | <.0001 | 1.635 |
| Urban | 0.043 | 0.040 | 0.272 | 1.044 |
| Letter | -0.019 | 0.047 | 0.687 | 0.981 |
| Checklist | -0.065 | 0.046 | 0.164 | 0.937 |
| **Software vs. Self on Paper** | | | | |
| Intercept | -0.240 | 0.038 | <.0001 | |
| AGI | 0.047 | 0.008 | <.0001 | 1.048 |
| Age 30-44 | -0.036 | 0.059 | 0.541 | 0.965 |
| Age 45-59 | -0.590 | 0.058 | <.0001 | 0.554 |
| Age 60-75 | -1.201 | 0.054 | <.0001 | 0.301 |
| Age > 75 | -1.649 | 0.072 | <.0001 | 0.192 |
| Complexity | 0.111 | 0.018 | <.0001 | 1.117 |
| Urban | -0.056 | 0.029 | 0.056 | 0.946 |
| Letter | -0.031 | 0.034 | 0.356 | 0.969 |
| Checklist | -0.094 | 0.034 | 0.006 | 0.911 |

### TABLE C.5. Cross Tabulation of the Treatment Groups and the Tax Preparation Method in Strata 1 (Frequent Filers)

| Tax Preparation Method | Treatment Group | | |
|---|---|---|---|
| | Control Group | Checklist Letter | Letter |
| Free E-File | 866 | 940 | 1,365 |
| Paid Preparer | 2,910 | 3,108 | 3,094 |
| Paper | 12,049 | 13,600 | 13,160 |
| Software | 7,063 | 7,830 | 7,630 |
| VITA | 665 | 838 | 770 |
| Not Filed | 12,197 | 9,310 | 9,657 |
| Total | 35,750 | 35,626 | 35,676 |

Note: V-Coded returns have been removed from the table due to disclosure concerns.

### TABLE C.6. Cross Tabulation of the Treatment Groups and the Tax Preparation Method in Strata 2 (New/Infrequent Filers)

| Tax Preparation Method | Treatment Group | | |
|---|---|---|---|
| | Control Group | Checklist Letter | Letter |
| Free E-File | 393 | 552 | 765 |
| Paid Preparer | 1,171 | 1,355 | 1,358 |
| Paper | 4,120 | 5,140 | 4,855 |
| Software | 2,783 | 3,479 | 3,509 |
| VITA | 136 | 101 | 116 |
| Not Filed | 9,247 | 7,116 | 7,192 |
| Total | 17,850 | 17,743 | 17,795 |

Note: V-Coded returns have been removed from the table due to disclosure concerns.

# 5

# Appendix

## Conference Program

# 14th Annual IRS-TPC Joint Research Conference on Tax Administration
# June 13, 2024

## Program

8:30–8:45—Opening

*Robert McClelland (Senior Fellow, Urban-Brookings Tax Policy Center*
*Barry Johnson (Deputy Chief Data and Analytics Officer, Research, Applied Analytics and Statistics (IRS)*

9:30–11:00—Session 1: Harnessing data for better research

Moderator: *Brittany Jefferson (IRS, Wage and Investment)*

» Improving linkages to individual income tax data
*Amy O'Hara, Stephanie Strauss, Maanasa Vatsavayi, Nathan Wycoff (Georgetown University)*

» A large scale, high quality US occupational database: Results from merged IRS and ACS write-ins
*Victoria Bryant, Thomas Hertz, Kevin Pierce (IRS, RAAS); Julia Beckhusen, Lynda Laughlin, Liana Christin Landivar, Carl Sanders (US Census Bureau); Josh Gagne, David Grusky, Sofia Jamesson (Stanford University); Michael Hout (New York University); Ananda Martin-Caughey (Brown University); Javier Miranda (University of Jena)*

» Disaggregating tax compliance burden: A comparative study
*Bizuayehu Bedane (IRS, RAAS)*

Discussant: *Leonard Burman (Urban-Brookings Tax Policy Center)*

10:15–11:45 a.m.—Session 2: Discovering the art of avoidance

Moderator: *Devi McKalko (IRS, RAAS)*

» Using a gravity model to predict cross-border tax avoidance
*Lori Stuntz, Michael Udell (IRS, RAAS)*

» Art in the age of tax avoidance
*Matthew Pierson (The Wharton School, University of Pennsylvania)*

» Staying on the wagon: Estimating indirect deterrence effects from filing and payment compliance programs
*Brett Collins, Chris Wilson, Corbin Miller, Mark Payne, Sean Roh, Yan Sun, Alex Turk (IRS, RAAS)*

Discussant: *William Boning (U.S. Department of the Treasury)*

11:45 a.m.–12:15 p.m. —Break/Lunch

12:15–12:45 p.m.—Keynote speaker

*Danny Werfel (Commissioner of Internal Revenue)*
*Interview with Tracy Gordon (Co-Director, Urban-Brookings Tax Policy Center)*

12:45–1 p.m.—Break

1–2:30 p.m.—Session 3: Trusting the tax man: Metrics, AI, and auditss

Moderator: *Melissa Vigil (IRS, RAAS)*

» Measuring success: New performance metrics for a new Internal Revenue Service
*Janet Holtzblatt (Urban-Brookings Tax Policy Center)*

» Tools to promote trustworthiness in a prototype AI system at the IRS
*Michael Szulczewski, M. Feldman, Steffani Silva (MITRE); Brandon Anderson, Alissa Graff (IRS, RAAS)*

Discussant: *Arnstein Øvrum (Norwegian Tax Administration)*

2:30–4 p.m.—Session 4: Simplifying the filing burden

Moderator: *John Guyton (IRS, RAAS)*

» Technical challenges in maintaining tax prep software with large language models
*Sina Gogani-Khiabani, Varsha Dewangan, Ashutosh Trivedi (CU Boulder); Nina Olson (Center for Taxpayer Rights); Saeid Tizpaz-Niari (UT El Paso)*

» Rethinking tax information: The case for quarterly 1099s
*Kathleen DeLaney Thomas (UNC School of Law)*

» Investigating the impact of Free e-File letter intervention on taxpayer's tax filing and preparation methods
*Pei-Hua Chen, Astin Cornwall, Anne D. Herlache, Scott Leary, Brenda Schafer, Melissa Vigil (IRS, RAAS); Rizwan Javaid (IRS Taxpayer Experience Office)*

Discussant: *Robert Weinberger (Urban-Brookings Tax Policy Center)*

4-4:05 p.m. —Wrap-up

Barry Johnson *(Deputy Chief Data and Analytics Officer, Research, Applied Analytics, and Statistics (IRS))*