The IRS Research Bulletin

Proceedings of the 2019 IRS / TPC Research Conference

Research, Applied Analytics & Statistics

Publication 1500 (Rev. 6-2020) Catalog Number 11546J Department of the Treasury **Internal Revenue Service** www.irs.gov

Papers given at the

9th Annual Joint Research Conference on Tax Administration

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

Held at the Urban Institute
Washington, DC
June 20, 2019

Compiled and edited by Alan Plumley*
Research, Applied Analytics, and Statistics, Internal Revenue Service

*Prepared under the direction of Barry W. Johnson, Acting Chief Research and Analytics Officer

Foreword

This edition of the *IRS Research Bulletin* (Publication 1500) features selected papers from the IRS-Tax Policy Center (TPC) Research Conference held at the Urban Institute in Washington, DC, on June 20, 2019. Conference presenters and attendees included researchers from many areas of the IRS, officials from other Government agencies, and academic and private sector experts on tax policy, tax administration, and tax compliance. In addition to those who attended in person, many participated live online, as the TPC broadcast video of the proceedings over the Internet. The videos are archived on their Website to enable additional participation. Online viewers participated in the discussions by submitting questions via e-mail as the sessions proceeded.

The conference began with welcoming remarks by Eric Toder, Co-Director of the Tax Policy Center, and by Barry Johnson, the Acting IRS Chief Research and Analytics Officer, who introduced a short video welcome from IRS Commissioner Charles Rettig. The remainder of the conference included sessions on estimating the effects of tax administration on compliance, the influence of external factors on compliance, improving the digital taxpayer experience, and understanding the drivers of taxpayer behavior. The keynote speaker was Richard Rubin, U.S. Tax Policy Reporter at the *Wall Street Journal,* who offered his insights on current tax issues.

We trust that this volume will enable IRS executives, managers, employees, stakeholders, and tax administrators elsewhere to stay abreast of the latest trends and research findings affecting tax administration. We anticipate that the research featured here will stimulate improved tax administration, additional helpful research, and even greater cooperation among tax administration researchers worldwide.

Acknowledgments

This IRS-TPC Research Conference was the result of preparation over a number of months by many people. The conference program was assembled by a committee representing research organizations throughout the IRS. Members of the program committee included: Alan Plumley, Ken Kaufman, José Colon De La Matta, and Michael Sebastiani (RAAS); Scott Rutz (Taxpayer Advocate); Derek Haynes (Human Capital Office Research); John Miller (Large Business & International Division); Diane Belanger (Wage & Investment Division); Debbie Schmidt (Small Business / Self-Employed Division); Andrew Kitchings (Tax Exempt, Government Entities Division); Stephanie Needham (Criminal Investigation Division); George Contos (Communications & Liaison); Alcora Walden (Online Services); and Rob McClelland (Tax Policy Center). In addition, Ann Cleven and Hailey Roemer from the Tax Policy Center and Jon Creem (RAAS) oversaw numerous details to ensure that the conference ran smoothly.

This volume was prepared by Clay Moulton, Camille Swick, and Lisa Smith (layout and graphics) and Beth Kilss and Georgette Walsh (editors), all of the IRS Statistics of Income Division. 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.

Barry Johnson
Acting IRS Chief Research and Analytics Officer

9th Annual IRS-TPC Joint Research Conference on Tax Administration

Contents

Foreword iii

1. Estimating the Effects of Tax Administration on Compliance

- Estimating the Specific Indirect Effect for Multiple Types of Correspondence Audits
  Ben Howard, Lucia Lykke, and Leigh Nicholl (MITRE Corporation), and Alan Plumley (IRS, RAAS) 3
- Enforcement vs. Outreach—Impacts on Tax Filing Compliance
  Anne Herlache, Ishani Roy, Alex Turk (IRS, RAAS), and Stacy Orlett (IRS, SB/SE) 30
- Exchange of Information and Bank Deposits in International Financial Centres
  Pierce O'Reilly and Michael A. Stemmer (OECD Centre for Tax Policy and Administration) and Kevin Parra Ramirez (Banque de France) 72

2. The Influence of External Factors on Compliance

- Recent Changes in the Paid Return Preparer Industry and EITC Compliance
  Emily Y. Lin (Office of Tax Analysis, U.S. Department of the Treasury) 107
- Effect of Recent Reductions in the Internal Revenue Service's Appropriations on Returns on Investment
  Janet Holtzblatt (Urban-Brookings Tax Policy Center) and Jamie McGuire (Joint Committee on Taxation) 128

3. Improving the Digital Taxpayer Experience

- IRS Online Account User Testing: Improving the User Experience Through Iterative Design and Research
  Heather Gay (Mediabarn Inc.) 147
- Usability of Biometric Authentication Methods for Citizens with Disabilities
  Ronna ten Brink and Becca Scollan (The MITRE Corporation) 154
- Customer Experience Research Leads to Better Design and Increased Adoption
  Nikki Kerber, Kristen Papa, and Jacob Sauser (Booz Allen Hamilton) 178

4. Understanding the Drivers of Taxpayer Behavior

- Underpayment of Estimated Tax: Understanding the Penalized Taxpayer Population
  Victoria Bryant, Brett Collins, Janet Li, Alicia Miller, Alex Turk, and Tomás Wind (IRS, Research, Applied Analytics, and Statistics), and Stacy Orlett (IRS, Small Business/Self-Employed Division) 193

- The Positive and Negative Effects of Burdensome Audits
  Amy Hageman (Kansas State University), Ethan LaMothe (Oklahoma State University), and Mary Marshall (Louisiana Tech University) 220

- Using a Graph Database To Analyze the IRS Databank
  Ririko Horvath and Rahul Tikekar (IRS, Research, Applied Analytics, and Statistics) 223

5. Appendix

- Conference Program 233

1

Estimating the Effects of Tax Administration on Compliance

Howard ◆ Lykke ◆ Nicholl ◆ Plumley

Herlache ◆ Roy ◆ Turk ◆ Orlett

O'Reilly ◆ Stemmer ◆ Ramirez

Estimating the Specific Indirect Effect for Multiple Types of Correspondence Audit

Ben Howard, Lucia Lykke, and Leigh Nicholl (The MITRE Corporation), and Alan Plumley (IRS, RAAS)

Introduction

Tax enforcement actions have a direct revenue effect: the tax collected from (or refunded to) the contacted taxpayer pertaining to the year that was the subject of the contact. These enforcement actions undoubtedly also have an indirect effect on revenues: a change in the current or future behavior of taxpayers who either have experienced an enforcement contact themselves (the "specific" indirect effect) or have some knowledge or perception about others' tax enforcement experiences (the "general" indirect effect). This study seeks to estimate and compare the magnitudes of the specific effects on taxpayers following one of three types of correspondence audits, using longitudinal taxpayer data obtained by the United States Internal Revenue Service (IRS) through operational audits conducted on tax returns filed within the Tax Years (TYs) 2006 through 2012 period. We study the effects of correspondence audits that examine three different types of taxpayers: those who were audited based on business expenses, itemized deductions, and self-employment tax, respectively. In each case, we compare the subsequent-year reporting on total tax and audit-specific line items between the audited group and a not-audited taxpayer "control" group that was otherwise eligible for the audit according to all operational eligibility criteria. In this way, we advance prior literature by estimating the specific indirect effect for three categories of audit, examining taxpayers who are designated for audit eligibility by operational criteria that vary by the audit category.

Comparing the subsequent reporting behavior of taxpayers who experienced different types of audits has both research and operational value. Much of the prior literature on the indirect effect of audit has focused on taxpayers who are self-employed (e.g., Beer *et al.* (2015); DeBacker *et al.* (2018a)), finding that these taxpayers increase reporting on measures such as taxable income following an audit, and the effect is more pronounced compared to taxpayers whose income is primarily subject to third-party reporting (DeBacker *et al.* (2018b); Kleven *et al.* (2011)). The key point from these studies is that taxpayers—including audited taxpayers—are not a homogenous group, and therefore have different underlying characteristics and may also respond differently to different types of audits. However, little exploration has been done among different types of taxpayers using operational data and selection criteria. Further, the selection mechanism that drives whether or not a taxpayer is audited varies among types of audit.

Additionally, knowing whether and how different types of audits yield different specific indirect effects may help to inform IRS resource allocation decisions. As the Government Accountability Office (GAO) recently pointed out (GAO (2012)), different types of audits yield different direct revenue benefit/cost. The GAO called for greater knowledge of the indirect effect of different audit types in order to understand whether increasing or decreasing audit coverage in various audit categories would result in a long-term increase or decrease in the overall revenue generated from taxpayers' voluntary reporting compliance. For these reasons, we explore the differential specific indirect effects of three different types of correspondence audits.

However, empirically observing these indirect effects is challenging. Operational audits, unlike research audits such as those conducted under the National Research Program (NRP), are not randomly distributed among the taxpayer population. This fact poses major challenges for causal inference (Kleven *et al.* (2011); Mazzolini *et al.* (2017)). Although we are unable to completely account for such endogeneity in this study, we advance existing knowledge by specifically controlling for the specific operational metrics applied to each return to determine eligibility for audit within each category *and* the priority given to it among all eligible returns in that category. That means that our control group was not drawn from the overall population of

unaudited returns, but only from the much smaller subpopulation of returns that met all operational eligibility criteria. And, being able to apply the specific method used operationally in each category to prioritize the returns in the eligible pool, we further controlled for this ranking, representing an advance over prior studies that have not had access to such information.

As such, the following are general research objectives that guide this study:

1. Assess whether there is an observable change in taxpayers' individual contributions to IRS revenue, as defined by total tax reporting, in the years subsequent to experiencing a correspondence audit. We do this by comparing audited taxpayers' post-audit tax reporting to the tax reporting of unaudited taxpayers who were eligible during the same tax year.
2. Assess whether reporting on other relevant items, specifically on the line items being examined in the different types of correspondence audits, changes in the years after a taxpayer experiences an audit when compared to the reporting of similarly eligible, but ultimately unaudited, taxpayers.
3. Explore potential differences in post-audit reporting behavior across three distinct categories of correspondence audits, each of which is associated with a different underlying population of U.S. taxpayers subject to different audit selection criteria.

Literature Review

Types of Indirect Effect

Much of the literature and research conducted on taxpayer compliance behavior has rested on the assumption that tax agencies' enforcement activities—particularly audits—encourage tax compliance by deterring tax evasion or, conversely, by assuring that the tax system is fair and just. Tax evasion may take the form of not filing or misreporting income or other information (such as deductions) on tax returns, and compliance refers to the behaviors of filing tax returns on time, accurately reporting information on tax returns, and paying taxes owed on time (Hallsworth (2014)). Much research has been done to test whether and how a taxpayer's experience of enforcement threat or activity (e.g., a visit from an IRS officer, an audit) will affect that taxpayer's future probability of compliance, an effect referred to as "specific deterrence" (Slemrod (2016)) or, more generally, as the specific indirect effect. Although an audit may result in immediate funds collected from a noncompliant taxpayer (a direct effect of the audit), that taxpayer will likely pay taxes for many years to come and therefore the audit may continue to affect taxes paid by that taxpayer in subsequent years. This specific indirect effect is the focus of this study.

Additionally, taxpayers may also have awareness of enforcement activities experienced by other taxpayers that affects their perception of the risk of noncompliance. This secondhand effect of enforcement activities is known as the "general indirect effect"[1] (Plumley (1996); see also Slemrod (2016)). Several studies have found that audit rates at the aggregate level are positively associated with greater tax compliance (Ali *et al.* (2001); Dubin *et al.* (1990); Plumley (1996)). However, field experiments have resulted in mixed findings. There is some evidence that information about the threat of audit for businesses travels through tax preparer networks and corporate relationships (i.e., between parent and subsidiary companies) (Boning *et al.* (2018)). However, studies of the general indirect effect within neighborhoods have not found evidence that neighbors' tax-related experiences spill over to each other (Meiselman (2018)). A full review of evidence for general indirect effect is outside the scope of this study.

Evidence for Specific Indirect Effect

Compliance is, in most cases, impossible to observe because in the absence of a repeat audit, it is difficult to know whether a taxpayer's reporting was accurate. This may be especially true for taxpayers who report self-employment income that is not subject to third-party reporting. As such, most studies of the specific effect examine trends in reporting proxy measures, including income, tax liability, or specific deductions or

[1] Other terms for these types of indirect effects: general indirect effect is also referred to as general deterrence or the "spillover effect," and specific indirect effect is also referred to as a "dynamic effect" of audits (Advani *et al.* (2015)).

adjustments. Several themes from this research are relevant to this study: (1) the use of operational versus research audits; (2) the observation of specific effects among the self-employed; and (3) the attenuation of specific indirect effects over time.

A major challenge for the study of indirect effects of enforcement activities is the fact that taxpayers are not usually selected randomly into the "treatment" of being audited. Several countries, including the U.S., conduct randomly assigned research audits, which might be used to circumvent this selection bias problem; however, if taxpayers know that they are audited randomly for research purposes, this may introduce a validity issue insofar as taxpayers may respond to a random audit differently from an operational audit (Slemrod (2016)).

Specific Indirect Effect Among the Self-Employed

Several studies using research program data from the U.S. and other countries suggest evidence for the specific effect on subsequent income reporting, with the strongest effect among the self-employed. In the U.S., a study using NRP data from randomly assigned audits as a "treatment" group along with general taxpayer return information as a "control" group found that being audited increases reported wage income the following year by 1.3 percent, on average, and increases reported Schedule C income by 14.2 percent. This effect begins to diminish 3 years after being audited and mostly disappears after 4 years (DeBacker *et al.* (2018a)).

Further, random audit data from a Danish program has shown that being randomly audited was associated with an increase in income reported the following year, and this increase was largely driven by the self-employed. The results of this study suggest that the self-employed are most likely to be noncompliant but also show the strongest adjustment in reporting 1 year[2] after an audit (Kleven *et al.* (2011)). Confirming the conclusion about the importance of third-party reporting for an indirect effect,[3] U.K. taxpayers audited at random increased reported tax liability substantially over a 4-year period for taxpayers who filed self-assessed[4] income tax returns, which includes individuals with self-employment income and landlords, among others (Advani *et al.* (2015)). Overall, the finding that self-employed taxpayers are more sensitive to the indirect effect of audit for subsequent year reporting suggests that the underlying characteristics of the taxpayer and the return itself are key for understanding how indirect effects work.

Two recent IRS studies examined the impact of audits on future compliance among the self-employed, using operational audit data. In both, audits were not randomly assigned, but rather happened as part of standard operational procedures. Focusing on sole proprietorship compliance, one study found that among taxpayers who all had high DIF scores, audited taxpayers saw decreases in their DIF scores (indicating increased compliance) over the following 5 years compared to not-audited taxpayers; this effect disappeared by the fifth year after audit (Nestor and Beers (2014)). In a second study, researchers used propensity score matching techniques to conduct a quasi-experiment. They found that being audited increased reported Schedule C net profit and taxable income of taxpayers whose previous audits resulted in additional tax liability assessments,[5] and this effect persisted over the next 3 years. Conversely, taxpayers who were audited previously but the audit did not result in a change in tax liability saw a decline in compliance 3 years after audit (Beer *et al.* (2015)).

Specific Indirect Effect Among Other Populations

In this study, we build on prior work that suggests that variations in population characteristics, and also the nature of dissimilar categories or types of audits, may be differently associated with subsequent-year reporting. In addition to the self-employment-focused studies above, a few studies have investigated the specific indirect effect of audits on other populations, such as those taxpayers who report capital gains and losses, list

[2] Unlike in the U.S., the Danish audit schedule completes audits in 1 year (U.S. audits can take anywhere from 1 to 3 years after the taxpayer has filed to initiate, and about another year or so after initiation to close). Kleven *et al.* (2011) therefore observed income reporting only 1 year after the audit. They did not test for attenuation in audit effects over time.

[3] This is because, as noted by many researchers, the lack of third-party reporting means that self-employed taxpayers have more room to be noncompliant, since there is no way to cross-reference the information on their returns (DeBacker *et al.* (2018a); Erard and Ho (2003); Kleven *et al.* (2011); also discussed in Slemrod (2016)).

[4] In the UK, not all taxpayers have to submit self-assessed tax returns. Those who do need to submit them tend to be individuals with income from self-employment, people with very high incomes, landlords, and people collecting pension income (Advani *et al.* (2015)).

[5] Beer *et al.* (2015) used the outcome of the audit as a proxy for whether the taxpayer was assessed as being compliant or noncompliant. That is, if the audit recommended additional tax assessments, the taxpayer was noncompliant (did not report enough tax liability); if the audit resulted in no recommended change, the taxpayer was compliant (reported appropriate tax liability).

supplemental income, itemize deductions, or claim the Earned Income Tax Credit (EITC) on their returns. Among taxpayers audited randomly in the U.S., there is evidence that Schedule A itemized deductions, adjustments to income, Schedule C income, and Schedule E income are all sensitive to a research audit; in all four cases, taxpayers report more income and fewer deductions after the audit and the effect was persistent for up to 6 years. Conversely, no evidence was found of Schedule D income changing in response to a research audit. Two studies have shown that Earned Income Tax Credit claiming decreases after experiencing an audit; after a random NRP audit, taxpayers who claimed EITC decrease their future EITC claiming (DeBacker *et al.* (2018b)), and taxpayers who were audited operationally for EITC credit validity also reduce EITC claiming in subsequent years, especially within the first year after audit (Guyton *et al.* (2018)).

Other studies have attempted to characterize the specific indirect effect that enforcement actions taken by the IRS have on filing compliance. Given that nonfiling is challenging to observe, research surrounding this topic tends to consider only known nonfilers. Specifically, this includes those who did not appear on a tax return but had income reported to the IRS by a third party, usually through Form W-2, Form 1099-R, or other documents (Datta *et al.* (2015); Guyton *et al.* (2017)). One such study considered the effect of the Automatic Substitute for Return (ASFR) process, a function of the IRS applied to eligible nonfilers who have not responded to prior notices by filing a return. In the study, researchers found evidence of increased timely filing compliance up to 4 years after the ASFR treatment, with the effect decreasing each year (Datta *et al.* (2015)).

Other Evidence for Specific Indirect Effects

Two additional areas of research provide further evidence for specific indirect effects: laboratory experiments that attempt to replicate the condition of being audited in an artificial setting, and field experiments using enforcement "contacts" as a proxy for audits.

Laboratory experiments have found results that do not correspond to results observed in natural settings. For example, two studies using university students found evidence of a "bomb crater" effect of compliance, in which compliance decreases immediately after an audit, then increases (Kastlunger *et al.* (2009); Maciejovsky *et al.* (2007)).

Enforcement contacts, typically in the form of letters or tax official visits, have been used to study specific indirect effects as well. Studies show that deterrence messages designed to make the threat of audit or other enforcement activity result in an increase in compliance, both immediately and over a period of several years after the fact. In a natural field experiment conducted in Minnesota, a random sample of taxpayers who received a letter alerting them that their returns would be "closely monitored" showed increased payments compared to a control group (Slemrod *et al.* (2001)).

Similar effects have been observed among nonfilers in the U.S. In Detroit, tax "ghosts" (i.e., nonfilers) who received a letter explaining noncompliance penalties were more likely to file back-year returns, remit payments, and report greater tax liability compared to nonfilers who received letters with no penalty message or with nondeterrence messages about civic pride (Meiselman (2018)). Similarly, tax delinquents in three U.S. States were 7 percent more likely to submit payments within 10 weeks after receiving a letter indicating their State's financial penalties for noncompliance compared to a control group (Perez-Truglia and Troiano (2015)).

Research Questions and Hypotheses

In this study, we address the following research questions. For research questions 1 and 2, we separately answer each question for each of the three types of correspondence audit analyzed.

Total Tax Reporting

1. How does tax reported by taxpayers who were audited on any of their returns for Tax Years 2006 through 2012 vary over time after audit compared to the tax reporting of taxpayers who were eligible for the same type of audit, but were not audited?

a. Hypothesis 1 (H1): We hypothesize that the indirect effect of the audit will have an association with tax reporting, measured in comparison to the reporting of eligible unaudited taxpayers, 3 to 5 years after the audit and the effect will subsequently attenuate.

Reporting on Audit-Specific Line Items

2. Is there evidence that audit-specific line item reporting by taxpayers who were audited changes over time compared to the reporting of similar taxpayers who were eligible for the same type of audit, but did not experience an audit?

   b. Hypothesis 2 (H2): We hypothesize that the indirect effect of the audit will have an association with specific line item reporting, measured in comparison to the reporting of eligible unaudited taxpayers, 3 to 5 years after the audit and the effect will subsequently attenuate. The audit-specific line items are some Schedule C line items, some Schedule A line items, and some Schedule SE line items, respectively, for the three audit categories.

Exploratory Comparison Among Audit Categories

For research questions 3 and 4, statistical testing among categories of audits is not conducted because the underlying populations are different. Our questions are therefore framed as exploratory and we do not present hypotheses.

1. Do the empirical results from our analyses of different categories of audits suggest that the effect of an audit on subsequent tax reporting varies by the category of audit experienced?
2. Do the empirical results from our analyses of different categories of audits suggest that the effect of an audit on subsequent specific line item reporting varies by the category of audit conducted?

Data and Methods

Categories of Correspondence Audits

Correspondence audits, unlike field audits, are conducted via mail and are designed to examine a small set of line items or issues on a taxpayer's return. As such, correspondence audits focus on narrowly defined candidate populations as being "eligible" for a category of correspondence audit.

In this study, we separately compare eligible/not audited and audited taxpayers for three distinct categories of correspondence audit. We selected categories of correspondence audit that were active for the full study period (Tax Years 2006–2012), for which we have access to operational eligibility and selection criteria, and for which there was a sufficient volume of audits each year.[6] We control for potential confounding factors by limiting our analysis population only to taxpayers who were part of the candidate population for a given correspondence audit category, as defined by IRS operational procedures. Due to data sensitivity, we cannot further elaborate on the creation of the eligible population.

Audit Category 1: Examines some Schedule C expenses among taxpayers who filed a Schedule C (to report nonfarm business income) and met other category-specific eligibility criteria.

Audit Category 2: Examines some Schedule A deductions among taxpayers who itemized deductions and met other category-specific eligibility criteria.

Audit Category 3: Examines Schedule SE self-employment tax among taxpayers who met certain category-specific eligibility criteria.

Additionally, in order to select which returns to audit from the overall candidate population, examiners for each type of audit rely on different prioritization metrics, typically characteristics of the return. These prioritization metrics are specific to the audit category and cannot be further explained here due to data sen-

[6] We define sufficient volume arbitrarily as having roughly 1,000 cases each tax year.

sitivity. We treat these prioritization criteria as control variables. As such, we exploit knowledge of operational criteria to help account for confounding factors that inform audit selection.[7]

Data

In this study, we combine data on the three types of correspondence audits described above with return information on the general taxpayer population in the U.S. that met operational eligibility criteria for each type of audit. We use tax return and audit record data for primary Taxpayer Identification Numbers from the IRS's Compliance Data Warehouse (CDW) for Tax Years 2006–2018. In our analyses, we define the "baseline" year as the tax year a given taxpayer entered the sample, either because that taxpayer had an audited return for that tax year, or because they fell into the sampled eligible-not-audited group for that audit type for that tax year. In cases where a taxpayer entered the analytical sample multiple times (due to being eligible for the category of audit for multiple years and/or due to being audited multiple years), we handled these taxpayers as follows:

1. For any taxpayers that our queries returned multiple times because they were captured as "eligible" multiple times and were not audited in Tax Years 2006–2012: we declare the most recent eligibility year as the "baseline" year.
2. For any taxpayers that our queries returned multiple times because they were audited multiple times under the same audit category: we declare the first audit record as the "baseline" year.
3. For any taxpayers that our queries returned as being eligible in one or more years and audited in one or more years: we declare the earliest (or only) audited record as the "baseline" year and consider them solely in the "audited" group.

Audit ("Treatment") Group

To define the audited group, all primary taxpayer identification numbers associated with one of the three types of audits for any tax year in the 2006–2012 period in the Enforcement Revenue Information System (ERIS) database were identified and retained. For these audited group taxpayers, we collected tax return information from the Form 1040, Schedule A, Schedule C, and Schedule SE for the tax year of the baseline year and up to 8 tax years after (up to TY2018). For example, for baseline year 2006, we compiled return data up through Tax Year 2014; for baseline year 2012, we compiled return data up to Tax Year 2018. We chose to examine 8 years after the baseline based on prior literature, which suggests that an indirect effect is present from 3 to 5 years after audit; this allows for a buffer window at the end to ensure any possible attenuation in effect can be captured.

Eligible, Not Audited ("Control") Group

To define the eligible, not audited group, we applied undisclosed operational filter criteria to return records from the full universe of nonaudited taxpayers available in CDW. We restricted the returned records to a random sample of 25,000 taxpayers from the eligible population in each of Tax Years 2006–2012, as this returned a sufficient sample size for our analysis based upon the known sizes of the audited or "treatment" group. In some tax years, there are fewer than 25,000 eligible taxpayers—in this case we selected all eligible taxpayers regardless of the population size. For these eligible group taxpayers, we collected tax return information from the Form 1040, Schedule A, Schedule C, and Schedule SE for the tax year of the baseline year and up to 8 tax years after (up to TY2018).

Dependent Variables

Total Tax. Our primary dependent variable is total tax as reported on Form 1040. Total tax is chosen as the dependent variable across audit categories, as the change in tax paid over time most closely represents the "return on investment" that the IRS reaps from any observable specific indirect effect that results from the audit. Total tax, along with all other variables measured in dollars, are all adjusted for inflation to 2018 U.S. Dollars

[7] We have access only to IRS operational documents from the most recent 1 to 3 tax ears. As such, we assume that operational criteria stayed relatively stable over time for each correspondence audit type. We cannot know for sure if this assumption is correct.

(USD).[8] Because total tax is strongly right skewed, we fit our analysis models using the natural logarithm of total tax plus one dollar to account for cases where the taxpayer has reported zero total tax. The one dollar is added before taking the natural logarithm. If an indirect effect is present, we would expect total tax reporting to increase.

Audit Category 1 Schedule C Items. In our secondary analyses of line items that may have an association with an indirect effect of an audit, we treat some Schedule C line items as the dependent variable for audit category 1 models. We sum these undisclosed line items to create one continuous quantity. Because the sum of these line items is again strongly right skewed, we use the natural logarithm of this sum plus one dollar. Note that in this case, increased reporting of these Schedule C line items should have the effect of decreasing overall tax liability; thus, we would expect a positive indirect effect to be associated with decreased reporting on these line items.

Audit Category 2 Schedule A Items. We next treat some Schedule A line items as the dependent variable for audit category 2 models. We sum these line items to create one continuous quantity. Because the sum of these line items is strongly right-skewed, we use the natural log of this sum plus one dollar. Note that in this case, increasing reporting of these Schedule A deductions should have the effect of decreasing overall tax liability; thus, we would expect a positive indirect effect to be associated with decreased reporting on these line items.

Audit Category 3 Schedule SE Items. Finally, we treat a relevant line item derived on the Schedule SE as the dependent variable for audit category 3. This line item is continuous, measured in dollars. Again, we use the natural logarithm of this line item plus one dollar. Note that in this case, increasing reporting of these Schedule SE line items should have the effect of increasing overall tax liability; thus, we would expect a positive indirect effect to be associated with increased reporting on these line items.

We will now refer to the line items relevant to audit categories 1, 2, and 3 as "relevant items."

Independent Variables

Audit-Time Interaction. The primary variables of interest are audit status and its interaction with time, specified as tax years since the baseline year. Audit status is a time-invariant variable for each taxpayer, as they can be considered only as "audited" or "not audited" in our sample. Years after baseline is time-varying, meaning that it takes on a different value for each of a taxpayer's returns to describe the time between that return and the audited or eligible return. We define the baseline year as Year 0, and we fit time as a categorical variable rather than a continuous, numeric variable, such that its slope is not constrained to be linear. This allows for any potential attenuation in indirect effect to be captured.

Control Variables. A variety of control variables were assessed with the intent to account for possible changes in taxpayer characteristics over time, including financial situation, living situation, and family structure. For all models, we control for *Total Positive Income* (TPI), adjusted to reflect 2018 U.S. dollars.[9] We treat *Filing Status* (FS) as a binary variable, with 1 being Married Filing Jointly and the reference level being other filing statuses collapsed into one category (Single, Married Filing Separately, Widow/er, Head of Household). We derive an urban/not urban (*Urban)* classification using zip code data and Census Bureau definitions.[10] A binary wage indicator is derived based on the presence of any nonwage income reported on Form 1040 (*any wages)*. We adjust for *total exemptions*, and the presence of claiming any *Child Tax Credit*. To account for home ownership, we control for a continuous measure of mortgage interest deductions (*mortgage interest)*. For audit categories 1 and 3, we adjust our estimates for whether the taxpayer itemized deductions as indicated by the presence of a Schedule A (*itemized deductions)*. TPI, FS, Urban, any wages, total exemptions, any Child Tax Credit, mortgage interest, and itemized deductions all are treated as time-varying covariates. We also fit *tax year* of the return as a categorical variable with possible values Tax Years 2006–2018. In the models predicting total tax only, we also control for *Priority,* a variable representing the metric used operationally by the audit category in question to rank and select returns for audit. For audit categories 1 and 3, this is measured in 2018 USD with

[8] Inflation adjustment was conducted with the following formula: value in 2018 USD = (Consumer Price Index (CPI) in 2018/CPI in the TY of interest) * value in TY of interest.

[9] Total Positive Income is defined as the sum of wages, salaries, interest, and dividends and does not subtract losses or deductions.

[10] https://www.census.gov/geographies/reference-files/2010/geo/relationship-files.html#par_textimage_470670252.

the interpretation that higher priority is more likely to be audited. This variable is distinct for each category of audit and is time-invariant, meaning that it is the taxpayer's assigned priority in the baseline year.

Statistical Analysis

To assess the relationship between audit status and the outcomes of interest over time, a linear mixed effect model is fit for each outcome and audit category. Linear mixed effects models are longitudinal models in which within-subject correlation is captured and accounted for in the standard errors (Moulton (1986), in Bell and Jones (2015)). A random effect (γ_{0i}) is included for TIN, which allows each taxpayer to have their own "baseline" intercept for the dependent variable. A mixed effects model specification also has the advantage of allowing both time-varying and time invariant predictor and outcome variables (Bell and Jones (2015)), unlike fixed-effects-only models. Within-taxpayer correlation is modeled with an autoregressive structure, as is common with evenly spaced repeated measures. The model specifications are provided in equations (1) and (2) for the i^{th} taxpayer and j^{th} return (years after baseline). Analyses were conducted with R version 3.4.4, using the modeling package *nlme* (Pinheiro (2019)).

Model 1: Total Tax Reporting Over Time

For each category of audit, we separately estimate model (1) below, in which $\ln(\text{total tax} + 1)_{ij}$ denotes the natural logarithm of total tax in U.S. dollars plus one dollar, adjusted for inflation, for each individual *i* at year *j*. $\beta_{11}\, audited_i$ is a time-invariant measure of whether the taxpayer was audited for the tax return filed at baseline year. Models for audit category 2 are not adjusted for whether the taxpayer itemized their deductions since eligibility for this audit necessitates itemizing deductions. γ_{0i} denotes a random effect on individual *i*.

$$
\begin{aligned}
(1)\qquad & \ln(\text{total tax} + 1)_{ij} \\
& = \beta_0 + \gamma_{0i} + \beta_1 FS_{ij} + \beta_2 TY_{ij} + \beta_3 TPI_{ij} + \beta_4 priority_i + \beta_5 any\ wages_{ij} \\
& + \beta_6 total\ exemptions_{ij} + \beta_7 any\ Child\ Tax\ Credit_{ij} + \beta_8 itemized\ deductions_{ij} \\
& + \beta_9 mortgage\ interest_{ij} + \beta_{10} urban_{ij} + \beta_{11} audited_i + \beta_{12} year.after.baseline_{ij} \\
& + \beta_{13} audited_i * year.after.baseline_{ij} + \epsilon_{ij}
\end{aligned}
$$

Model 2: Audit-Specific Line Items Reporting Over Time

Next, for each category of audit, we separately estimate model (2) below, in which the natural logarithmic transformation of the sum of relevant items + 1 denotes a single or sum of relevant line items for the category of audit in U.S. dollars plus one dollar, adjusted for inflation, for each individual *i* at year *j*. $\beta_5\, audited_i$ is a time-invariant measure of whether or not the taxpayer was audited for the tax return filed at baseline year. γ_{0i} denotes a random effect on TIN.

$$
\begin{aligned}
(2)\qquad & \ln\Big(\sum \text{relevant items} + 1\Big)_{ij} \\
& = \beta_0 + \gamma_{0i} + \beta_1 FS_{ij} + \beta_2 TY_{ij} + \beta_3 TPI_{ij} + \beta_6 any\ wages_{ij} + \beta_7 total\ exemptions_{ij} \\
& + \beta_8 any\ Child\ Tax\ Credit_{ij} + \beta_9 itemized\ deductions_{ij} + \beta_{10} mortgage\ interest_{ij} \\
& + \beta_{11} urban_{ij} + \beta_{12} audited_i + \beta_{13} year.after.baseline_{ij} + \beta_{14} audited_i \\
& * year.after.baseline_{ij} + \epsilon_{ij}
\end{aligned}
$$

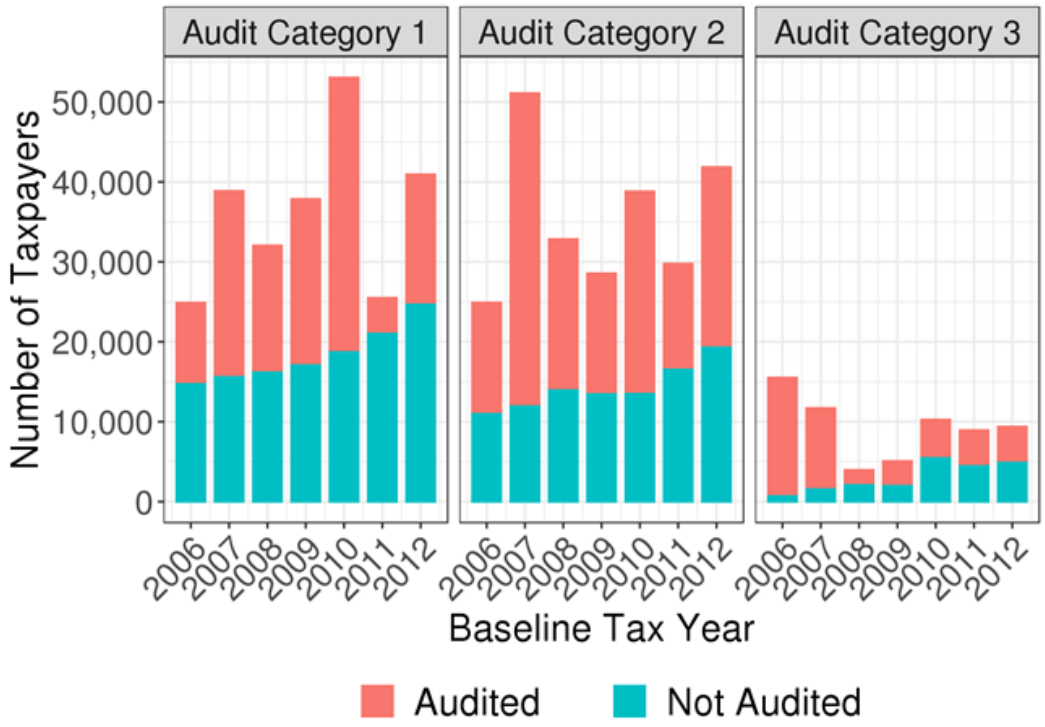
Results

The sample sizes of each audit category and baseline year are shown in Figure 1. Audit categories 1 and 2 have 253,132 and 247,837 unique taxpayers, respectively. For audit category 1, audits were most common in Tax Year 2010 and least common in Tax Year 2011. Similarly, for audit category 2, Tax Year 2011 was a lighter year for audits, while Tax Year 2007 has the highest audit frequency. Audit category 3 is a less common audit for which relatively few taxpayers were eligible but not audited. This category has in total 64,823 audited and not audited taxpayers.

Tables 1-3 summarize return characteristics by audit status for all data in the baseline year by audit category. All values are shown in 2018 USD. For continuous variables, we conducted a Wilcoxon rank-sum test to assess whether the difference between the audited and not-audited groups for each audit category are statistically significant. Categorical variables are assessed for association with audit status using a Chi-Square test. All differences are statistically significant at the $p < 0.05$ level.

For audit category 1, there is evidence to suggest that audited taxpayers at baseline have a statistically significantly higher total tax than the not-audited group. Additionally, the audited taxpayers at baseline appear to have a higher TPI compared to the not audited. In terms of audit priority, which is sensitive, unsurprisingly the audited taxpayers have on average significantly higher priority. However, it is important to note that there is still some overlap between groups: there are audited taxpayers with zero priority and a not-audited taxpayer with an extremely high priority. We cannot access information that would explain why a taxpayer with low priority would be audited and a taxpayer with high priority would not be. It is possible that these audits met some unknown exclusion criteria, the return was selected for a different audit, or something having to do with the timing of the return filing. The relevant items variable, which is the sum of some Schedule C line items, is also significantly higher for the audited group.

FIGURE 1. Sample sizes for all baseline years and audit categories



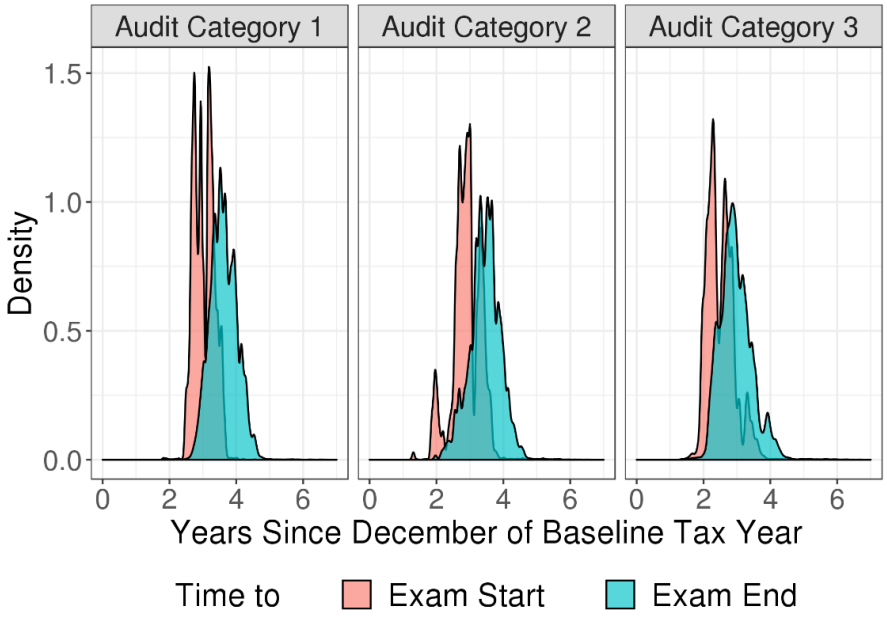
For audit category 2, there is again evidence to suggest that audited taxpayers at baseline have a statistically significantly different value for total tax compared to the not audited taxpayers. However, the audited taxpayers at baseline appear to have a slightly lower TPI compared to the not audited. The not audited group has a slightly higher priority than the audited group, which is further considered in the Discussion section.

Finally, for audit category 3, there is evidence to suggest that audited taxpayers at baseline have a significantly different distribution of characteristics for all variables considered. For example, the audited group appears to have on average higher TPI and total tax.

Timing of Audits

Figure 2 summarizes the time to audit exam start and end by audit category. We assume that exam start date coincides with when the taxpayer is notified that their return is being examined, and thus marks when we might expect to observe a behavioral response to the audit. The distribution of time to exam start in Figure 2 indicates that for most taxpayers and all three audit categories, taxpayers are notified of their audit approximately 2 to 3 years after the December of the TY for which they filed the audited return. Almost all taxpayers are aware that they are being audited within 4 years after the TY of the audited return. This suggests that if an indirect effect is present, it will mostly likely not manifest until 2 or 3 years after the TY of the audited return. For example, if a taxpayer is audited for their Tax Year 2008 return, which encompasses taxes paid through December 2008, they are likely to know about this audit by December 2011. They will file their Tax Year 2011 return between January 2012 and April 2012, meaning that we can expect this taxpayer to be aware they are being audited and exhibit any potential behavior change in their Tax Year 2011 return (3 years after baseline).

FIGURE 2. Density plots of the timing of audit exam start and end dates, relative to December of the TY of the audited return (audited taxpayers only)

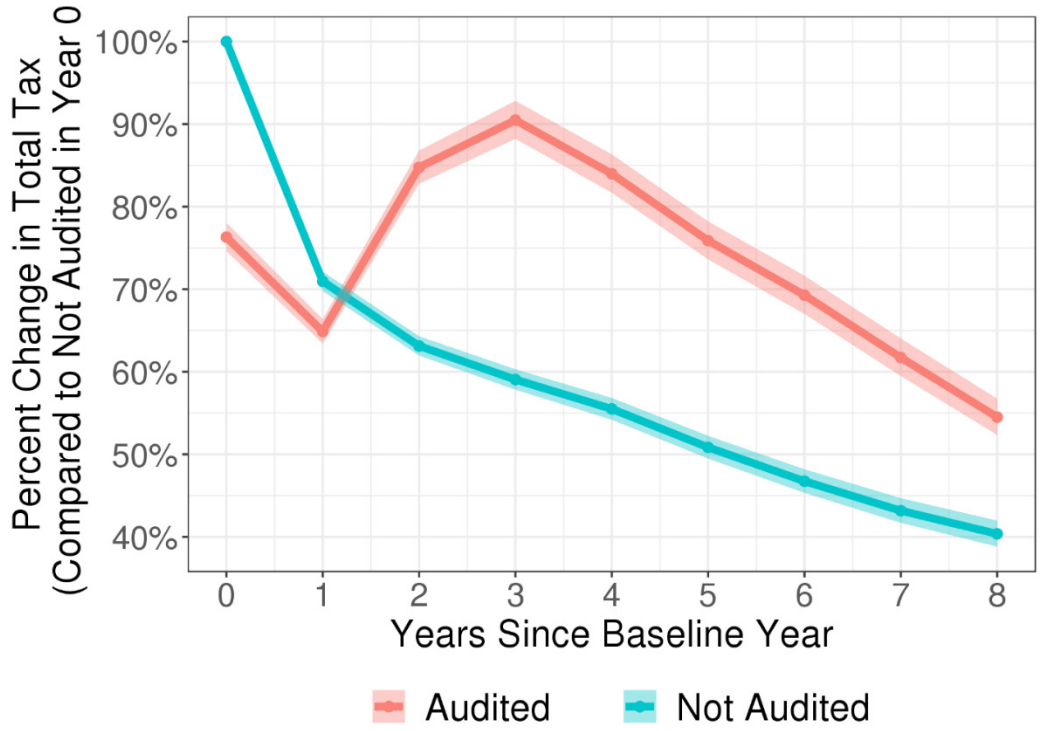


Modeling Results: Total Tax

Audit Category 1

Table 4 displays the estimates from the total tax model for audit category 1, which deals with Schedule C line items. Figure 3 shows the predicted changes in total tax over time for the audited and not-audited groups based on the estimated coefficients for the audited, years after baseline, and audit*years after baseline interaction variables. There is sufficient evidence to suggest a difference in total tax reporting for the baseline year: on average the audited taxpayers remit 76.3 percent of that of the not-audited taxpayers (95 percent confidence interval (CI) 74.6–78.0), while holding the control variables constant. One year after baseline, it is estimated that both groups have more similar values of total tax: the audited group paying 64.9 percent (CI 63.4–66.4) of that of the not audited in baseline's total tax and the not-audited group paying 70.9 percent (CI 69.8–72.1) of that of their own baseline total tax. However, in year 2, the audited group's predicted total tax increases sharply to 84.8 percent of that of the not audited in baseline (CI 82.8–86.8), while the not-audited group's estimated total tax decreases in slope. By 3 years and more after baseline, both groups show evidence of decreasing total tax over time when adjusting for control variables.

Figure 3. Predicted values for the linear mixed effects model of ln(total tax) in audit category 1

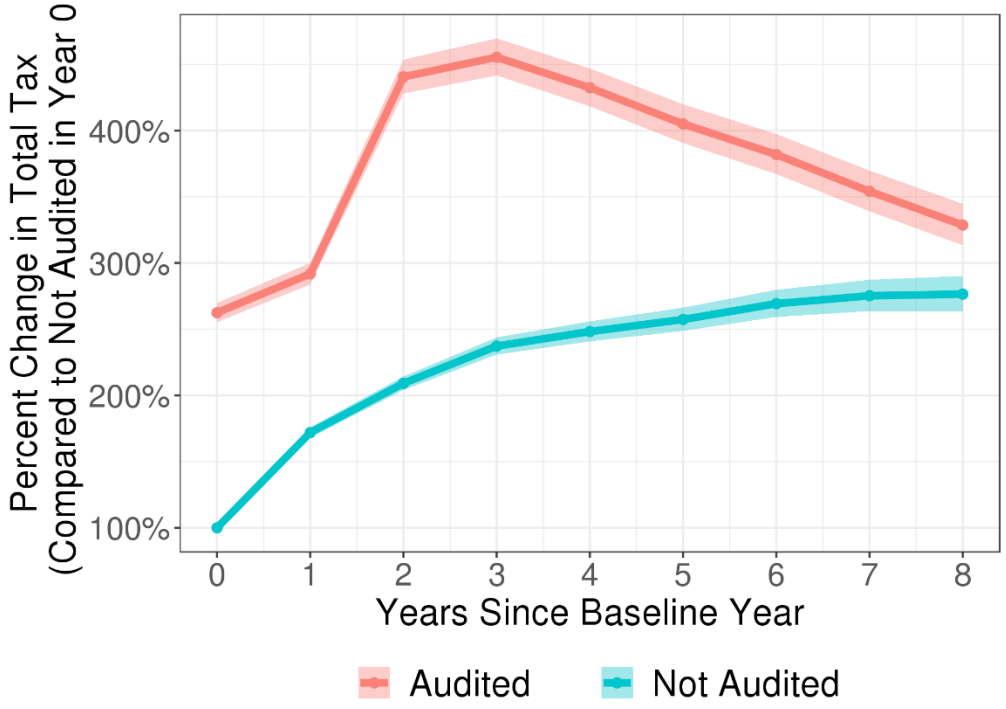


NOTE: Shading represents 95% confidence intervals. Not audited, year 0 is the reference group.

Audit Category 2

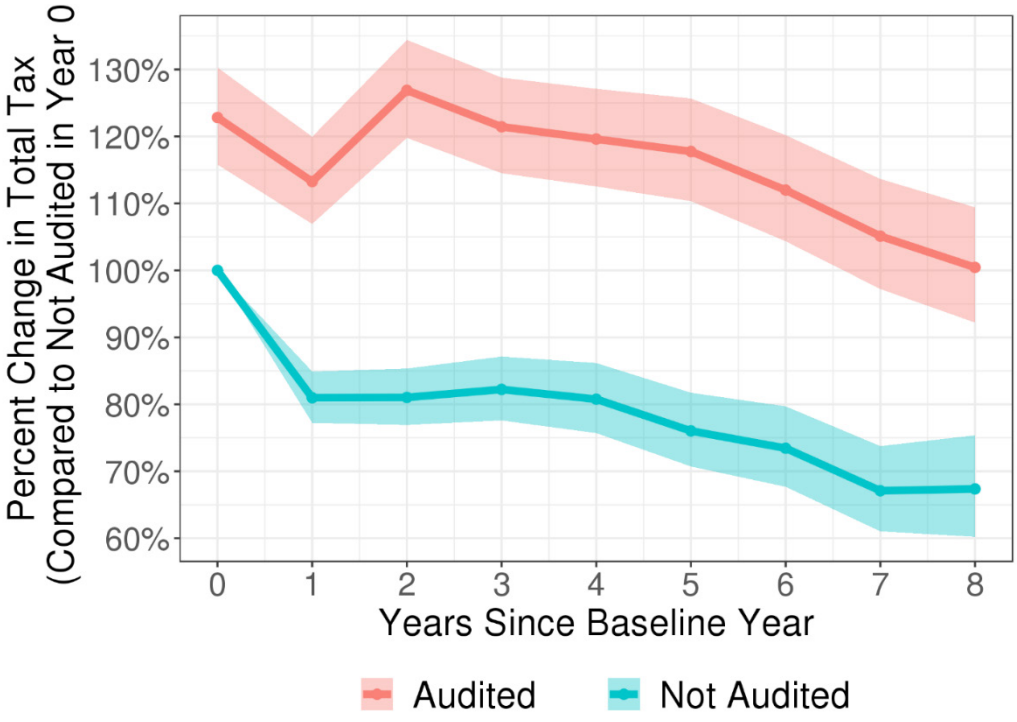
For audit category 2, which deals with Schedule A line items, the results of the total tax model are also presented in Table 4. Figure 4 shows the predicted values over time for the audited and not-audited groups. In year 0, for taxpayers with the same values of control variables and in the same tax year, it is estimated that the audited taxpayer on average has a total tax 2.62 times that of the not-audited taxpayer in the same year (CI 2.55–2.70). While there is evidence that not-audited taxpayers increase their total tax over time, there is also evidence of a significant jump in the audited taxpayers' total tax between 2 and 3 years after baseline. Two years after baseline, it is expected that audited taxpayers have a total tax 4.41 times that of not-audited taxpayers in year 0 (CI 4.28–4.54), while not-audited taxpayers are expected to increase their total tax just 2.09 times relative to their baseline tax (CI 2.04–2.14). The slope of the audited taxpayers is estimated to decrease beginning 3 years after baseline, while the not-audited estimated total tax is still increasing.

FIGURE 4. Predicted values for the linear mixed effects model of ln(total tax) in audit category 2



NOTE: Shading represents 95% confidence intervals. Not audited, year 0 is the reference group.

FIGURE 5. Predicted values for the linear mixed effects model of ln(total tax) in audit category 3



NOTE: Shading represents 95% confidence intervals. Not audited, year 0 is the reference group.

Audit Category 3

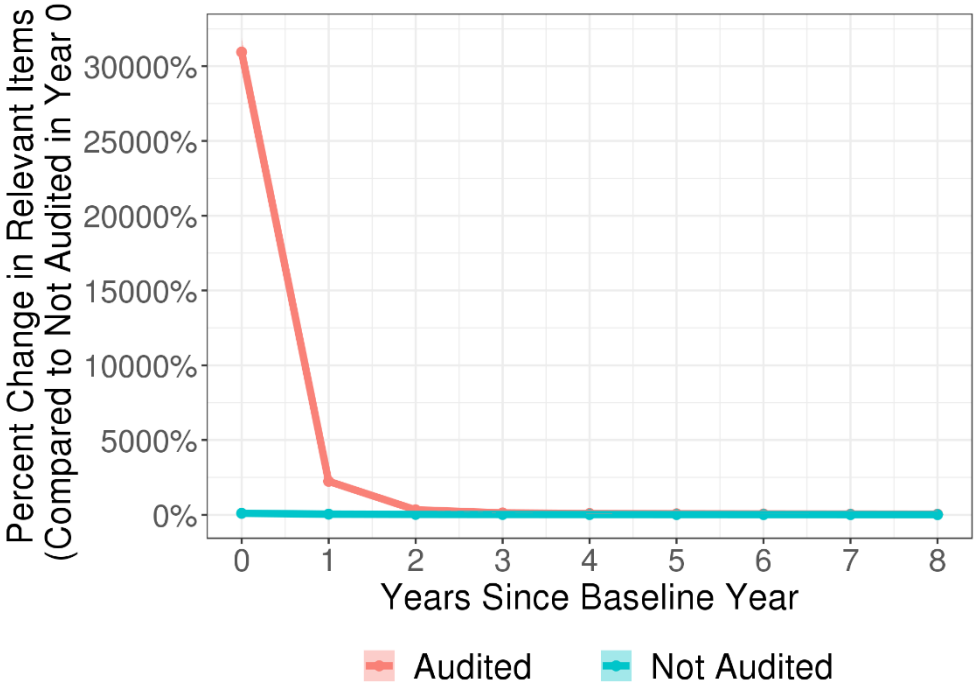
Finally, the results of the linear mixed effects model for the log of total tax in audit category 3 is also presented in Table 4 with predicted values plotted in Figure 5. In year 0, for taxpayers with the same values of control variables, and in the same tax year, it is estimated that the audited taxpayer on average has a total tax 22.8 percent more than that of the not audited taxpayer in the same year (CI 15.8–1.30). After both groups dip 1 year after baseline, the audited group's estimates increase at 2 years after baseline while the not-audited group remains approximately the same. By 3 years after baseline, the audited group's estimated total tax is decreasing.

Modeling Results: Audit Category Relevant Line Items

Audit Category 1

Table 5 displays the estimates from the model for audit category 1's sum of relevant items outcome, which deals with Schedule C line items. Figure 6 shows the predicted changes in relevant line items over time for the audited and not-audited groups. There is evidence to suggest a difference in relevant items reporting for the baseline year: on average the audited taxpayers have 309 times more in relevant items (CI 299.95–319.16) than that of the not-audited taxpayers, while holding the control variables constant. In year 1, the audited taxpayers are estimated to have 22.41 times more in relevant items compared to the not-audited group (CI 21.71–23.13). However, by year 2, there is insufficient evidence to suggest a difference in relevant items reporting between the audited and not-audited groups.

FIGURE 6. Predicted values for the linear mixed effects model of ln(relevant items) in audit category 1



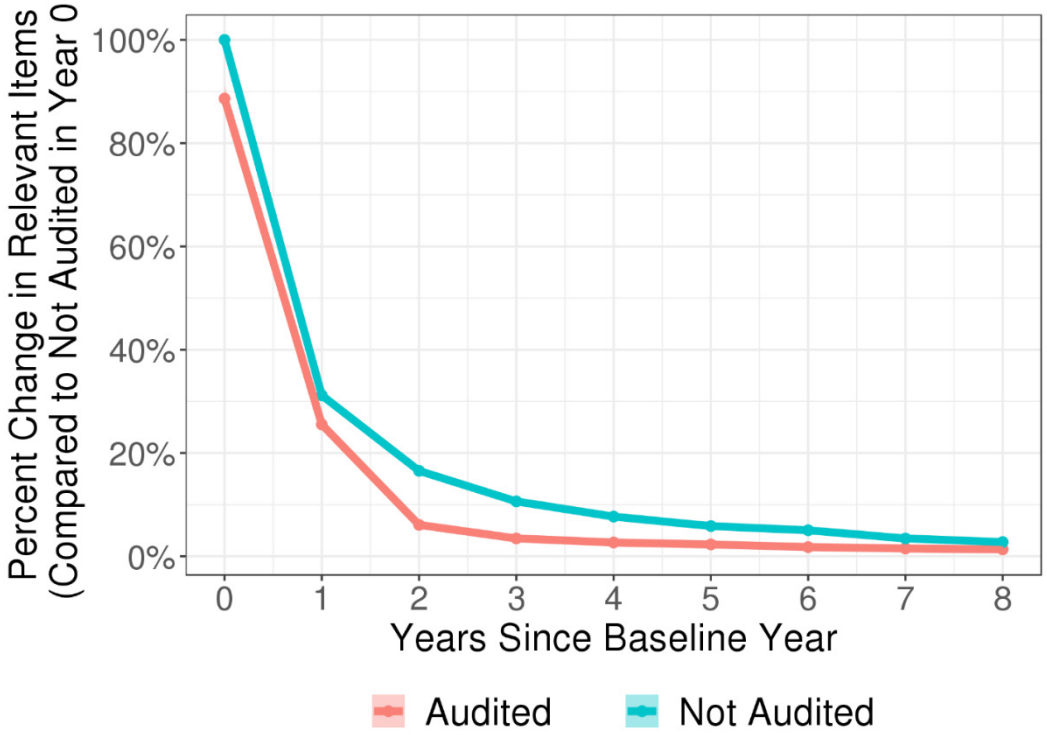
NOTE: Shading represents 95% confidence intervals. Not audited, year 0 is the reference group.

Audit Category 2

The predicted values for the model of relevant Schedule A items for audit category 2 are displayed in Figure 7. On average, it is estimated that in the baseline year the audited group has a sum of relevant items on average 11 percent lower than that of the not-audited group (estimate: 0.89, CI 0.86–0.91). One year later, the audited

taxpayers have a relevant sum 74 percent less than the not-audited group in the baseline year (estimate: 0.26, CI (0.25–0.26)), while the not-audited taxpayers report 70 percent less in relevant items compared to their prior year (CI 0.30–0.32). After years 2 and 3, both groups appear to trend towards no longer reporting the relevant items.

FIGURE 7. Predicted values for the linear mixed effects model of ln(relevant items) in audit category 2

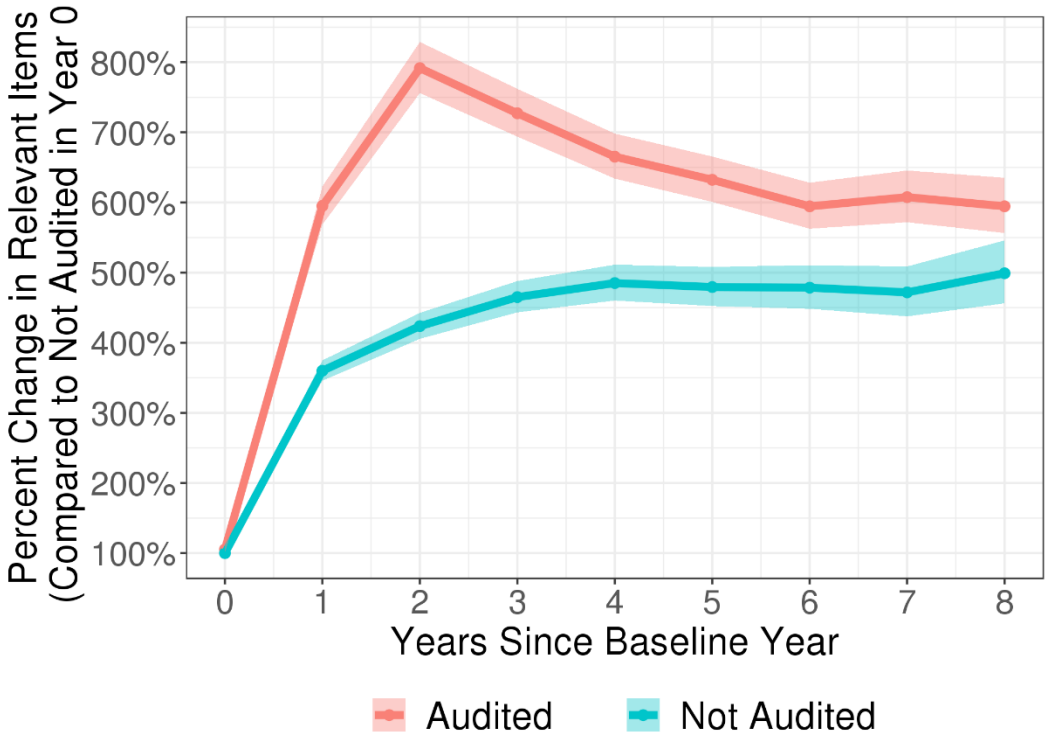


NOTE: Shading represents 95% confidence intervals. Not audited, year 0 is the reference group.

Audit Category 3

For audit category 3, regarding Schedule SE items, the audited and not-audited groups have relatively similar reporting at baseline assuming the same values of the control variables (see Figure 8). The audited group has on average 5 percent higher reporting of relevant items (estimate: 1.05, CI (1.01–1.10), p = 0.026). However, the audited group has a marked increase to 5.95 times that of the not audited baseline group by year 1 (CI 5.69–6.23), while the not-audited group increases less to 3.60 (CI 3.46–3.75). By year 2, the audited group peaks in its relevant items reporting to a multiplicative change of 7.92 (CI 7.56–8.29). Following that jump, the predicted values begin to have a negative slope and approach the estimates of the control group.

FIGURE 8. Predicted values for the linear mixed effects model of ln(relevant items) in audit category 3



NOTE: Shading represents 95% confidence intervals. Not audited, year 0 is the reference group.

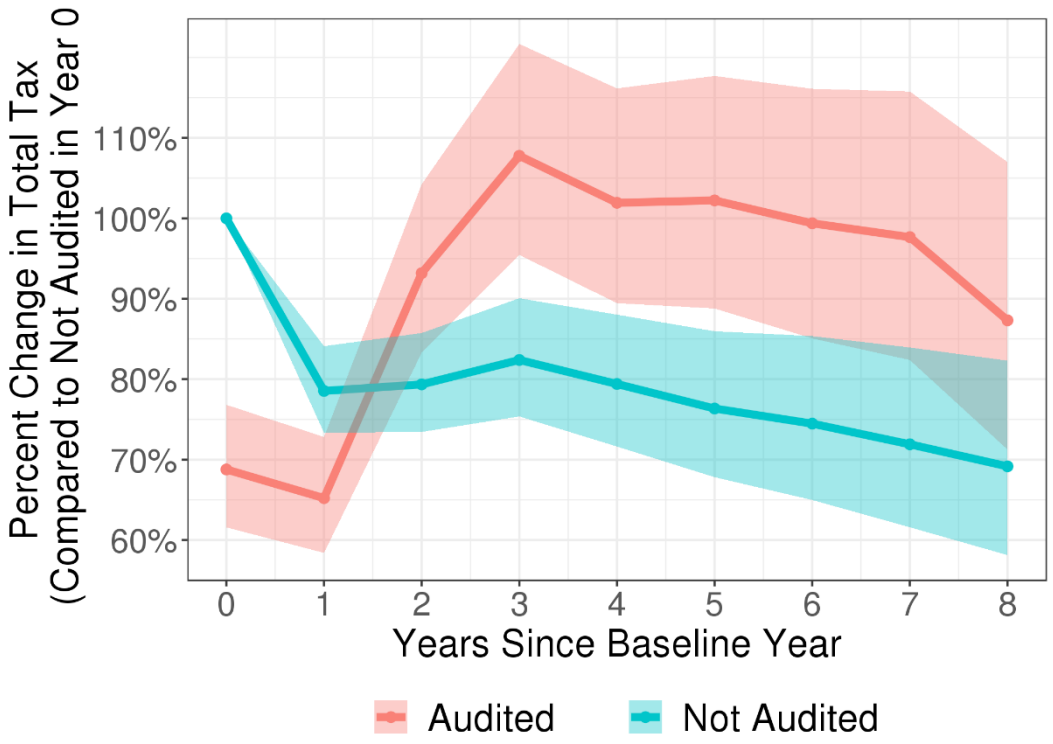
Sensitivity Analysis: Audit Category 1

To address the disparity in baseline characteristics for Audit Category 1 taxpayers, a sensitivity analysis is conducted on taxpayers with similar baseline priority. This is defined as taxpayers with a priority in the baseline year between $5,000 and $8,000 (2018 USD), a range chosen upon inspection of the distribution of priority by audit status. In total, 12,308 taxpayers fall into this range: 4,376 from the audited population and 7,932 from the not-audited population.

Total Tax

Table 6 displays the estimates from the total tax model for this sensitivity analysis of Audit Category 1 taxpayers with similar baseline priority. Figure 9 shows the predicted changes in total tax over time for the audited and not-audited groups based on the estimated coefficients for the audited, years after baseline, and audit*years after baseline interaction variables. There is enough evidence to suggest a difference in total tax reporting for the baseline year: on average the audited taxpayers remit 68.77 percent of that in total tax compared to the not-audited taxpayers (CI 61.58–76.81), while holding the control variables constant. However, in years 2 and 3, the audited group's predicted total tax increases further while the not-audited group appears to remain constant in slope. By 3 years after baseline, both groups show evidence of decreasing total tax over time when adjusting for our control variables.

FIGURE 9. Predicted values for the linear mixed effects model of ln(total tax) in sensitivity analysis of audit category 1

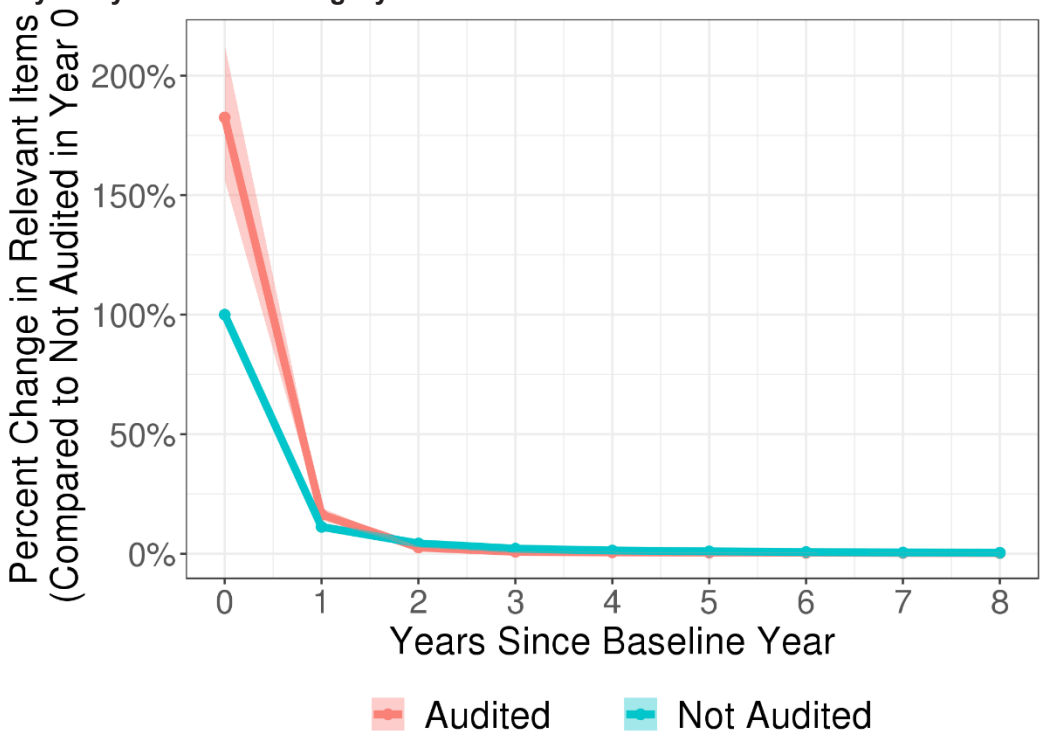


NOTE: Shading represents 95% confidence intervals. Not audited, year 0 is the reference group.

Relevant Line Items

Table 6 also displays the estimates from the model for audit category 1's sensitivity analysis on the sum of relevant items outcome, which deals with Schedule C line items. Figure 10 shows the predicted changes in line item reporting over time for the audited and not-audited groups. There is adequate evidence to suggest a difference in relevant items reporting for the baseline year: on average the audited taxpayers have 82 percent higher reporting in relevant line items (estimate 1.82 CI 1.56–2.13) than that of the not-audited taxpayers, while holding the control variables constant. In year 1 both groups decrease in their claiming of relevant line items: the audited taxpayers are estimated to report 16.37 percent of the not-audited group's baseline reporting (CI 14.04–19.01) and the not-audited reporting is on average 11.22 percent of their reporting the year prior (CI 10.18–12.36). However, by year 2, there is insufficient evidence to suggest a difference in relevant items reporting between the audited and not audited groups.

FIGURE 10. Predicted values for the linear mixed effects model of ln(relevant items) in sensitivity analysis of audit category 1



NOTE: Shading represents 95% confidence intervals. Not audited, year 0 is the reference group.

Discussion

In this study, we investigated the indirect effect of experiencing an audit on subsequent total tax reporting and on reporting of other relevant line items for three categories of correspondence audit. We advance prior literature in two ways: 1) by accounting for operational selection criteria and 2) by expanding the focus to include taxpayers who do not report self-employment income, but rather are examined for other types of reporting characteristics. Prior studies that use operational data to construct "treatment" and "control" groups *ex post* have typically relied on DIF scores when considering the likelihood of experiencing an audit (e.g., Beer (2015); Nestor and Beers (2014)); however, in the case of correspondence audits, other criteria are used instead of DIF, and we are able to account for these in this study. As in any study using operational data, we grapple with the challenge that taxpayers are selected into the "treatment" condition of audit based on criteria that is only partially known (Slemrod (2016)), even from within a narrowly defined candidate population.

For all three audit categories, we find evidence suggestive of an indirect effect. In audit category 1, which deals with Schedule C items, there is an increase in predicted total tax for the audited group around 1 to 3 years after baseline, followed by an attenuation out to year 8. Considering that most audit category 1 exams will have started 3 years after baseline, we assume that most of the audited taxpayers have been notified by the peak in reporting observed in Figure 3 at 3 years after baseline. In this way, our results mirror prior findings from both research audit data on Schedule C filers (DeBacker *et al.* (2018a)) and findings using operational data on Schedule C filers (Beer (2015)). Interestingly, we find similar evidence of a specific indirect effect for audit category 2 (Schedule A itemizers) and weak evidence of a specific indirect effect for audit category 3 (self-employment tax). To our knowledge, these specific taxpayer populations have not been explicitly examined in other studies, which have tended to focus on taxpayers who report self-employment income to other taxpayers more generally. Our findings suggest that when other populations of taxpayers who are audited are

compared to unaudited taxpayers with similar characteristics (i.e., the "eligible" group), specific indirect effects may emerge more clearly.

Further in alignment with prior studies, our results show evidence of attenuation of specific indirect effects across all three audit categories. We observe peak indirect effects around the time that taxpayers' audits generally start—around year 3 after the audited return was filed—and we then see convergence between audited and not audited groups starting about 5 years after audit. It is a notable contribution of this research that this attenuation appears to hold in three separate populations of taxpayers.

Although Table 1 (see the Appendix) shows differences in the underlying characteristics of the audited and not audited groups, it is important to note that all models are controlling for the audit priority variable. This allows us to account for a degree of selection bias in the "treatment" condition of being audited. Although the IRS could have applied further exclusion criteria of which we are unaware, the priority variable reflects knowledge that is typically unknown to researchers using operational audit data and represents a step in the right direction of accounting for the endogeneity inherent in using nonrandom audit data, and our future work aims to continue building on this. Because priority is included in the model as a control variable, interpretation of the estimated coefficients is for taxpayers *with the same audit priority in baseline*. Therefore, the modeling results apply to relatively homogenous taxpayers who are similarly likely to be audited. Our sensitivity analysis on an overlap in Audit Category 1 taxpayers confirms this, in which we see results similar for this subset of taxpayers who are more similar according to baseline priority as we see for the full model of total tax. While the audited and not-audited groups in the sensitivity analysis have different baseline estimates for total tax, there is still evidence to suggest that their slopes are significantly different over time. This interaction between audit and time is the primary predictor of interest in this study; the audited and not-audited groups may have different baseline values, but we are assessing whether they are parallel in reporting behavior over time.

Our final research questions ask whether the specific indirect effects on total tax reporting and relevant line item reporting vary by category of audit. Because we did not conduct formal hypothesis testing between these disparate populations, our results are more qualitative in nature.

First, in comparing total tax model results for the three audit categories, we see that the trajectories and magnitudes of difference in total tax paid by the audited and not-audited groups appear distinct across categories. For audit category 1 (Schedule C), audited taxpayers are estimated to start out at baseline paying less tax than the control group, but the groups' trajectories cross as audited taxpayers increase their total tax reporting, peaking at year 3 after baseline (when most taxpayers' audits start) and then attenuating. Audit category 1 is the only category in which the not-audited group has a consistently negative total tax slope. This may suggest that these audit-eligible taxpayers were reporting higher than usual tax liability in their baseline year, and this tax liability was correlated with the operational criteria that made them eligible for the audit in the first place. However, based on the multiplicative interpretation of the exponentiated model coefficients, the magnitude of the difference between groups appears relatively smaller compared to audit categories 2 and 3. In contrast, the trajectory of the audited group's reporting under audit category 2 (Schedule A) seems to show the most responsiveness among audited taxpayers across all audit categories: these taxpayers are predicted to increase their total tax reporting at years 2 and 3 after audit to more than four times the total tax reporting of unaudited taxpayers at the baseline year. Interestingly, there is evidence of attenuation of the specific indirect effect across all categories of audit, where we see the audit group's estimated change in tax approach that of the not-audited group in the later years. However, the timing of this appears earlier for audit category 3 (SE tax) than for categories 1 and 2, which could suggest that the latter have longer-lasting effects on tax reporting.

Comparing specific line item models, we see more pronounced differences between audit categories in the reporting trends for line items specific to the type of audit. First, we see that for some audit categories, "baseline" values of relevant items more closely align between the audited and not-audited groups. When adjusting for potential confounders, audit category 3 taxpayers are most alike in their baseline reporting of Schedule SE relevant items, and the audited group increases their reporting after audit more sharply than the unaudited group. This may indicate an "education effect"—that is, audited taxpayers learn about what they should have reported in order to submit a correct tax return, and they adjust their reporting accordingly. Similarly, audited Schedule A taxpayers appear to decrease their reporting of certain deductions in years 1 to 2 more steeply

after audit when compared to not-audited taxpayers. It is interesting to note that both audited and not-audited groups substantially decrease their deduction reporting after the baseline year; this is likely an artifact of the selection criteria that made all of these taxpayers eligible for the audit category at the baseline year when they entered our analytical sample. That is to say, they may have been selected into the audited or eligible group on the basis of having an "unusual" or outlier year for reporting certain items.

In contrast, under audit category 1, the trends in the reporting of certain Schedule C expenses tell a different story. The results shown in Figure 6 portray a heavy selection bias toward taxpayers who report higher values on these Schedule C line items when being designated for audit—one of the challenges of using operational data. By 2 years after audit, it appears that most of the audited taxpayers are no longer claiming these relevant Schedule C line items. Additional work needs to be done in constructing a more comparable control group for further analysis of line items.

Limitations

We applied operational eligibility criteria to construct a "control" group. In doing so, we operate under the assumption that the categories of audit we analyze here have been relatively stable over time, especially with regard to the types of line items examined in the audits. Still, it is possible that the current selection filters did not apply to all historic tax years: we are informed of current filters (e.g., those used for Tax Year 2018), but these filters may not necessarily apply to Tax Years 2006–2012, and we do not have knowledge of the eligibility criteria used in historic years for all audit categories. Similarly, formulation of the prioritization variables may have changed over time, but, without easy access to this knowledge, we must assume that the current prioritization for each audit category applies to Tax Years 2006–2012. Further, there appears to be some overlap between the distributions of priority for the audited and not audited groups, as observed in all three audit categories. This could potentially be due to the date the returns were filed and how quickly they were picked up in the correspondence audit cycle. However, discrepancies between priority and audit status could also imply that there exist additional audit selection criteria unknown to us.

Additionally, audited taxpayers have varying notification times, even for audits of returns from the same tax year, and results must be interpreted while considering the fact that not every taxpayer is aware of their audit by the time they are preparing their tax return for a subsequent tax year. Finally, not all taxpayers have a complete set of returns after the baseline year; this absence is assumed to be Missing at Random (MAR).

Finally, a mixed effects model assumes that the random effects are independent from the residuals. In the presence of unobserved confounders this assumption is not likely to be met, and therefore the estimation could be biased.

Future Research

Our plans for future research include executing analyses comparable to the ones presented here over additional categories of correspondence audit, as well as across other types of audits beyond correspondence. We will also continue to explore whether and how the audit category and underlying differences in population matter in terms of the form that a specific indirect effect takes. This approach has the operational potential of providing new information about which categories of audit have the greatest specific indirect effect on IRS revenue.

We acknowledge that there exist further control variables to be considered in future models, such as those that would better account for tax policy changes. Other data points available to us, such as whether a taxpayer used a tax preparer, might also have some degree of explanatory power in the relationship between audit experience and subsequent tax reporting, and future research should continue to investigate these relationships. Additionally, despite using the best filter criteria to select the control group, there appear to be different underlying characteristics between the audited and not-audited groups; thus, an assumption of exchangeability is unlikely to hold here. Ensuring we have comparable control groups for all audit categories is a priority of our research going forward. Given this, we have already arranged for a purely random control group to not be audited among returns filed for a recent tax year that meet all of the selection criteria of one of the three categories of audit we featured in this paper. That should allow us to evaluate how much our current results overstate or understate the indirect effect.

References

Advani, Arun, William Elming, and Jonathan Shaw. 2015. "How Long-Lasting Are the Effects of Audits?" *Tax Administration Research Centre* Discussion Paper 011-15. https://tarc.exeter.ac.uk/media/universityofexeter/businessschool/documents/centres/tarc/publications/discussionpapers/How_long_lasting_are_the_effects_of_audits_v3.pdf.

Ali, Mukhtar M., H. Wayne Cecil, and James A. Knoblett. 2001. "The Effects of Tax Rates and Enforcement Policies on Taxpayer Compliance: A Study of Self-Employed Taxpayers." *Atlantic Economic Journal* 29 (2): 186–202. https://doi.org/10.1007/BF02299137.

Beer, Sebastian, Mathias Kasper, Erich Kirchler, and Brian Erard. 2015. "Audit Impact Study." *TAS Research and Related Studies* 2. https://taxpayeradvocate.irs.gov/Media/Default/Documents/2015ARC/ARC15_Volume2_3-AuditImpact.pdf.

Bell, Andrew, and Kelvyn Jones. 2015. "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data." *Political Science Research and Methods 3*(1): 133–153.

Boning, William C., John Guyton, Ronald H. Hodge, II, Joel Slemrod, and Ugo Troiano. 2018. "Heard It Through the Grapevine: Direct and Network Effects of a Tax Enforcement Field Experiment." Working Paper 24305. National Bureau of Economic Research. https://www.nber.org/papers/w24305.pdf.

Datta, Saurabh, Stacy Orlett, and Alex Turk. 2015. "Individual Nonfilers and IRS-Generated Tax Assessments: Revenue and Compliance Impacts of IRS Substitute Assessments When Taxpayers Don't File." Small Business/Self-Employed Division, Internal Revenue Service. https://www.irs.gov/pub/irs-soi/15rescondatta.pdf.

DeBacker, Jason, Bradley T. Heim, Anh Tran, and Alexander Yuskavage. 2018a. "Once Bitten, Twice Shy? The Lasting Impact of Enforcement on Tax Compliance," *The Journal of Law and Economics* 61(1): 1-35. https://www.journals.uchicago.edu/doi/abs/10.1086/697683.

DeBacker, Jason, Bradley T. Heim, Anh Tran, and Alexander Yuskavage. 2018b. "The Effects of IRS Audits on EITC Claimants." *National Tax Journal* 71 (3): 451–484. https://doi.org/10.17310/ntj.2018.3.02.

Dubin, Jeffrey A., Michael J. Graetz, and Louis L. Wilde. 1990. "The Effect of Audit Rates on the Federal Individual Income Tax, 1977-1986." *National Tax Journal* 43 (4): 395–409.

Erard, Brian, and Chih-Chin Ho. 2003. "Explaining the U.S. Tax Compliance Spectrum." *eJournal of Tax Research* 1 (2): 93–109.

Guyton, John, Pat Langetieg, Day Manoli, Mark Payne, Brenda Schafer, and Michael Sebastiani. 2017. "Reminders and Recidivism: Using Administrative Data To Characterize Nonfilers and Conduct EITC Outreach." *American Economic Review* 107 (5): 471–75. https://doi.org/10.1257/aer.p20171062.

Guyton, John, Kara Leibel, Day Manoli, Ankur Patel, Mark Payne, and Brenda Schafer. 2018. "Tax Enforcement and Tax Policy: Evidence on Taxpayer Responses to EITC Correspondence Audits." Working Paper 24465. National Bureau of Economic Research. https://www.nber.org/papers/w24465.pdf.

Hallsworth, M. 2014. "The Use of Field Experiments To Increase Tax Compliance." *Oxford Review of Economic Policy* 30 (4): 658–79. https://doi.org/10.1093/oxrep/gru034.

Kastlunger, Barbara, Erich Kirchler, Luigi Mittone, and Julia Pitters. 2009. "Sequences of Audits, Tax Compliance, and Taxpaying Strategies." *Journal of Economic Psychology* 30 (3): 405–18. https://doi.org/10.1016/j.joep.2008.10.004.

Kleven, Henrik Jacobsen, Martin B. Knudsen, Claus Thustrup Kreiner, Søren Pedersen, and Emmanuel Saez. 2011. "Unwilling or Unable To Cheat? Evidence From a Tax Audit Experiment in Denmark." *Econometrica* 79 (3): 651–92. https://doi.org/10.3982/ECTA9113.

Maciejovsky, Boris, Erich Kirchler, and Herbert Schwarzenberger. 2007. "Misperception of Chance and Loss Repair: On the Dynamics of Tax Compliance." *Journal of Economic Psychology* 28 (6): 678–91. https://doi.org/10.1016/j.joep.2007.02.002.

Mazzolini, Gabriele, Laura Pagani, and Alessandro Santoro. 2017. "The Deterrence Effect of Real-World Operational Tax Audits." Working Paper No. 359. University of Milan Bicocca Department of Economics, Management and Statistics. http://dx.doi.org/10.2139/ssrn.2914374.

Meiselman, Ben. 2018. "Ghostbusting in Detroit: Evidence on Nonfilers from a Controlled Field Experiment." *University of Michigan Working Paper*. http://www-personal.umich.edu/~mdbmeis/MeiselmanJMP.pdf.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics 32 (3) 385–97*. https://doi.org/10.1016/0304-4076(86)90021-7.

Nestor, Mike, and Tom Beers. 2014. "Estimating the Impact of Audits on the Subsequent Reporting Compliance of Small Business Taxpayers: Preliminary Results." *TAS* 2. https://taxpayeradvocate.irs.gov/Media/Default/Documents/2014-Annual-Report/Estimating-the-Impact-of-Audits-on-the-Subsequent-Reporting-Compliance-of-Small-Business-Taxpayers-Preliminary-Results.pdf.

Perez-Truglia, Ricardo, and Ugo Troiano. 2015. "Shaming Tax Delinquents." Working Paper 21264. National Bureau of Economic Research. https://doi.org/10.3386/w21264.

Pinheiro, José. 2019. Package "nlme." Available at https://cran.r-project.org/web/packages/nlme/nlme.pdf.

Plumley, Alan. 1996. "The Determinants of Individual Income Tax Compliance." Department of the Treasury, Internal Revenue Service, Publication 1916 (Rev. 11–96).

United States Government Accountability Office. 2012. "IRS Could Significantly Increase Revenues by Better Targeting Enforcement Resources." Report GAO-13–151.

Slemrod, Joel. 2016. "Tax Compliance and Enforcement: New Research and Its Policy Implications." Ross School of Business Working Paper 1302. University of Michigan.

Slemrod, Joel, Marsha Blumenthal, and Charles Christian. 2001. "Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota." *Journal of Public Economics* 79 (3): 455–83. https://doi.org/10.1016/S0047-2727(99)00107-3.

Appendix

TABLE 1. Baseline Characteristics for Audit Category 1

Variable (Unit)	Statistic	Audited (N = 123,292)	Not Audited (N = 129,840)	Total (N = 253,132)	p-value
Total Tax (2018 USD)	Mean (SD)	51512 (397378)	11273 (33118)	30872 (279068)	<0.00011
TPI (2018 USD)	Mean (SD)	316583 (2101042)	96045 (142498)	203462 (1473991)	<0.00011
Priority (2018 USD)	Mean (SD)	31063 (1402755)	2023 (9696)	16167.9 (979115)	<0.00011
Filing Status (N (%))	Married Filing Jointly Single/Other	73668 (59.75%) 49624 (40.25%)	81615 (62.86%) 48225 (37.14%)	155283 (61.34%) 97849 (38.66%)	<0.00012
Relevant Items (2018 USD)	Mean (SD)	26625 (1215624)	1419 (3053)	13696 (848481)	<0.00011
Mortgage Interest (2018 USD)	Mean (SD)	10041 (15416)	5050 (8389)	7481 (12573)	<0.00011
Any Wage Income (N (%))	No Yes	8490 (6.89%) 114802 (93.11%)	12961 (6.54%) 116879 (88.42%)	21451 (8.47%) 231681 (91.53%)	<0.00012
Any Child Tax Credit (N (%))	No Yes	97111 (78.77%) 26181 (21.23%)	98300 (75.71%) 31540 (24.28%)	195411 (77.20%) 57721 (22.8%)	<0.00012
Itemized Deductions (N (%))	No Yes	40949 (33.21%) 82343 (66.79%)	66041 (50.86%) 63799 (49.14%)	106990 (42.27%) 146142 (57.73%)	<0.00012
Urban ZIP Code (N (%))	No Yes	2694 (2.19%) 120598 (97.81%)	2483 (1.91%) 127357 (98.09%)	5177 (2.05%) 247955 (97.95%)	<0.00012
Total Exemptions (N (%))	0 1 2 3 4 5+	108 (0.09%) 31828 (25.82%) 40532 (32.87%) 20952 (16.99%) 18996 (15.41%) 10876 (8.82%)	1703 (1.13%) 38915 (29.97%) 45748 (35.23%) 17106 (13.17%) 18519 (14.26%) 7849 (6.05%)	1811 (0.72%) 70743 (27.95%) 86280 (34.08%) 38058 (15.03%) 37515 (14.82%) 18725 (7.4%)	<0.00012

NOTES: [1]Wilcoxon rank sum test [2]Chi-Square test

TABLE 2. Baseline Characteristics for Audit Category 2

Variable (Unit)	Statistic	Audited (N = 146,337)	Not Audited (N = 101,500)	Total (N = 247,837)	p-value
Total Tax (2018 USD)	Mean (SD)	14443 (75399)	10998 (34016)	13032 (61915)	<0.00011
TPI (2018 USD)	Mean (SD)	148079 (394170)	161959 (289591)	153764 (355149)	<0.00011
Priority (2018 USD)	Mean (SD)	0.43 (3.24)	0.68 (18.81)	0.53 (12.30)	<0.00011
Filing Status (N (%))	Married Filing Jointly Single/Other	82211 (56.18%) 64126 (43.82%)	61090 (60.19%) 40410 (39.81%)	143301 (57.82%) 104536 (42.18%)	<0.00012
Relevant Items (2018 USD)	Mean (SD)	35756 (152083)	46705 (481205)	40240 (329421)	<0.00011
Mortgage Interest (2018 USD)	Mean (SD)	7391 (12644)	7194 (15438)	7311 (13857)	<0.00011
Any Wage Income (N (%))	No Yes	16260 (11.11%) 130077 (88.89%)	29360 (28.92%) 72140 (71.07%)	45620 (18.41%) 202217 (81.59%)	<0.00012
Any Child Tax Credit (N (%))	No Yes	119873 (81.92%) 26464 (18.08%)	88470 (87.16%) 13030 (12.84%)	208343 (84.06%) 39494 (15.94%)	<0.00012
Urban ZIP Code (N (%))	No Yes	2728 (1.86%) 143609 (98.14%)	2667 (2.63%) 98833 (97.37%)	5395 (2.18%) 242442 (97.82%)	<0.00012
Total Exemptions (N (%))	0 1 2 3 4 5+	119 (0.081%) 40249 (27.5%) 57382 (39.21%) 22781 (15.57%) 16431 (11.23%) 9375 (6.41%)	87 (0.086%) 27298 (26.89%) 42225 (41.60%) 14223 (14.01%) 10944 (10.78%) 6723 (6.62%)	206 (0.083%) 67547 (27.25%) 99607 (40.19%) 37004 (14.93%) 27375 (11.05%) 16098 (6.5%)	<0.00012

NOTES: [1]Wilcoxon rank sum test [2]Chi-Square test

TABLE 3. Baseline Characteristics for Audit Category 3

Variable (Unit)	Statistic	Audited (N = 41,849)	Not Audited (N = 22,974)	Total (N = 64,823)	p-value
Total Tax (2018 USD)	Mean (SD)	13922 (203452)	19597 (610147)	15933 (398329)	<0.00011
TPI (2018 USD)	Mean (SD)	120250 (2701267)	149923 (2250823)	130767 (2550763)	<0.00011
Priority (2018 USD)	Mean (SD)	4438.8 (68700)	5967 (158729)	4943 (107166)	<0.00011
Filing Status (N (%))	Married Filing Jointly Single/Other	15632 (37.35%) 26217 (62.65%)	8956 (39.98%) 14018 (61.02%)	24588 (37.93%) 40235 (62.07%)	<0.00012
Relevant Items (2018 USD)	Mean (SD)	21.65 (382.21)	33.71 (616.93)	25.92 (478.78)	<0.00011
Mortgage Interest (2018 USD)	Mean (SD)	4156 (23719)	4544 (11014)	4293 (20155)	<0.00011
Any Wage Income (N (%))	No Yes	17433 (41.66%) 24416 (58.34%)	12805 (55.74%) 10169 (44.26%)	30238 (46.65%) 34585 (53.35%)	<0.00012
Any Child Tax Credit (N (%))	No Yes	33654 (80.42%) 8195 (19.58%)	17949 (78.13%) 5025 (21.87%)	51603 (79.61%) 13220 (20.39%)	<0.00012
Itemized Deductions (N (%))	No Yes	27132 (64.83%) 14717 (35.17%)	14472 (63.00%) 8502 (37.00%)	41604 (64.18%) 23219 (35.82%)	<0.00012
Urban ZIP Code (N (%))	No Yes	998 (2.38%) 40851 (97.62%)	721 (3.14%) 22253 (96.86%)	1719 (2.65%) 63104 (97.35%)	<0.00012
Total Exemptions (N (%))	0 1 2 3 4 5+	535 (1.28%) 17886 (42.74%) 10321 (24.66%) 3715 (8.88%) 4173 (9.97%) 5219 (12.47%)	213 (0.93%) 6484 (27.80%) 5038 (21.93%) 2184 (9.51%) 3839 (16.71%) 5216 (22.70%)	748 (1.15%) 24370 (37.59%) 15359 (23.69%) 5899 (9.1%) 8012 (12.36%) 10435 (16.1%)	<0.00012

NOTES: [1]Wilcoxon rank sum test [2]Chi-Square test

TABLE 4. Estimates from Linear Mixed Model Predicting the Natural Log of Total Tax

	Audit Category 1		Audit Category 2		Audit Category 3	
Variable	Estimate (95% CI)	p	Estimate (95% CI)	p	Estimate (95% CI)	p
Audited	0.763 (0.746, 0.780)	0	2.624 (2.553, 2.697)	0	1.228 (1.158, 1.303)	0
Married filing jointly	4.660 (4.578, 4.743)	0	5.387 (5.281, 5.494)	0	11.033 (10.609, 11.474)	0
Urban zip code	1.331 (1.288, 1.375	0	1.092 (1.048, 1.138)	0	1.47 (1.37, 1.577)	0
Any wage income	2.946 (2.907, 2.987)	0	3.38 (3.327, 3.434)	0	1.993 (1.944, 2.042)	0
Itemized deductions	2.136 (2.117, 2.156)	0	NA	NA	3.262 (3.179, 3.347)	0
Mortgage interest	1 (1, 1)	0	1 (1, 1)	0	1 (1, 1)	0
Any Child Tax Credit	0.775 (0.765, 0.784)	0	0.975 (0.96, 0.991)	0.002	1.024 (0.993, 1.056)	0.14
Total exemptions (reference = 0)						
1	3.127 (2.909, 3.362)	0	3.752 (3.188, 4.415)	0	3.805 (3.384, 4.278)	0
2	1.451 (1.348, 1.563)	0	1.406 (1.194, 1.656)	0	0.949 (0.84, 1.072)	0.40
3	1.089 (1.011, 1.173)	0.026	0.851 (0.723, 1.003)	0.053	0.463 (0.409, 0.525)	0
4	0.851 (0.789, 0.918)	0	0.57 (0.483, 0.671)	0	0.178 (0.157, 0.202)	0
5+	0.538 (0.498, 0.581)	0	0.333 (0.282, 0.393)	0	0.067 (0.059, 0.076)	0
Tax year (Reference = 2006)						
2007	0.948 (0.919, 0.979)	0.0009	0.986 (0.95, 1.023)	0.46	1.018 (0.965, 1.075)	0.5054
2008	0.855 (0.829, 0.883)	0	0.601 (0.579, 0.624)	0	0.832 (0.783, 0.884)	0
2009	0.771 (0,746, 0.796)	0	0.421 (0.405, 0.438)	0	0.71 (0.666, 0.756)	0
2010	0.772 (0.746, 0.797)	0	0.47 (0.451, 0.489)	0	0.788 (0.738, 0.841)	0
2011	0.876 (0.845, 0.907)	0	0.511 (0.49, 0.533)	0	0.756 (0.705, 0.811)	0
2012	1.093 (1.052, 1.135)	0	0.607 (0.58, 0.635)	0	0.823 (0.763, 0.888)	0
2013	1.254 (1.204, 1.306)	0	0.647 (0.616, 0.679)	0	0.926 (0.851, 1.007)	0.0734
2014	1.489 (1.425, 1.555)	0	0.761 (0.722, 0.801)	0	1.13 (1.03, 1.239)	0.0094
2015	1.654 (1.578, 1.733)	0	0.735 (0.695, 0.777)	0	1.215 (1.098, 1.344)	0.0002
2016	1.741 (1.655, 1.832)	0	0.687 (0.647, 0.73)	0	1.236 (1.107, 1.38)	0.0002
2017	1.944 (1.841, 2.052)	0	0.806 (0.756, 0.859)	0	1.323 (1.174, 1.491)	0
2018	2.267 (2.138, 2.403)	0	0.997 (0.93, 1.068)	0.93	1.598 (1.401, 1.822)	0
Priority at Baseline	1 (1, 1)	0.47	0.997 (0.996, 0.997)	0	1 (1, 1)	0.024
TPI	1 (1, 1)	0	1 (1, 1)	0	1 (1, 1)	0
Reference = Year 0						
Year 1	0.709 (0.698, 0.721)	0	1.719 (1.681, 1.759)	0	0.81 (0.772, 0.849)	0
Year 2	0.631 (0.620, 0.643)	0	2.091 (2.04, 2.143)	0	0.81 (0.769, 0.853)	0
Year 3	0.591 (0.578, 0.603)	0	2.372 (2.308, 2.438)	0	0.822 (0.776, 0.871)	0
Year 4	0.555 (0.542, 0.568)	0	2.482 (2.408, 2.559)	0	0.808 (0.757, 0.862)	0
Year 5	0.508 (0.495, 0.522)	0	2.573 (2.487, 2.663)	0	0.76 (0.707, 0.817)	0
Year 6	0.467 (0.453, 0.482)	0	2.693 (2.593, 2.797)	0	0.734 (0.677, 0.797)	0
Year 7	0.432 (0.417, 0.447)	0	2.752 (2.637, 2.873)	0	0.671 (0.61, 0.738)	0
Year 8	0.404 (0.388, 0.420)	0	2.765 (2.635, 2.901)	0	0.674 (0.602, 0.754)	0
Audited * Years after baseline						
Audited*Year 1	1.198 (1.171, 1.226)	0	0.647 (0.629, 0.665)	0	1.139 (1.073, 1.208)	0
Audited*Year 2	1.76 (1.720, 1.800)	0	0.803 (0.78, 0.826)	0	1.275 (1.198, 1.357)	0
Audited*Year 3	2.008 (1.962, 2.055)	0	0.731 (0.711, 0.753)	0	1.202 (1.128, 1.282)	0
Audited*Year 4	1.983 (1.938, 2.030)	0	0.664 (0.644, 0.684)	0	1.206 (1.13, 1.287)	0
Audited*Year 5	1.956 (1.911, 2.003)	0	0.6 (0.582, 0.618)	0	1.261 (1.18, 1.347)	0
Audited*Year 6	1.943 (1.897, 1.990)	0	0.54 (0.524, 0.557)	0	1.241 (1.158, 1.33)	0
Audited*Year 7	1.874 (1.827, 1.922)	0	0.49 (0.474, 0.506)	0	1.275 (1.181, 1.378)	0
Audited*Year 8	1.769 (1.723, 1.817)	0	0.453 (0.437, 0.469)	0	1.214 (1.108, 1.33)	0

NOTE: Coefficients are exponentiated and represent a multiplicative change in total tax.

TABLE 5. Estimates from Linear Mixed Model Predicting the Natural Log of Relevant Items for the Three Audit Categories

	Audit Category 1		Audit Category 2		Audit Category 3	
Variable	*Estimate (95% CI)*	*p*	*Estimate (95% CI)*	*p*	*Estimate (95% CI)*	*p*
Audited	309 (300, 319)	0	0.886 (0.861, 0.912)	0	1.054 (1.006, 1.105)	0.026
Married filing jointly	2.102 (2.051, 2.154)	0	3.398 (3.327, 3.471)	0	1.54 (1.492, 1.589)	0
Urban zip code	0.988 (0.944, 1.034)	0.61	1.069 (1.022, 1.117)	0.0032	1.075 (1.014, 1.139)	0.014
Any wage income	0.439 (0.431, 0.447)	0	1.176 (1.156, 1.196)	0	0.261 (0.256, 0.266)	0
Itemized deductions	0.904 (0.893, 0.916)	0	NA	NA	0.916 (0.897, 0.936)	0
Mortgage interest	1 (1, 1)	0.47	1 (1, 1)	0	1 (1, 1)	0
Any Child Tax Credit	0.832 (0.818, 0.846)	0	0.7 (0.688, 0.712)	0	0.867 (0.845, 0.889)	0
Total exemptions (reference = 0)						
1	2.595 (2.348, 2.869)	0	3.187 (2.669, 3.807)	0	1.523 (1.379, 1.682)	0
2	2.779 (2.509, 3.078)	0	2.124 (1.778, 2.539)	0	1.556 (1.404, 1.724)	0
3	3 (2.705, 3.327)	0	1.847 (1.545, 2.208)	0	1.937 (1.744, 2.151)	0
4	2.961 (2.666, 3.287)	0	1.709 (1.428, 2.044)	0	1.785 (1.606, 1.984)	0
5+	2.732 (2.455, 3.041)	0	1.497 (1.25, 1.794)	0	1.668 (1.5, 1.856)	0
Tax year (Reference = 2006)						
2007	0.994 (0.952, 1.038)	0.80	0.9 (0.865, 0.938)	0	0.993 (0.949, 1.04)	0.78
2008	1.087 (1.040, 1.135)	0.0002	0.839 (0.805, 0.875)	0	0.861 (0.818, 0.907)	0
2009	1.092 (1.045, 1.142)	0.0001	0.736 (0.705, 0.768)	0	0.831 (0.788, 0.875)	0
2010	1.118 (1.068, 1.171)	0	0.851 (0.815, 0.889)	0	0.86 (0.815, 0.907)	0
2011	1.186 (1.129, 1.246)	0	0.876 (0.837, 0.917)	0	0.831 (0.786, 0.879)	0
2012	1.252 (1.188, 1.319)	0	0.881 (0.84, 0.924)	0	0.887 (0.836, 0.942)	0.0001
2013	1.406 (1.329, 1.489)	0	0.965 (0.916, 1.016)	0.17	0.837 (0.784, 0.894)	0
2014	1.515 (1.425, 1.611)	0	1.159 (1.097, 1.224)	0	0.845 (0.786, 0.907)	0
2015	1.649 (1.543, 1.761)	0	1.367 (1.288, 1.45)	0	0.86 (0.796, 0.93)	0.0002
2016	1.712 (1.595, 1.838)	0	1.604 (1.506, 1.708)	0	0.862 (0.792, 0.938)	0.0006
2017	1.786 (1.655, 1.926)	0	1.866 (1.745, 1.995)	0	0.849 (0.775, 0.93)	0.0004
2018	1.775 (1.636, 1.927)	0	0.083 (0.077, 0.089)	0	0.783 (0.708, 0.866)	0
TPI	1 (1, 1)	0.54	1 (1, 1)	0	1 (1, 1)	0.67
Reference = Year 0						
Year 1	0.432 (0.422, 0.442)	0	0.312 (0.304, 0.319)	0	3.601 (3.459, 3.749)	0
Year 2	0.284 (0.277, 0.292)	0	0.165 (0.161, 0.17)	0	4.237 (4.056, 4.425)	0
Year 3	0.22 (0.214, 0.226)	0	0.106 (0.103, 0.109)	0	4.651 (4.433, 4.879)	0
Year 4	0.182 (0.176, 0.188)	0	0.077 (0.074, 0.079)	0	4.851 (4.601, 5.113)	0
Year 5	0.155 (0.150, 0.161)	0	0.059 (0.056, 0.061)	0	4.794 (4.524, 5.081)	0
Year 6	0.133 (0.127, 0.139)	0	0.05 (0.048, 0.052)	0	4.785 (4.484, 5.107)	0
Year 7	0.119 (0.113, 0.125)	0	0.034 (0.033, 0.036)	0	4.719 (4.376, 5.087)	0
Year 8	0.105 (0.099, 0.110)	0	0.028 (0.026, 0.029)	0	4.99 (4.562, 5.459)	0
Audited * Years after baseline						
Audited*Year 1	0.168 (0.163, 0.173)	0	0.924 (0.896, 0.953)	0	1.568 (1.49, 1.649)	0
Audited*Year 2	0.038 (0.036, 0.039)	0	0.415 (0.402, 0.428)	0	1.772 (1.68, 1.869)	0
Audited*Year 3	0.018 (0.018, 0.019)	0	0.367 (0.355, 0.379)	0	1.483 (1.405, 1.566)	0
Audited*Year 4	0.015 (0.015, 0.016)	0	0.393 (0.38, 0.405)	0	1.301 (1.231, 1.375)	0
Audited*Year 5	0.014 (0.013, 0.014)	0	0.442 (0.428, 0.457)	0	1.251 (1.182, 1.324)	0
Audited*Year 6	0.013 (0.013, 0.013)	0	0.4 (0.387, 0.413)	0	1.178 (1.111, 1.25)	0
Audited*Year 7	0.012 (0.012, 0.013)	0	0.494 (0.476, 0.511)	0	1.221 (1.143, 1.305)	0
Audited*Year 8	0.012 (0.012, 0.013)	0	0.55 (0.529, 0.571)	0	1.13 (1.045, 1.222)	0.002

NOTE: Coefficients are exponentiated and represent a multiplicative change in relevant items reporting.

TABLE 6. Estimates from Linear Mixed Models for Sensitivity Analysis of Audit Category 1

	Outcome: Total Tax		Outcome: Relevant Line Items	
Variable	*Estimate (95% CI)*	*p*	*Estimate (95% CI)*	*p*
Audited	0.688 (0.616, 0.768)	0	1.825 (1.564, 2.129)	0
Married filing jointly	4.201 (3.872, 4.558)	0	2.512 (2.239, 2.818)	0
Urban zip code	1.14 (0.987, 1.317)	0.075	0.977 (0.797, 1.196)	0.82
Any wage income	2.799 (2.639, 2.97)	0	0.435 (0.4, 0.473)	0
Itemized deductions	1.873 (1.79, 1.959)	0	0.919 (0.863, 0.98)	0.010
Mortgage interest	1 (1, 1)	0	1 (1, 1)	0.0001
Any Child Tax Credit	0.911 (0.861, 0.964)	0.0013	0.865 (0.798, 0.937)	0.0004
Total exemptions (reference = 0)				
1	3.813 (2.293, 6.339)	0	1.913 (0.931, 3.931)	0.0777
2	1.735 (1.038, 2.9)	0.0354	1.795 (0.867, 3.715)	0.1151
3	1.266 (0.755, 2.122)	0.3712	2.002 (0.963, 4.162)	0.0632
4	0.855 (0.508, 1.438)	0.5547	2.055 (0.984, 4.291)	0.0552
5+	0.508 (0.3, 0.859)	0.0115	1.714 (0.814, 3.608)	0.1558
Tax year (Reference = 2006)				
2007	1.171 (0.977, 1.404)	0.0883	0.759 (0.586, 0.982)	0.0355
2008	0.91 (0.759, 1.09)	0.3069	0.764 (0.591, 0.988)	0.04
2009	0.783 (0.651, 0.941)	0.009	0.691 (0.532, 0.896)	0.0053
2010	0.748 (0.62, 0.904)	0.0026	0.888 (0.68, 1.158)	0.3798
2011	0.78 (0.641, 0.949)	0.0129	1.038 (0.788, 1.368)	0.7915
2012	0.864 (0.705, 1.058)	0.1568	1.075 (0.808, 1.429)	0.6197
2013	0.966 (0.778, 1.198)	0.7516	1.335 (0.986, 1.807)	0.0615
2014	1.082 (0.862, 1.358)	0.4986	1.506 (1.095, 2.072)	0.0118
2015	1.189 (0.934, 1.513)	0.1596	1.772 (1.265, 2.482)	0.0009
2016	1.189 (0.922, 1.535)	0.1824	1.877 (1.315, 2.68)	0.0005
2017	1.261 (0.963, 1.652)	0.0913	2.014 (1.383, 2.932)	0.0003
TPI	1 (1, 1)	0	1 (1, 1)	0.81
Priority at baseline	1 (1, 1)	0.27	NA	NA
Reference = Year 0				
Year 1	0.785 (0.733, 0.841)	0	1.825 (1.564, 2.129)	0
Year 2	0.794 (0.734, 0.857)	0	0.112 (0.102, 0.124)	0
Year 3	0.824 (0.754, 0.901)	0	0.043 (0.039, 0.048)	0
Year 4	0.794 (0.716, 0.88)	0	0.021 (0.019, 0.024)	0
Year 5	0.764 (0.678, 0.86)	0	0.014 (0.012, 0.016)	0
Year 6	0.745 (0.65, 0.854)	0	0.01 (0.008, 0.012)	0
Year 7	0.719 (0.616, 0.839)	0	0.007 (0.006, 0.009)	0
Year 8	0.692 (0.581, 0.823)	0	0.005 (0.004, 0.006)	0
Audited * Years after baseline				
Audited*Year 1	1.207 (1.067, 1.365)	0.0027	0.799 (0.671, 0.952)	0.012
Audited*Year 2	1.708 (1.511, 1.929)	0	0.331 (0.278, 0.394)	0
Audited*Year 3	1.902 (1.698, 2.131)	0	0.208 (0.177, 0.244)	0
Audited*Year 4	1.867 (1.654, 2.106)	0	0.238 (0.2, 0.282)	0
Audited*Year 5	1.947 (1.724, 2.198)	0	0.261 (0.22, 0.311)	0
Audited*Year 6	1.94 (1.716, 2.195)	0	0.311 (0.261, 0.371)	0
Audited*Year 7	1.975 (1.738, 2.244)	0	0.345 (0.288, 0.414)	0
Audited*Year 8	1.836 (1.564, 2.154)	0	0.37 (0.294, 0.464)	0

NOTE: Coefficients are exponentiated and represent multiplicative changes in outcomes.

Enforcement Versus Outreach—Impacts on Tax Filing Compliance

Anne Herlache, Ishani Roy, and Alex Turk (IRS Research, Applied Analytics, and Statistics) and Stacy Orlett (IRS, Small Business/Self-Employed Division)[1]

Introduction

The Internal Revenue Service (IRS) commonly uses mailed outreach to communicate with taxpayers who have compliance issues. For example, notices are sent through the U.S. Mail to inform people of their unpaid tax liabilities or their unmet requirements to file a return. Such notices are part of larger enforcement processes that are costly for both the IRS and the taxpayer in terms of the time spent and resources used to address noncompliance and promote voluntary compliance. To reduce both the costs of tax administration and the burden on taxpayers, the IRS is exploring alternative ways to promote voluntary compliance. By encouraging people to meet their tax responsibilities through low-cost outreach, time and resources can be refocused on more difficult cases that could not be resolved through lighter-touch methods.

In this paper, we report the initial results from a pilot test, comparing the impact of tax enforcement via notices regarding delinquent tax returns to outreach encouraging taxpayers to file their current tax return. This study began in April 2018, during the 2017 filing season, and focuses on filing behavior for Tax Years (TYs) 2016, 2017, and 2018. Over three waves of outreach, we contacted taxpayers with reminders, "soft" notices (i.e., letters worded less forcefully than the typical enforcement notices), and/or delinquent return notices. This design yielded insight into the tradeoffs of addressing past noncompliance versus encouraging current and future compliance, the potential benefit of repeated contact, and the impacts of delaying treatments in terms of future compliance ramifications. Furthermore, we sampled two populations of taxpayers: 1) known nonfilers (i.e., those who did not file their prior-year return and who were likely to have a filing requirement in the current tax year), and 2) potential stopfilers (i.e., taxpayers who filed their prior-year return but were at risk for becoming nonfilers in the current tax year). We garnered insights not only of the trade-off in timing and type of contact for noncompliant taxpayers, but we also learned how to help at-risk taxpayers avoid the challenges associated with becoming noncompliant in the first place.

Background and Related Research

IRS Context

Nonfiler Sample

The IRS designed a pilot to test alternative outreach that would promote voluntary filing compliance among those who had failed to file their previous year's tax return. These taxpayers were identified via the Individual Case Creation Nonfiler Identification Process (CCNIP) as having an unmet filing requirement for TY 2016.

Third-party income and other information provided via information returns (e.g., Form W-2) allows the IRS to identify individuals with a potential filing requirement. These potential nonfilers can then be prioritized for selection into the enforcement stream. Those who are selected will enter the Return Delinquency (RD) notice process and receive up to two notices requesting that they file their return. Nonfilers who do not file in response to the notices may enter the Taxpayer Delinquency Investigation (TDI) process, which can lead to enforcement action carried out in a variety of ways; for example, the Automated Collection System or Field

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

Collection. Each step in this process can be costly for the IRS and burdensome for the taxpayer, making earlier resolution desirable from both enforcement and service aspects.

Stopfiler Sample

To expand the scope of the project, the research team developed a model to identify taxpayers at risk for becoming "stopfilers" (hereafter referred to as stopfilers). By using multiple years of data to refine the model, we were able to score taxpayers propensity to nonfile in TY 2017 (model described below). Prior research has shown that nonfiling can become habitual; that is, once taxpayers become noncompliant, they frequently remain that way until the IRS intervenes (Rosage (1995)). By identifying potential stopfilers and testing encouragement methods for remaining compliant, we were able to determine if low-cost outreach can help prevent taxpayers from going down a costly path.

Concepts Related to Low-Cost Outreach

While there are many gradations within each path, taxpayers can become nonfilers by one of two routes: an active decision not to file or a passive decision not to file. The challenge posed to tax administration is quite different depending on the route taken. In this section, we focus on some theoretical underpinnings of passive nonfiling, or nonfiling that results from something other than a deliberate choice to shirk one's obligation to file (i.e., evasion). We expect that these nonfilers may be the most likely to respond to lighter-touch communications, followed by less entrenched deliberate nonfilers.

Characterizing passive nonfiling as the absence of a decision to file may be more fruitful than thinking about it as the result of any particular decision or action. The day-to-day concerns of life consume many of the cognitive resources available to adults at any given time. Infrequent events, such as filing your tax return, can easily be overlooked—even when initially salient—through repeated instances of "I'll get to it tomorrow" and other forms of avoidance (Anderson (2003)). Such procrastination can derive from many related concepts; we cover two such concepts and a potential solution below. We turn first to the planning fallacy.

The planning fallacy refers to the tendency to underestimate the time it will take to complete a task (Kahneman and Tversky (1977)). This is something we all have fallen prey to at one time or another (and it has certainly played a role in the writing of this section), and the consequences can vary as much as the intended task. In the case of needing to block some time to author a portion of a paper, the consequences are slight; however, when procrastination leads to, say, an unfiled tax return, the ramifications can be severe. One proposed mechanism by which the planning fallacy impacts task completion is through a failure to unpack the components of a task (Forsyth and Burt (2008); Kruger and Evans (2004)). However, breaking a task down into its pieces is not a cure for procrastination, nor does it prevent other psychological factors at play, such as moral licensing.

Moral licensing originated in the study of people behaving morally and later using that moral behavior as justification to behave immorally. Since then, the concept has broadened in scope to include, generally, any completed beneficial task being used as an excuse to do a less beneficial thing (Blanken, van de Ven, and Zeelenberg (2015)). Moral licensing in the context of health can be illustrated thusly: "I went to the gym this morning, so I've earned this extra-large order of mozzarella sticks." Considering this in the lens of this paper, this type of reasoning could easily be applied after any number of completed tasks to form a justification not to file one's taxes (or, rather, not to file one's taxes right now). "I got caught up on my other bills and paperwork; filing my taxes can wait for a different day." "I worked overtime this week; I'm justified in taking this weekend for myself and not filing my taxes." The possible list of completed taxing tasks (pun very much intended) that can provide an "out" for not filing a return is nearly endless. Justification comes in many forms, whether it is perfectly reasonable or a bit of a stretch; the outcome is the same—pushing off an undesirable task because you feel you have "earned it."

Now combine these factors. If you underestimate how long it will take to finish something and push it back repeatedly, it is easy to see how it could slip your attention entirely. Enter Just-in-Time Adaptive Interventions (JITAI). Developed out of health and educational psychology, JITAI strive to do what the name implies—reach the recipient with the right message at the right time to prompt the largest behavioral change

or maintenance of prior adaptive changes (Nahum-Shani, *et al.* (2014)). JITAI often make use of personalized information and electronic delivery mechanisms, which are beyond the scope of this paper, but the underlying concepts can still be harnessed for use in taxpayer outreach. Like many point-of-decision messages, well-timed outreach can motivate a behavioral change. Such techniques have been valuable in promoting behaviors, such as using stairs instead of elevators or making the most energy efficient decisions in one's home (Soler, *et al.* (2010); Intille (2002)). By issuing clear, nonthreatening reminders at a time in which filing a tax return is likely to be salient (i.e., proximal to major points in the tax season), we are likely to increase the odds of our message not only being read, but also being acted upon.

The above review gives a brief, nonexhaustive overview of some concepts that can inform the outreach the IRS develops to communicate with taxpayers. In this pilot test, we focused on developing clear, simple outreach messages and the timing of their delivery. This expands upon previous outreach research conducted at the IRS and other tax administrations.

Outreach and Reminders from Tax Administrations

This pilot study follows upon the filing reminder outreach established by Orlett, *et al.* (2017). Their study involved field tests using postcards or letters to promote voluntary filing compliance during the TY 2015 filing season among taxpayers who had prior filing delinquencies. They found that preemptively contacting these taxpayers can improve their future filing compliance, and there was a positive effect from multiple nudges. The results also suggested that receiving a letter may be more effective than receiving a postcard for some taxpayers. In their direction for further research, they stated the need to better understand the impact of the message in the outreach; a clearer message may be more effective in increasing taxpayer response. As noted above, this was incorporated into the current study.

Previous research has investigated the role of inattention and reminders in filing behavior and the take-up of the Earned Income Tax Credit (EITC). Gutyon *et al.* (2016) focused on prior nonfiling in TYs 2011 and 2012, with outreach during the 2014 and 2015 filing seasons. Individual taxpayers selected for this study were potentially eligible for the EITC during the years they were nonfilers (TYs 2011 and 2012). Taxpayers sampled from those specifications were randomly assigned to either a control group or one of six treatment groups. The treatment groups received outreach (informational postcard or brochure) with instructions on filing current and prior returns; the outreach also included information about the EITC. Reaching out to the treatment groups resulted in an increase to the filing rates among those receiving the EITC and among those who had a balance due, indicating the importance of outreach in promoting voluntary compliance even among taxpayers who had a tax liability. Furthermore, this study tested if the effects of a one-time contact would last into the following tax year, or if repeated reminders were necessary to promote EITC take-up and filing in general. The results indicated that inattention was a factor and recidivism was common among those who did not receive a reminder in the subsequent year, particularly among those who had had a balance due in the first test year.

Other studies conducted by tax administrations have varied the message being sent in mailed outreach. These have often tested different categories of behavioral insights with varying effectiveness (e.g., civic duty messaging, social norms messaging, and loss aversion via threat of penalties). The efficacy of specific messaging seems to depend on contextual factors, for example, social norms messaging was broadly effective in the United Kingdom, compared to its effectiveness with specific taxpayer segments in the United States (Behavioral Insights Team (2012); Blumenthal, Christian, and Slemrod (2001); Herlache, *et al.* (2018)). However, there is compelling evidence that reminders in general have a beneficial impact on promoting compliance (Kettle, *et al.* (2016); Orlett, *et al.* (2017)).

The present research provides insight into outreach and enforcement options available to the IRS. Building repeated contact into the design allowed us to assess the timing and combination of treatments that best prompted filing behavior. By escalating the language of the contact, we could also assess if progressively stricter language helped to move harder to reach nonfilers (i.e., those who did not respond to earlier treatment) into compliance where repeated "soft" contact may not. Moreover, this study moves beyond assessing the value of reminders into a direct comparison of enforcement and outreach. To our knowledge, this is the first study to include an established enforcement procedure and preemptive outreach as treatment arms in the same

study. Other studies have compared softer contact methods with enforcement (e.g. Collins, *et al.* (2018)), but the comparison focused on a single behavior. By measuring the impact on filing both prior-year delinquent returns and current-year returns, we were able to compare the tradeoffs of focusing on past noncompliance versus encouraging current and future compliance.

Method

Treatments

This study is the first and second wave of a larger, ongoing randomized control trial spanning three points of contact (waves) with taxpayers (see Appendix for treatments). We used mailed outreach to contact all three waves. Wave 1 was sent prior to the April 2018 filing deadline for Tax Year 2017 returns (treatment mailed April 2, 2018), Wave 2 was sent near the October 2018 filing extension deadline for 2017 tax returns (treatment mailed October 12, 2018), and Wave 3 was sent in December of 2018 (treatment mailed December 17, 2018).

Wave 1

TY 2016 delinquent return notice process: The RD notice process begins with a mailed notice informing taxpayers that IRS records indicate that they have an unmet filing requirement and the steps they need to take to address that requirement. Depending on their responses (or nonresponses) to the notice, taxpayers may receive an additional notice or progress to a Taxpayer Delinquency Investigation (TDI), which typically occurs within 6 to 8 weeks after the initial notice is sent.

Simple letter: This was a short letter focusing on reminding taxpayers to file their TY 2017 tax returns if they had not yet done so. The letter included a URL for the IRS Website and a toll-free customer service number for the IRS.

Simple postcard: The content of this postcard matched the simple reminder letter; it was a generic reminder to the taxpayers to file a TY 2017 return if they had not yet done so.

Complex letter: This letter contained the same TY 2017 filing reminder as the simple letter. In addition, it included information about filing prior-year returns and the necessary form (Form 4506-T) for obtaining prior-year tax information, such as W-2s and other tax documents.

Complex postcard: This postcard matched the complex letter's content: it contained the same TY 2017 filing reminder and additional information about filing prior-year returns.

Control: The control condition received a non-IRS postcard containing a public service announcement provided by a Federal agency partner.

Wave 2

Soft letter: This letter was similar to the simple letter in that it was a generic reminder to file a TY 2017 return; however, the wording used was more forceful than the letter sent in Wave 1. That is, it noted that the filing deadline had passed and had more sections with general IRS information (e.g., payment options).

Wave 3

TY 2017 Delinquent return notice process: Described above under Wave 1, TY 2016.

Soft letter: Described above (nonfiler sample only) under Wave 2.

The Wave 1 treatments allowed us to compare the impact of enforcement versus outreach on addressing past noncompliance and encouraging current-year compliance (nonfiler sample only). We were also able to

evaluate the effectiveness of the outreach format (postcard versus letter) and content (simple versus complex reminder) on filing behavior—both for the current filing season (TY 2017: nonfiler and stopfiler samples) and for delinquent returns (TY 2016: nonfiler sample only).

The additions of Waves 2 and 3 added greater nuance to what we were able to discern about the effectiveness of IRS written communications in prompting filing behavior (i.e., enforcement treatments via mailed notices versus the recently designed mailed outreach). We compared the impact of repeated treatments to a single treatment as well as the impact of repeated treatments with an escalating tone of severity to repeated soft treatments. Within the nonfiler sample, this design also provided a comparison of the timing of treatment, addressing whether earlier outreach was more effective at driving compliance.

Sample and Experimental Design

Nonfiler Sample

As mentioned above, the taxpayers sampled for this study were drawn from administrative data, in which they were identified as having a potential unmet filing requirement for TY 2016. Using the treatment effect of a similar study by Orlett, *et al.* (2017) as an estimate, we structured this study to be able to detect a 1.5-percent difference between the treatment conditions and the control condition (collapsing across treatments) and included a 10-percent oversample to address undeliverable mail concerns. Adding a buffer for unforeseen complications, we arrived at a minimum sample size of 5,000 per treatment condition when building the three-wave experimental design. When accounting for voluntary filing rates and filing in response to treatment, as well as accommodating the four preemptive treatment conditions included under the larger heading of preemptive outreach, we arrived at the design illustrated in Table 1.

TABLE 1. Nonfiler Experimental Design Across Waves 1, 2, and 3

Treatment Group	Sample Size	Wave 1 (April 2018)	Wave 2 (Oct. 2018)	Wave 3 (Dec. 2018)
1	5,000	TY 2016 RD start		
2	5,000	Reminder		
3	5,000		Soft letter	
4	5,000			Soft letter
5	5,000			TY 2017 RD start
6	5,000	Reminder	Soft letter	
7	10,000	Reminder	Soft letter	Soft letter
8	10,000	Reminder	Soft letter	TY 2017 RD notice start
9 (Control)	15,000	Control postcard		
Total	65,000			

NOTE: Taxpayers who had filed their TY 2017 return (as determined by the latest data available prior to mailing) were removed from subsequent treatment. RD stands for return delinquency.

Taxpayers were randomly assigned to either one of the treatment groups or to the control group. Within Wave 1, group one was assigned to the delinquent return process treatment; groups two, six, seven, and eight were then further randomly assigned to a specific preemptive outreach treatment group, resulting in the sample sizes noted in Table 2.

TABLE 2. Nonfiler Wave 1 Treatment Groups, by Sample Size

Condition	Sample Size
Delinquent return notice process	5,000
Simple letter	7,500
Simple postcard	7,500
Complex letter	7,500
Complex postcard	7,500
Control	15,000

Groups three, six, seven, and eight received treatment at Wave 2 in the form of a soft letter. Groups four, five, seven, and eight received treatment at Wave 3. Groups four and seven received a soft letter; groups five and eight entered the TY 2017 RD notice process. Taxpayers who had filed prior to the mailing dates were excluded from the mailing lists.

Stopfiler Sample

As mentioned above, the research team developed a model leveraging existing IRS administrative data to predict which taxpayers may become stopfilers. The model was trained on TY 2012 and 2013 data, which was validated using TY 2014 data (predicting TY 2015 filing behavior). The training data had roughly 7.7 million records and 287 variables. After applying a variety of variable selection techniques (factor analysis, simple correlation, and LASSO), the final logistic regression model included 15 variables, resulting in a reasonable goodness of fit (AUC = .80). The model generates a probability score of becoming a stopfiler in the subsequent tax year. Because late-filers (taxpayers who file but miss the April filing deadline) can also benefit from outreach, we included them when assessing the 2014 validation.

Of the top 1 percent of taxpayers identified by our stopfiler model, roughly 21 percent had been actual stopfilers in TY 2015 and about 40 percent had been either stopfilers or late-filers. Within the top 1 percent, we identified 12 percent of all the expected stopfilers in the population. Expanding our view to the top 5 percent as identified by the model, we captured roughly a third of all expected stopfilers. When applying this model to TY 2016 in predicting TY 2017 filing behavior, we used the top 5 percent as the threshold for the sample. Of the top 5 percent, 26,500 were randomly sampled and then randomly assigned to a treatment or control condition. Including additional information about filing prior-year tax returns was not relevant to this sample, so taxpayers assigned to treatment received either the simple letter or simple postcard (see Table 3).

TABLE 3. Stopfiler Experimental Design Across Waves 1, 2, and 3

Treatment Group	Sample Size	Wave 1 (April 2018)	Wave 2 (Oct. 2018)	Wave 3 (Dec. 2018)
Wave 1 letter	10,000	Reminder	Soft letter	TY 2017 RD start
Wave 1 postcard	10,000	Reminder	Soft letter	TY 2017 RD start
Control	6,500	Control postcard		

NOTE: Taxpayers who had filed their TY 2017 return (as determined by the latest data available prior to mailing) were removed from subsequent treatment. RD stands for return delinquency.

Analysis and Modeling

Nonfiler Treatments

We used a logistic regression model to estimate the treatment effects of the contacts on three dichotomous outcomes: 1) taxpayers filing their TY 2016 delinquent returns, 2) taxpayers filing their TY 2017 income tax returns (the Wave 1 outcome is filing or filing for an extension; Waves 2 and 3 are filing), and 3) taxpayers filing

their TY 2018 income tax returns or filing for extensions to file. The Wave 1 outcomes were observed until just prior to the Wave 2 mailing in the beginning of October 2018; the Wave 2 outcomes were observed until roughly the time of the Wave 3 mailing, which was toward the end of calendar-year 2018. Wave 3 outcomes were observed through May 2019.

We removed from the analysis taxpayers who filed before the initial treatments were mailed. Table 4 provides the number of taxpayers who filed their TY 2017 return before treatment, and Table 5 provides the same for TY 2016.

TABLE 4. Nonfilers Who Filed Returns Before Wave 1 Treatment, TY 2017

Treatment	Sample Size	Filed TY 2017 Before Treatment	Percent Filed Before Treatment
TY 2016 return delinquency notice process	5,000	302	6.0%
Simple letter	7,500	494	6.6%
Simple postcard	7,500	492	6.6%
Complex letter	7,500	471	6.3%
Complex postcard	7,500	458	6.1%
Soft letter (Wave 2 only)	5,000	305	6.1%
Soft letter (Wave 3 only)	5,000	252	5.0%
TY 2017 return delinquency notice process (Wave 3 only)	5,000	323	6.5%
Control postcard	15,000	840	5.6%

SOURCE: IRS Compliance Data Warehouse, Individual Return Transaction File. Data extracted May 2019.

TABLE 5. Nonfilers Who Filed TY 2016 Returns Before Wave 1 Treatment

Treatment	Sample Size	Filed TY 2016 Before Treatment	Percent Filed Before Treatment
TY 2016 return delinquency notice process	5,000	422	8.4%
Simple letter	7,500	657	8.8%
Simple postcard	7,500	629	8.4%
Complex letter	7,500	680	9.1%
Complex postcard	7,500	654	8.7%
Soft letter (Wave 2 only)	5,000	459	9.2%
Soft letter (Wave 3 only)	5,000	397	7.9%
TY 2017 return delinquency notice process (Wave 3 only)	5,000	428	8.6%
Control postcard	15,000	1,343	9.0%

SOURCE: IRS Compliance Data Warehouse, Individual Return Transaction File. Data extracted May 2019.

We also removed taxpayers from the regression analysis (in addition to those who filed before the treatments were sent) when their mailed outreach was identified as undeliverable. We sent postcards unrelated to tax to both the control group and to taxpayers who had been sent the delinquent return notices.[2] As is evident in Table 6, the rate of undeliverable mail is somewhat higher in the control group and the delinquent return notice group. We attributed this to differences in how updated address information was used when the control group postcard was sent. While we can identify most of the undeliverable addresses in our control group, there is some inconsistency. Thus, we included statistical controls for the likelihood to be undeliverable.

[2] We tracked undeliverable for delinquent return notices via the unrelated postcards in an attempt to make tracking consistent with other treatments and the control group.

TABLE 6. Nonfiler Undeliverable Mail in Waves 1, 2, and 3, TYs 2016 and 2017

Treatment	Wave	Sample Size	Number Undelivered	Percent Undeliverable
TY 2016 return delinquency notice process	1	5,000	847	16.9%
Simple letter	1	7,500	1,113	14.8%
Simple postcard	1	7,500	1,010	13.5%
Complex letter	1	7,500	1,063	14.2%
Complex postcard	1	7,500	925	12.3%
Soft letter (Wave 2 only)	2	5,000	620	12.4%
Soft letter (Wave 3 only)	3	5,000	781	15.6%
TY 2017 return delinquency notice process (Wave 3 only)	3	5,000	854	17.1%
Control postcard	1	15,000	2,516	16.8%

NOTE: Mail returned as undeliverable was tracked via a unique identifier included in the mailing address, which was scanned and recorded upon receipt.
SOURCE: Results of the outreach.

We estimate treatment effects as follows. Let U be equal to 1 if a taxpayer's mail was returned as undeliverable. We model the probability of being undeliverable as

$$P(U = 1) = F(Z\beta_u + T\alpha),$$

where Z is a vector of case characteristics, and T is a vector of the five treatment dummies, excluding the control group, and F is the logistic distribution function. We can then calculate the control group estimated probability of being undeliverable as

$$U = F(Z\hat{\beta}_u).$$

For Wave 1, we then estimate filing response model for tax year t as

$$P(F_t = 1) = F(X\beta_t + T\alpha_t + \delta_t U),$$

where β_t, α_t, and δ_t are parameter vectors to be estimated and t represents either the current tax year or the delinquent tax year.

We can then calculate average treatment effects for treatment j as

$$\partial P(F_t = 1) \Big/ \partial T_j = \frac{1}{n}\sum_i \alpha_{jt} f(X_i\beta_t).$$

For the combined Wave 1 and Wave 2 treatments, we estimate the filing response model for tax year t as

$$P(F_t) = F(X\beta_t + T_1\alpha_{1t} + T_2\alpha_{2t} + T_{1*}T_2\alpha_{12t} + \delta_t U),$$

where β_t, α_{it}, and δ_t are parameter vectors to be estimated, t represents either the current tax year or the delinquent tax year, T_1 is a vector of the five Wave 1 treatments, and T_2 is the Wave 2 treatment, excluding the control group.

For the combined Wave 1, 2, and 3 treatments, we estimate the filing response model for tax year t as

$$P(F_t) = F(X\beta_t + T_1\alpha_{1t} + T_2\alpha_{2t} + T_3\alpha_{3t} + T_{1*}T_2\alpha_{12t} + T_{2*}T_3\alpha_{23t} + T_{1*}T_{2*}T_3\alpha_{123t} + \delta_t U),$$

where β_t, α_{it}, and δ_t are parameter vectors to be estimated, t represents either the current tax year, delinquent tax year, or the subsequent tax year, T_1 is a vector of the five Wave 1 treatments, T_2 is the Wave 2 treatment, and T_3 is a vector of the two Wave 3 treatments, excluding the control group.

Stopfiler Treatments

We used a logistic regression model to estimate the treatment effects of the contacts on two dichotomous outcomes: taxpayers filing their TY 2017 income tax return (the Wave 1 outcome is filing or filing for an extension;

Wave 2 is filing) and taxpayers filing their TY 2018 income tax return or filing for an extension. The outcomes of the three waves were, as in the nonfiler sample, observed through roughly the beginning of October 2018, the end of December 2018, and the end of May 2019, respectively.

As with the nonfiler analysis, we removed from the analysis taxpayers who filed before the Wave 1 treatments were mailed and where contact was identified as undeliverable. Table 7 provides the numbers for both of these conditions. Similar to the nonfiler sample, the rate of undeliverable mail is somewhat higher in the control group than the treatment groups. We again attribute differences in how updated the address information was between the treatment and control groups. While we can identify most of the undeliverable addresses in our control group, there is some inconsistency; we include statistical controls for the likelihood to be undeliverable.

TABLE 7. Stopfilers Who Filed TY17 Returns Before Treatment and Undeliverable Mail

Treatment	Sample Size	Filed TY 2017 Before Treatment	Percent Filed Before Treatment	Number Undelivered	Percent Undeliverable
Simple letter	10,000	2,872	28.7%	520	5.2%
Simple postcard	10,000	2,863	28.6%	644	6.4%
Control	6,500	1,930	29.7%	583	9.0%

NOTE: Mail returned as undeliverable was tracked via a unique identifier included in the mailing address, which was scanned and recorded upon receipt.
SOURCE: Results of the outreach.

We estimate treatment effects as follows. Let U be equal to 1 if a taxpayer's mail was undeliverable. We model the probability of being undeliverable as

$$P(U = 1) = F(Z\beta_u + T\alpha),$$

where Z is a vector of case characteristics and T is a vector of the two treatment dummies, excluding the control group, and F is the logistic distribution function. We can then calculate the control group estimated probability of being undeliverable as

$$U = F(Z\hat{\beta}_u).$$

For Wave 1, we then estimate filing response model for TY 2017 as

$$P(F_t = 1) = F(X\beta + T\alpha + \delta U),$$

where β, α, and δ are parameter vectors to be estimated. We can then calculate average treatments effects for treatment j as

$$\partial P(F_t = 1) \Big/ \partial T_j = \frac{1}{n}\sum_i \alpha_{jt} f(X_i \beta_t) \ .$$

For the stopfiler treatments, all treatment groups receive the same treatments in Wave 2 and Wave 3. Thus, for Waves 2 and 3, we estimate a similar filing response model for TY 2017 as specified above (and TY 2018 with Wave 3), but the outcome window is extended to the end of Calendar Year 2018 for Wave 2 and until the end of May 2019 for Wave 3.

Results and Interpretation

Nonfilers

Descriptives

For current-year filings, we found that at least 37 percent of the taxpayers in Wave 1 had either filed or filed for an extension of their TY 2017 return after treatment (excluding those where the treatment was undeliverable).

The simple reminder letter had the highest rate at 41.9 percent, followed by the complex reminder letter at 39.9 percent. The other Wave 1 treatments trailed the letter outreach. This is echoed in Wave 2 where we focus specifically on filing rather than also including filing for an extension; again we see the simple letter outperforming the other treatment options in securing the most returns at 31.8 percent. Adding the soft notice to early outreach in Wave 1 does not appear to have much of an impact, but receiving the soft notice in October as a first contact did bring in more returns than the control condition (29.4 percent filed returns versus 28.8 percent in the control condition).

Regarding TY 2016, the RD notice process stands out as the clear leader in securing past returns in both Waves 1 and 2 (21.7 percent and 25.3 percent, respectively). This was followed by the simple and complex reminder letters, the latter having additional information about filing prior-year returns, at 16.1 percent and 16.0 percent, respectively, in Wave 1 and 19.1 percent and 18.7 percent, respectively, in Wave 2. At Wave 2, the RD notice process, which also included additional contact for some nonfilers, brought in roughly 7.0 percent more TY 2016 returns than the control condition, while the simple letter hovered around 1 percent additional returns filed.

Table 8 provides Wave 1 figures by treatment group for TYs 2017 and 2016 filing behavior. Table 9 shows updated results for Wave 2. Table 10 shows the same type of information updated for Wave 3, where we can see a similar pattern of filing percentages. The TY 2016 RD notice process continues to lead in securing TY 2016 returns and the simple letter leads in securing TY 2017 returns. Moving into TY 2018, we see descriptive evidence that early outreach can prompt filing subsequent year returns.

TABLE 8. Nonfiler Wave 1 Filings or Extensions To File After Treatment, TYs 2017 and 2016

Treatment	TY 2017 Sample*	TY 2017 Filing or Extension	Percent TY 2017 Filing or Extension	TY 2016 Sample*	TY 2016 Filing	Percent TY 2016 Filing
TY 2016 return delinquency notice process	3,909	1,482	37.9%	3,767	816	21.7%
Simple letter	5,953	2,497	41.9%	5,760	927	16.1%
Simple postcard	6,062	2,360	38.9%	5,900	834	14.1%
Complex letter	6,040	2,408	39.9%	5,792	926	16.0%
Complex postcard	6,185	2,323	37.6%	5,955	885	14.9%
Control	11,820	4,356	36.9%	11,269	1,756	15.6%

*Excluding undeliverables and taxpayers who filed before Wave 1 treatment.
SOURCE: IRS Compliance Data Warehouse, Individual Return Transaction File. Data extracted May 2019.

TABLE 9. Nonfiler Wave 2 Filings After Treatment, TYs 2017 and 2016

Treatment (First Contact)	TY 2017 Sample Size*	TY 2017 Filing	TY 2017 Filing Percentage	TY 2016 Sample Size*	TY 2016 Filing	TY 2016 Filing Percentage
TY 2016 return delinquency notice process	3,909	1,207	30.9%	3,767	952	25.3%
Simple letter	5,953	1,896	31.8%	5,760	1,101	19.1%
Simple postcard	6,062	1,744	28.8%	5,900	1,011	17.1%
Complex letter	6,040	1,835	30.4%	5,792	1,081	18.7%
Complex postcard	6,185	1,745	28.2%	5,955	1,041	17.5%
Soft notice (Wave 2 only)	4,075	1,200	29.4%	3,932	708	18.0%
Control	11,820	3,400	28.8%	11,269	2,092	18.6%

*Excluding undeliverables and taxpayers who filed before Wave 1 treatment.
SOURCE: IRS Compliance Data Warehouse, Individual Return Transaction File. Data extracted May 2019.

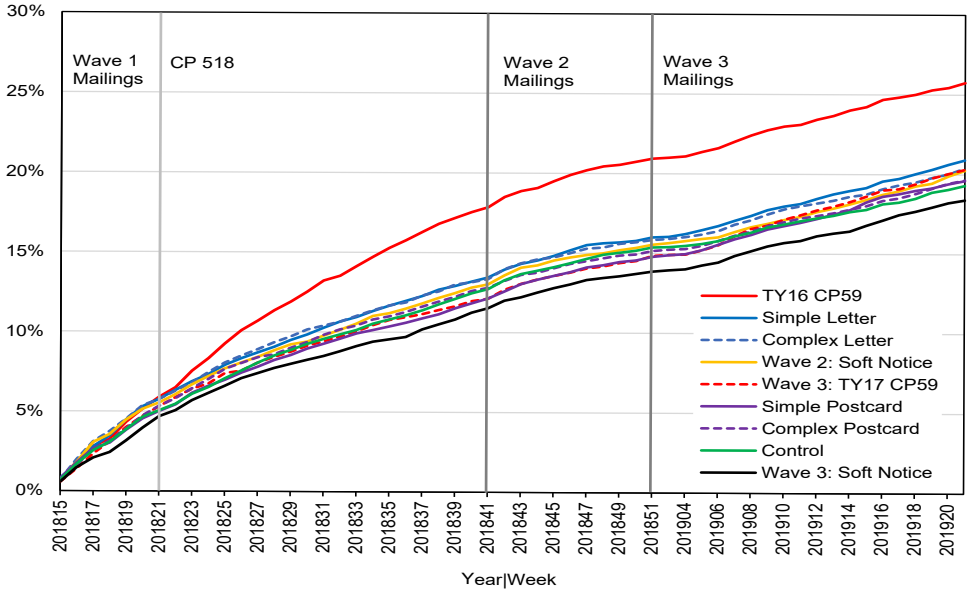
TABLE 10. Nonfiler Wave 3 Filings After Treatment, TYs 2018, 2017, and 2016

Treatment (First Contact)	TY 2018 Sample Size*	TY 2018 Filing or Extension Percentage	TY 2017 Sample Size*	TY 2017 Filing Percentage	TY 2016 Sample Size*	TY 2016 Filing Percentage
TY 2016 return delinquency notice process	4,153	45.6%	3,909	37.5%	3,767	30.9%
Simple letter	6,387	46.2%	5,953	38.7%	5,760	24.6%
Simple postcard	6,490	45.4%	6,062	35.7%	5,900	22.6%
Complex letter	6,437	44.2%	6,040	37.2%	5,792	23.8%
Complex postcard	6,575	43.2%	6,185	34.7%	5,955	22.5%
Soft notice (Wave 2 only)	4,380	43.8%	4,075	35.7%	3,932	23.3%
Soft notice (Wave 3 only)	4,219	43.2%	3,967	35.0%	3,833	21.8%
TY 2017 return delinquency notice process (Wave 3 only)	4,146	45.9%	3,892	36.9%	3,766	24.5%
Control	12,484	43.5%	11,820	34.2%	11,269	23.2%

*Excluding undeliverables and taxpayers who filed before Wave 1 treatment.
SOURCE: IRS Compliance Data Warehouse, Individual Return Transaction File. Data extracted May 2019.

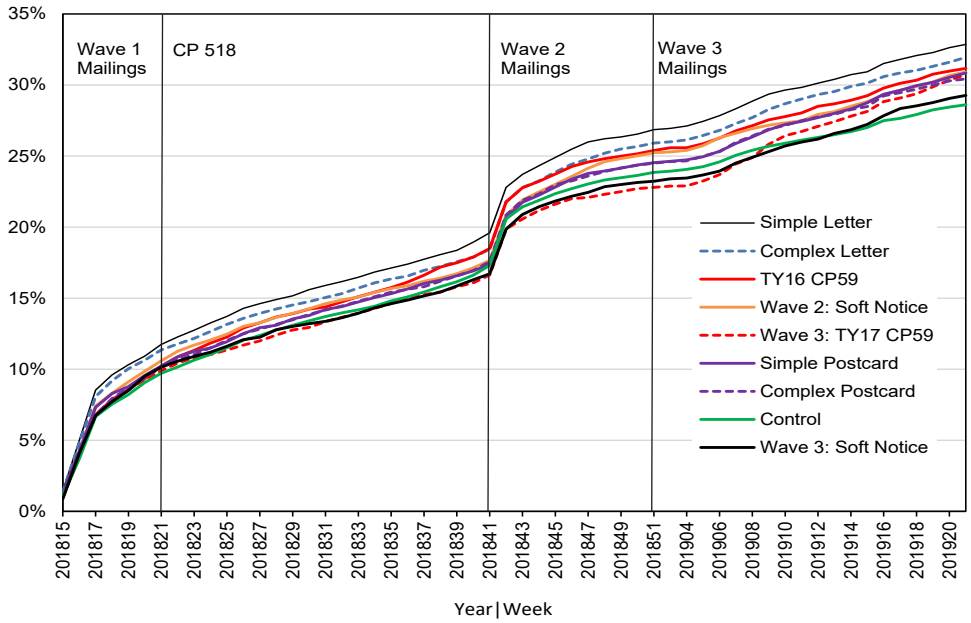
The timing of a taxpayer's behavior following treatment also provides some insight into the individual's response to the contact. Figures 1-3 provide the cumulative weekly filings for Tax Years 2016, 2017, and 2018 individual tax returns through the end of May 2019, displayed by the first treatment received. For the delinquent TY 2016 returns, we found that taxpayers who received the delinquent return notice treatment had the largest increase in filing, which appeared to accelerate around the end of May 2018 (cycle 201821). This corresponds with another IRS action triggered by the delinquent return notice process—the delivery of the final return delinquency notice would have been issued around this time for those who had not yet responded. The final notice could be perceived as more severe than the first notice (the notice issued as part of the study design in Wave 1), as it contained more information on enforcement actions and other consequences of not filing. For TY 2017 filings, we found that taxpayers who received a simple reminder letter at Wave 1 filed the most returns over time, and the relative difference between the rate of filing among treatment conditions remained fairly constant. Extending our view to TY 2018, we see that most treatment conditions appear to influence the subsequent tax year as well.

FIGURE 1. Percentage of Nonfilers Filing Returns Across Waves 1, 2, and 3 (April 2018–May 2019), by Week, TY 2016



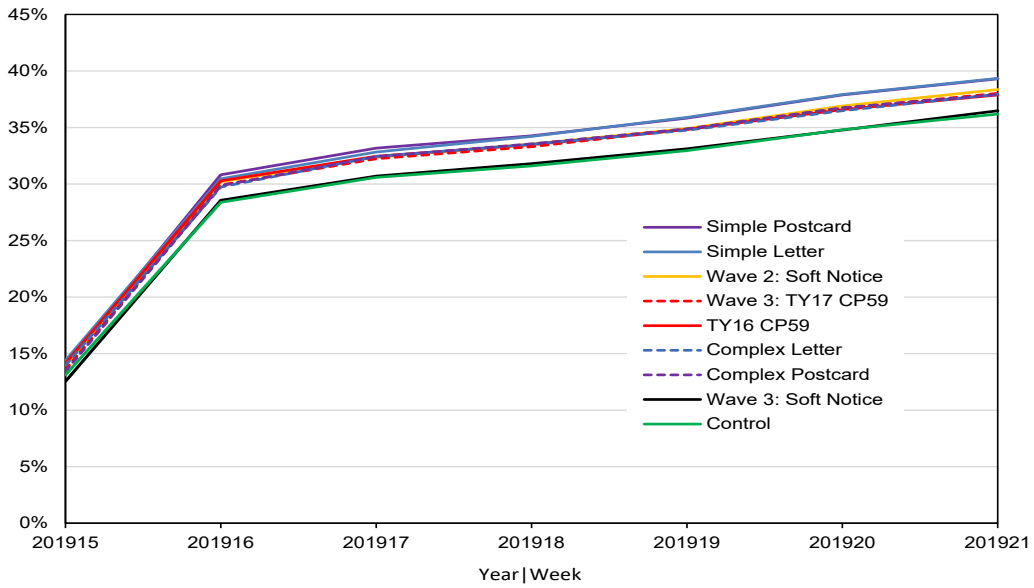
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

FIGURE 2. Percentage of Nonfilers Filing Returns Across Waves 1, 2, and 3 (April 2018–May 2019), by Week, TY 2017



NOTE: The spike around Cycle/Week 42 corresponds to the extension due date in October 2018.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

FIGURE 3. Percentage of Nonfilers Filing Returns Across Waves 1, 2, and 3 (April 2019–May 2019), by Week, TY 2018



SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

Model Results

The regression model estimates are provided in the Appendix. We estimated the models with and without the undeliverable control. While some model estimates change (e.g., the "secured return" risk score), the estimates remained similar for the treatments, especially for the Tax Year 2017 filings. Thus, for simplicity, we focused our discussion on the treatment effects from the models without the undeliverable control.

We start by presenting the results from Waves 1 and 2, focusing first on the delinquent tax year (2016), then turning to the current tax year (2017). The Waves 1 and 2 estimated treatment effects are reported in Table 11. By breaking the results apart by wave for the delinquent and current tax years, we emphasize the chronological nature of this pilot and present the results as they unfolded over time. It is important to keep in mind that Wave 1 results are based on observed outcomes from the start of the study until the Wave 2 treatments begin. Wave 2 outcomes are responses observed from the start of the study up until the Wave 3 treatments begin.

With the results from the first two waves covered, we then layered on our final analyses with outcomes tracked through May of 2019 (Wave 3 results). This includes how our treatment effects extended into the subsequent tax year (2018). Ending on the overarching view of treatment effects across three waves of contact and 3 tax years underscores the importance of considering the multidimensional impact of treatments; taxpayer behavior is not an event isolated to one tax year or one period in time.

TABLE 11. Nonfiler Marginal Effects for Waves 1 and 2 From Models Without Undeliverable Controls, TYs 2016 and 2017

Treatments	Wave 1 TY 2016	Wave 2 TY 2016	Wave 1 TY 2017	Wave 2 TY 2017
TY 2016 return delinquency notice process	0.031*	0.056*	0.009	0.018*
Simple letter	0.008	0.013	0.050*	0.035*
Simple postcard	-0.009	-0.011	0.015*	0.009
Complex letter	0.009‡	0.009	0.034*	0.019‡
Complex postcard	0.001	-0.004	0.006	0.003
Soft notice (Wave 2 only)	NA	0.004	NA	0.024*
Additional from soft notice after Wave 1 letter (either version)	NA	-0.002	NA	0.003
Additional from soft notice after Wave 1 postcard (either version)	NA	0.004	NA	0.006

NOTES: Wave 1 outcomes are TY17 filing or filing for an extension and TY16 filing; Wave 2 outcomes are TY17 filing and TY16 filing. * Indicates significance at the 95-percent level; ‡ indicates significance at the 90-percent level. N/A=Not Applicable.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

Focusing first on the filing of delinquent returns (TY 2016), the RD notice process shines as the most effective. This process increased the filing of individual tax returns for TY 2016 by 3.1 percentage points by the end of Wave 1, and 5.6 percentage points by the end of the 2018 calendar year. A distant second at Wave 1 was the complex letter, with an effect roughly one-third the size, followed by the simple letter at Wave 2, with an effect roughly one-fifth the size of the RD notice process. At Wave 1, both the complex postcard and complex letter were more effective than the simple versions in securing delinquent returns; however, only the effect of the complex letter was significantly different from the control.

At Wave 2, only the RD notice process was a significant predictor of securing delinquent returns. This pattern suggests that the benefit of the Wave 1 outreach in addressing delinquent returns was in prompting nonfilers to file their TY 2016 returns earlier than they would have otherwise; neither the complex letter nor the complex postcard treatments secured significantly more delinquent returns overall. The RD notice process, on the other hand, secured more delinquent returns than would have otherwise been submitted; however, it did not secure as many current year (TY2017) returns as the simple reminder letter, suggesting that there is a tradeoff in addressing past and current compliance.

Turning to TY 2017 results, the letters provided a significant increase in the filing of current tax year returns at both Wave 1 and Wave 2. The simple reminder letter increased filing behaviors (filing or filing for an extension) by 5 percentage points at Wave 1; the complex letter, which contained additional information on

filing delinquent returns, increased TY 2017 filing by roughly 3.4 percentage points at Wave 1. Both letters continued to increase the filing of current tax returns at Wave 2. The simple reminder letter increased filing by 3.5 percentage points; the complex letter increased filing by roughly 1.9 percentage points (marginally significant). The treatment effect appears to decline somewhat over time (keep in mind that Wave 1 results included extensions as well as filed returns), but is persistent nonetheless. The simple letter continued to have the largest effect on current filing compliance through the end of the calendar year. In general, the correspondence with information about delinquent returns had a smaller treatment effect on securing TY 2017 returns, whether it was in a letter or a postcard. Turning to the soft notice, we see a significant increase of 2.4 percentage points when the soft notice was the only treatment received. This suggests that there is an opportunity cost of delaying treatment, at least in the short run. However, the addition of the soft notice to early outreach did not significantly increase the number of returns submitted to the IRS beyond what was accounted for by the Wave 1 treatment.

The effect of the RD process on filing TY 2017 returns was the smallest of the Wave 1 treatments and was not significantly different from the control. Taken together with the effects of the complex letter and postcard, this suggests that, to some degree, references to past unfiled returns are not helpful in getting taxpayers to return to filing compliance in the current tax year—at least not when contact is made close to the filing due date. The addition of Wave 3 data sheds additional light on this question, particularly around the influence of enforcement efforts on prior versus current and future filing. Table 12 presents the Wave 3 treatment effects for Tax Years 2016, 2017, and 2018.

TABLE 12. Nonfiler Marginal Effects for Wave 3 From the Model Without an Undeliverable Control, TYs 2016, 2017, and 2018

Treatments	Wave 3 TY 2016	Wave 3 TY 2017	Wave 3 TY 2018
TY 2016 return delinquency notice process	0.067*	0.029*	0.022*
Simple letter	0.019‡	0.036*	0.041*
Simple postcard	-0.004	0.016	0.025*
Complex letter	0.013	0.022‡	0.021‡
Complex postcard	-0.002	0.005	0.003
Soft notice (Wave 2 only)	0.015‡	0.036*	0.033*
Additional from Wave 2 soft notice after Wave 1 letter (either version)	0.011	0.016	-0.008
Additional from Wave 2 soft notice after Wave 1 postcard (either version)	0.001	0.001	-0.005
Soft notice (Wave 3 only)	0.003	0.042*	0.026*
Additional from Wave 3 soft notice after Wave 1 letter (either version) and Wave 2 soft notice	-0.012	-0.004	-0.0002
Additional from Wave 3 TY 2017 return delinquency notice process after Wave 1 letter (either version) and Wave 2 soft notice	-0.004	0.020	0.014
Additional from Wave 3 soft notice after Wave 1 postcard (either version) and Wave 2 soft notice	0.007	0.018	0.020
Additional from Wave 3 TY 2017 return delinquency notice process after Wave 1 postcard (either version) and Wave 2 soft notice	0.009	0.029*	0.023‡
Soft notice (Wave 3 only)	0.003	0.042*	0.026*
TY 2017 return delinquency notice process (Wave 3 only)	0.013‡	0.034*	0.023*

NOTES: TY16 and TY17 outcomes are filing; TY18 refers to filing or filing for an extension. * Indicates significance at the 95-percent level; ‡ indicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

Extending our view of the treatment effects to include Wave 3 treatments, we see an affirmation of some of our earlier results. Beginning with TY 2016 filing, the TY 2016 RD notice process continues to be the front-runner in securing additional returns at a 6.7-percentage-point increase, followed by marginally significant increases among those who received a simple letter in Wave 1 or a soft letter in Wave 2. It also seems that starting a nonfiler in a TY 2017 RD notice process yielded some gains regarding the previous year's (TY 2016) delinquent return (a 1.3-percentage-point increase). However, the TY 2017 RD process, understandably, shows greater gains within the year it was targeting.

The majority of the treatments in this pilot study focused on TY 2017, but Wave 3 was the first time we had a direct comparison of enforcement and softer outreach referencing the same tax year and being sent at the same time. Focusing on the nonfilers who received either the Wave 3 soft notice or the Wave 3 TY 2017 RD process, we see that the soft notice brought in more Tax Year 2017 returns by the end of May 2019 than the enforcement process brought in during this same period (increases of 4.2 and 3.4 percentage points, respectively).

The TY 2017 RD treatment effect on TY 2017 returns echoes the effect size we saw among TY 2016 returns at Wave 1 for the TY 2016 RD process (the TY 2016 RD process increased TY 2016 filing at Wave 1 by 3.6 percentage points; the TY 2017 RD process increased TY 2017 filing by 3.4 percentage points in a similar length of time). If this pattern continues, we might expect the TY 2017 RD treatment to gain traction over the course of the process; as additional notices go out, the treatment effect could parallel the increases we've seen over time in the TY 2016 RD condition. However, in viewing the data through May of 2019, we saw several informational outreach treatments bringing in returns at a similar level to the RD start (the simple letter and the Wave 2 soft notice both increased TY 2017 filings by 3.6 percentage points). These results highlight the importance of considering alternative treatment routes outside of the traditional enforcement methods that may free up additional resources that can be directed toward cases that are harder to reconcile.

Likewise, it is important to consider the combination of treatments available in promoting filing compliance. By considering the cumulative impact of the Wave 1 simple outreach with the follow-up treatments in Waves 2 and 3, we get a sense of how mixed the results of repeated treatments can be. Generally speaking, following two waves of informational outreach with a start in the RD process secured additional returns for the tax year being addressed as well as the prior and subsequent tax years. This makes sense, as the escalating nature of the treatment language is likely to sway reticent taxpayers who may not be convinced by repeated softer contacts. Specifically, looking at a Wave 1 simple letter followed by a soft notice in Wave 2 and an RD start in Wave 3 seems to be the most promising combination of treatments when considering the impact across 3 tax years, with a cumulative 2.6-percentage-point increase in TY 2016 returns submitted to the IRS, a 7.2-percentage-point increase in TY 2017 returns, and a 4.7-percentage-point increase in TY 2018 returns or extensions. The same level of promise is not shown in repeated informational outreach. With lighter-touch outreach, it seems that two contacts (a Wave 1 simple letter plus a Wave 2 soft notice) may be an effective route to bringing in additional returns across the 3 tax years being studied (with 3.0-, 5.2-, and 3.3-percentage-point increases across TYs 2016, 2017, and 2018, respectively). While only the incremental effects that were noted in Table 12 as significant or marginally significant offered increases in their own right, it is important to consider them as part of a package of correspondence when weighing different options for treating nonfiling.

Another way to consider this is to translate the above effects to estimating how many returns would have been submitted to the IRS if we had contacted 100,000 taxpayers in that manner (Table 13). By projecting the results this way, we can get a better sense of the results in terms of the outcomes of interest, which are either secured returns or extensions to file. If we consider purely the projected amount of filed tax returns or extensions to file, the TY 2016 RD notice process is among the leaders, but the complexity of the landscape requires a closer look. The 2016 RD process is securing more returns, but the gains are being driven by TY 2016 at the expense of potential TY 2017 returns and TY 2018 returns or extensions. Likewise, a start in any RD process comes at a much higher administrative resource cost and taxpayer burden.

The choice in how to address noncompliance needs to take many factors into account, in particular the time and resources required (for example, the time and cost of mailed reminders versus the RD notice process and the calls received by the IRS in response to either treatment) and the priority outcome (securing

delinquent tax returns versus current year tax returns). One factor that should not be undervalued in considering how to address noncompliance is timing. By viewing the results of this pilot across 3 tax years, we can see the value of a single, well-timed piece of mail. A simple letter sent just prior to the TY 2017 filing deadline yielded gains across 3 tax years. Likewise, a soft notice sent around the TY 2017 extension deadline prompted nonfilers to act on their TY 2016, 2017, and 2018 filing obligations.

TABLE 13. Estimated Impacts From Contacting 100,000 Nonfilers with Treatment Based Wave 3 Results, TYs 2016, 2017, and 2018

Treatments	Wave 3 TY 2016	Wave 3 TY 2017	Total: Wave 3 TYs 2016–2017*	Wave 3 TY 2018	Total: Wave 3 TYs 2016–2018‡
TY 2016 return delinquency notice process	6,700	2,900	9,600	2,200	11,800
Simple letter	1,900	3,600	5,500	4,100	9,600
Plus wave 2 soft notice	3,000	5,200	8,200	3,300	11,500
Plus Wave 3 soft notice	1,800	4,800	6,600	3,300	9,900
Plus Wave 3 RD process	2,600	7,200	9,800	4,700	14,500
Simple postcard	-400	1,600	1,200	2,500	3,700
Plus Wave 2 soft notice	-300	1,700	1,400	2,000	3,400
Plus Wave 3 soft notice	400	3,500	3,900	4,000	7,900
Plus Wave 3 RD process	600	4,600	5,200	4,300	9,500
Soft notice only (Wave 2)	1,500	3,600	5,100	3,300	8,400
Soft notice only (Wave 3)	300	4,200	4,500	2,600	7,100
TY 2017 return delinquency notice process only (Wave 3)	1,300	3,400	4,700	2,300	7,000

NOTES: TY 2016 and TY 2017 outcomes are filing; TY 2018 refers to filing or filing for an extension. *totals reflect secured returns; ‡totals reflect secured returns or extensions. Using analyses reported for Wave 3, marginal effects were projected to 100,000 contacts (see the Appendix and Table 12 above).
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

Setting aside combinations of outreach for now, if we decide to send a single piece of mail, then this pilot study suggests that good timing can boost its intended effect. Tying our outreach to when tax is likely to be more salient naturally (at major points in the tax cycle) seems to increase the chances of the message hitting home, so-to-speak (i.e., as compared to a soft notice sent at another point in time that is less associated with taxes, for example, near the end of the calendar year). While there is a lot of information to parse from the results of the nonfiler portion of this pilot study, the general take-home message is that low-cost outreach can be a viable way to treat nonfiling. Taken together, these results leave tax administrators better able to weigh some of the options available to them. Next, we turn to an even more proactive approach for dealing with potential noncompliance: addressing it before it starts.

Potential Stopfilers

Descriptives

Recall that the individuals in this sample were drawn from taxpayers who filed in Tax Year 2016, but who were predicted to be at the highest risk of stopfiling in Tax Year 2017 (recall also that we use stopfiler as a shorthand for potential stopfiler). These taxpayers were either assigned to one of two treatments in Wave 1 (a simple letter or a postcard), or were part of a no-treatment control. All taxpayers treated in Wave 1 were then contacted with a soft notice in Wave 2 and with the RD notice process in Wave 3 if they had not yet filed. We found at least 63 percent of the stopfilers in Wave 1 had filed and at least 16 percent had filed for an extension to file their TY 2017 tax return after receiving treatment. Both treatments yielded an increase in filing behavior (filing or

filing for an extension), but neither rose to the level of statistical significance. At Wave 2, we see the reminder postcard (with a soft letter contact) bringing in more returns than the letter combination at 79.7 percent. This pattern continued at Wave 3, where those in the treatment conditions who had not yet filed were started in the TY 2017 RD process.

TABLE 14. Stopfiler Filing Behavior at Waves 1, 2, and 3, TY 2017

Treatment	TY 2017 Sample*	Percent Wave 1 Filing	Percent Wave 1 Extension	Percent Wave 2 Filing	Percent Wave 3 Filing
Reminder letter (Plus Waves 2 and 3)	6,757	63.6%	17.7%	78.3%	82.8%
Reminder postcard (Plus Waves 2 and 3)	6,697	64.9%	16.1%	79.7%	84.5%
Control	4,204	63.9%	16.8%	79.7%	83.5%

NOTES: The Wave 1 outcome is TY 2017 filing or filing for an extension; the Wave 2 outcome is TY 2017 filing. *Excluding undeliverables and taxpayers who filed before Wave 1 treatment.
SOURCE: IRS Compliance Data Warehouse, Individual Return Transaction File. Data extracted May 2019.

TABLE 15. Stopfiler Filing Behavior at Wave 3, TY 2018

Treatment	TY 2018 Sample*	Percent Filing	Percent Extension
Reminder letter (Plus Wave 2 soft notice)	9,480	59.2%	17.2%
Reminder postcard (Plus Wave 2 soft notice)	9,356	59.3%	17.4%
Control	5,917	58.0%	18.2%

*Excluding undeliverables and taxpayers who filed before Wave 1 treatment.
SOURCE: IRS Compliance Data Warehouse, Individual Return Transaction File. Data extracted May 2019.

Model Results

The estimated treatment effects for filing the current year return are reported in Table 16. The regression model estimates are reported in the Appendix. As with the nonfiler models, we estimate the models with and without the undeliverable control. For the stopfiler sample, the undeliverable control was highly correlated with the stopfiler score and influenced the model results. The direction of the predictors remained the same, but the magnitude of the estimated impact of the stopfiler model score, and therefore the estimated interaction of the stopfiler model and the treatment, was sensitive to the inclusion of the undeliverable control. For that reason, we concluded that the stopfiler model score was adequately controlling for difference in the potential for a treatment being undeliverable between the treated groups and the control group, and we focused our discussion on the treatment effects from the models without the undeliverable control.

TABLE 16. Stopfiler Marginal Effects From Models Without Undeliverable Controls for Waves 1, 2, and 3 for TY 2017 and Wave 3 for TY 2018

Treatments	Wave 1 TY 2017	Wave 2 TY 2017	Wave 3 TY 2017	Wave 3 TY 2018
Reminder letter (+Wave 2 soft notice + Wave 3 RD start)	0.008	0.003	0.016*	0.004
Reminder postcard (+Wave 2 soft notice + Wave 3 RD start)	0.003	0.008‡	0.024*	0.003

NOTES: TY 2017 Wave 1 outcome is filing or filing for an extension; the TY 2017 Waves 2 and 3 outcomes are TY 2017 filing; the TY 2018 Wave 3 outcome is filing or filing for an extension. * Indicates significance at the 95-percent level; ‡ indicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

The regression analysis removes taxpayers who filed before the start of the treatment sequence or whose mail was returned as undeliverable. The detailed model estimates are reported in the Appendix. At Wave 1, the simple reminder letter was trending toward a significant increase in filing behavior while the postcard was lagging but still indicated an increase. Recall that at Wave 2, all stopfilers assigned to treatment who had not yet filed were mailed a soft letter. With the Wave 2 results, we see the estimated treatment effects of the letter treatment remaining fairly constant and just shy of significance. The postcard treatment followed by a soft notice, however, had a larger, and now marginally significant, estimated treatment effect on filing a Tax Year 2017 return.

There is a strong and consistent relationship between the impacts of the treatments and the stopfiler model score. Recall that we sampled from the top 5 percent of taxpayers identified as being at risk for stopfiling in Tax Year 2017. Interestingly, but not unexpectedly, we see different behavior resulting from treatment depending on the risk level for stopfiling (see Figures 4-6; note the significant interaction term in the stopfiler regression results in the Appendix, Tables 24-27). The estimated relationship between the risk of becoming a stopfiler (the stopfiler model score) and the treatment effects is consistent across all three treatment timeframes.

FIGURE 4: Wave 1 Marginal Effects on Filing Returns Across Stopfiler Model Score, TY 2017

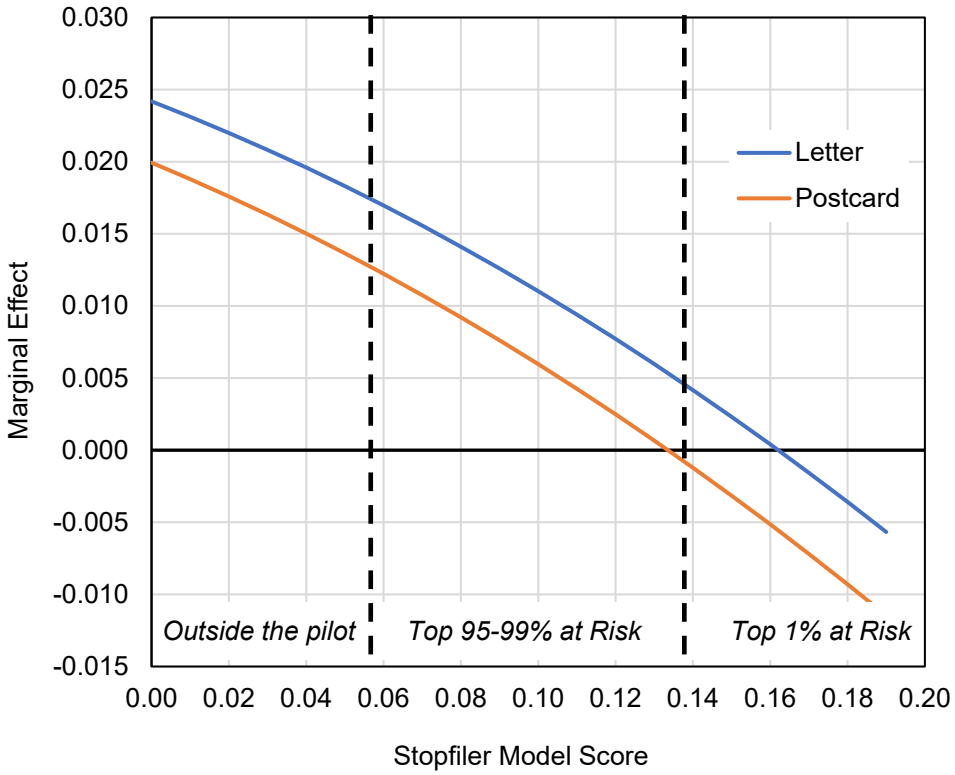


FIGURE 5: Wave 2 Marginal Effects on Filing Returns Across Stopfiler Model Score, TY 2017

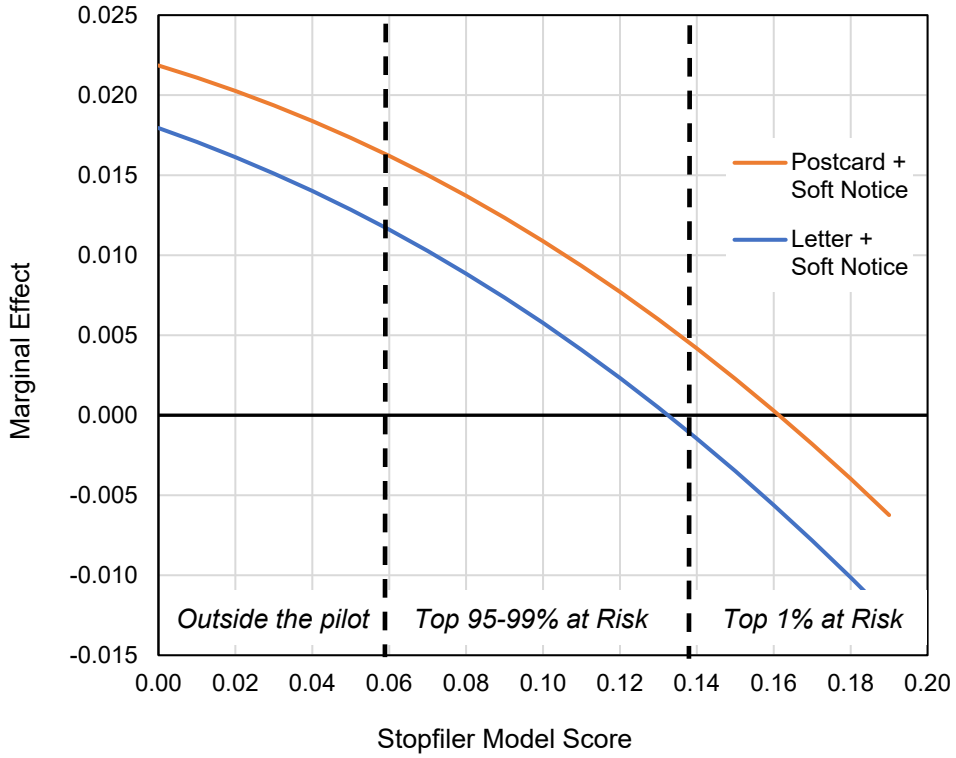
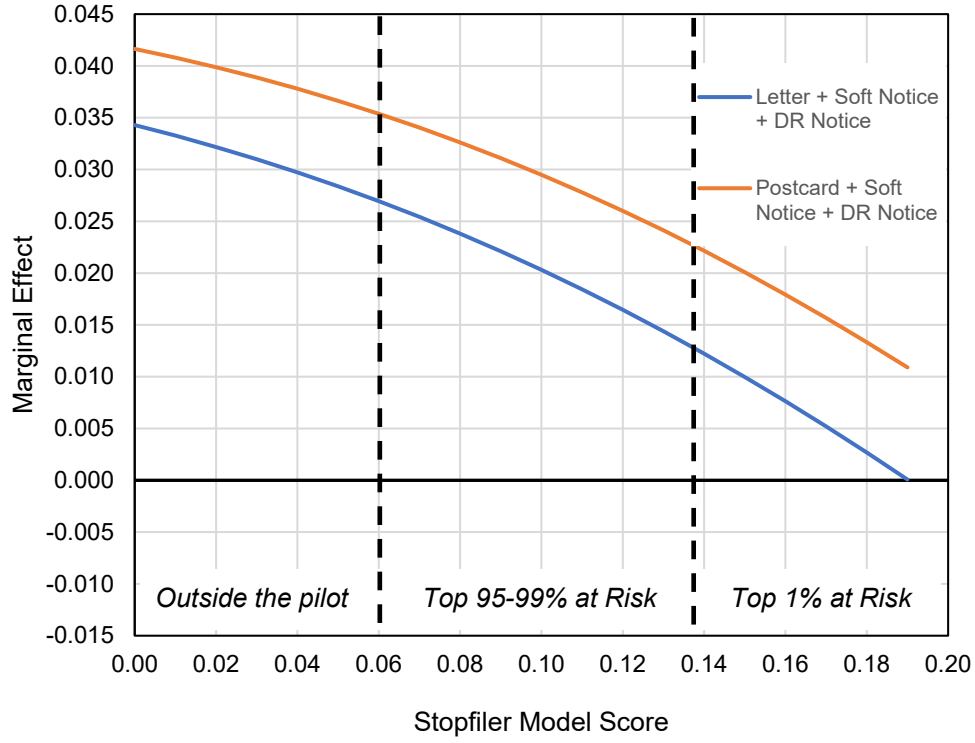


FIGURE 6: Wave 3 Marginal Effects on Filing Returns Across Stopfiler Model Score, TY 2017



The stopfiler model score is highly predictive of TY 2017 filing behavior across all waves of treatment. As we can see in Figures 4-6, those at the highest risk for stopfiling are not moved by the simple reminders at Waves 1 and 2. However, sending a straightforward prompt to file worked among those individuals predicted to be in the top 95-99 percent at risk for stopfiling—in the sense that such a reminder results in many taxpayers filing sooner than they would have otherwise, and a 0.8-percent increase in returns submitted to the IRS by the end of the calendar year.

Finally, after adding the RD notice process to the prior soft contacts at Wave 3, we find now larger and positive treatment effects across the range of risk scores for both treatment sequences. (Recall that in Wave 3, all treated stopfilers who had not yet filed their TY 2017 return were started in the TY 2017 RD process.) The addition of the Wave 3 treatment was successful in securing additional returns from the stopfilers who were not moved by earlier, softer outreach, increasing filed returns on average by 1.6 to 2.4 percentage points. Looking across the risk scores, we find larger treatment effects for taxpayers with (relatively) lower risk scores. In other words, the percentage-point increase in the filing rate resulting from the treatment declines as you become more certain that the taxpayer will become a stopfiler.

Even without a risk and treatment interaction term, a standard logistic (or other nonlinear probability models) would produce a similar relationship between the marginal treatment effect and risk (i.e., where an individual case is on the probability density function impacts the estimated marginal effect for a given case). It is worth noting that the standard linear probability model that is commonly used in the tax administration literature does not produce such a relationship, and the use of the linear probability model may need to be reconsidered, especially when focusing on populations in the tails of a risk distribution.

The regression results and marginal treatment effects for the filing of future tax returns (TY 2018) are reported in Table 27 in the Appendix. We did not detect any significant subsequent compliance effect of the treatments in the data through May 2019. An interesting note is that the estimated coefficient for the stopfiler risk model (scored on TY 2016 data and predicting TY 2017 filing behavior) was statistically significant.

Further research is needed to better understand the benefits of preemptive contact for the population of potential stopfilers. Sending reminders to those most likely to respond to treatment could increase the overall benefit for the IRS and the service it provides to taxpayers. This would yield more voluntary compliance than focusing solely on those at the highest risk for stopfiling, who may require additional treatment to return to compliance. These results also reinforce the importance of including a broader sample than you might initially suppose, as an argument could have been made to contact only people at the highest risk of becoming stopfilers. If we had done so in this pilot, we would have a less complete understanding of this treatment option.

Conclusions

The tax compliance landscape is a complex one: each taxpayer brings a unique background to bear on their decision to file or not file (or, as noted in the introduction, the absence of a decision to file). Among known nonfilers, there may be a stark tradeoff to addressing past noncompliance versus returning to compliance in the current tax year. Time and financial factors likely play a role in such a choice, as does the salience stemming from pointed contact made by the IRS. Tax administrators have choices to make in directing that focus. Does one wish to draw attention to delinquent returns or to filing timely in the current tax year? The results of this pilot study indicate that there is indeed a tradeoff to that decision. Focusing on prior-year returns has the intended effect of securing more of those delinquent returns, but taxpayers from this group were less likely to file their current-year return than their compatriots who received timely reminders regarding filing current-year returns, and vice versa. Extending the view of these treatment effects into the subsequent tax year sheds additional light on treatment options. While the start in the TY2016 RD notice process secured many returns overall, it fell short of other treatment options in securing returns from the current or subsequent tax year (TYs 2017 and 2018). Also, the largest subsequent compliance effects resulted from those contacted earliest in the process. This is consistent with the notion that the sooner taxpayers satisfy their filing obligations, the more attention they can devote to meeting next year's tax obligations (e.g., adjusting their withholdings and other prepayments, changing recordkeeping practices, etc.). This underscores the need to be thoughtful in selecting which behavior to treat, as each treatment option differs across tax years.

Perhaps an alternate or complementary route is to focus attention on compliant taxpayers who are at risk of becoming stopfilers. A gentle reminder at the right time could help draw their attention to their filing obligations. Taxpayers have many responsibilities to juggle and filing may be perceived as something that can wait, but noncompliance can quickly snowball into a burdensome and potentially overwhelming situation. Encouraging taxpayers on the edge of nonfiling to remain voluntarily compliant can help alleviate the burden placed on both the IRS and the taxpayer.

Turning to the content of lower-cost outreach, future work could further refine the reminder options available to the IRS. Among nonfiler or stopfiler populations, our results are consistent with the planning fallacy being a factor in the failure to file. Future outreach could include something to help break the task of filing into its component pieces. By seeing a task as a single unit—one thing to be completed—we underestimate all that goes into reaching the finish line and all that can go wrong en route. "I need to file my taxes," is viewing filing as a single task, potentially undervaluing the various steps that go into accomplishing it. "I need to find my W-2 and other forms, coordinate with my partner and/or dependents, decide if I'm e-filing, using a paid preparer, or paper filing, and schedule an uninterrupted block of time in which to file," is a more comprehensive list of items leading to the same outcome. Research suggests that breaking a goal into its components leads to a lower degree of planning fallacy than would otherwise exist (i.e., segmentation effect; Forsyth and Burt (2008)). Furthermore, the more complex the task, the more breaking it down in that fashion helps to decrease bias (Kruger and Evans (2004)). To the extent that the IRS wishes to further explore effective, low-cost outreach, providing taxpayers with a visual and organizational aid may help lower their overall burden and increase their likelihood to file timely.

Being mindful of both the timing and content of outreach can help the IRS's message reach the right audience at the right time. In this pilot study we saw evidence of additional information in reminder letters (about prior-year returns) crowds out the effect of the reminder to file the current year return; likewise, starting the return delinquency process focuses the taxpayer's attention on the missed return at the expense of the current-year return (as compared to simple outreach). While taxpayers in the RD notice process begin to address their current-year tax returns later in the calendar year, they do not do so at the rate of those who received a straightforward reminder directly before the filing deadline. This is an important consideration for tax administrations as administrators are frequently in a position of simultaneously trying to promote voluntary compliance and ensure that noncompliant taxpayers meet their past obligations. Knowing the boundaries of the effects of mailed communications and being intentional with the behaviors being addressed leaves tax administrations on firmer footing with securing returns, whether it be through outreach or enforcement.

References

Anderson, C. (2003). The Psychology of Doing Nothing: Forms of Decision Avoidance Result From Reason and Emotion. *Psychology Bulletin*, 129(1), 139–167.

Behavioural Insights Team. (2012). Applying Behavioural Insights To Reduce Fraud, Error, and Debt. London: Cabinet Office.

Blanken, I., N. van de Ven, and M. Zeelenberg. (2015). A Meta-Analytic Review of Moral Licensing. *Personality and Social Psychology Bulletin*, 41(4), 1–19.

Blumenthal, M., C. Christian, and J. Slemrod. (2001). Do Normative Appeals Affect Tax Compliance? Evidence From a Controlled Experiment in Minnesota. *National Tax Journal*, 54(1), 125–138.

Collins, B., A. Plumley, I. Roy, A. Turk, T. Ashley, and J. Wilson. (2018). Federal Tax Liens and Letters: Effectiveness of the Notice of Federal Tax Liens and Alternative IRS Letters on Individual Tax Debt Resolution. *2018 IRS Research Bulletin* (Publication 1500), 83–122. Washington, DC: Internal Revenue Service, Statistics of Income Division.

Forsyth, D.K., and C.D.B Burt. (2008). Allocating Time to Future Tasks: The Effect of Task Segmentation on Panning Fallacy Bias. *Memory and Cognition*, 36(4), 791–798. Retrieved from doi: 10.3758/MC.36.4.791.

Guyton, J., D.S. Manoli, B. Schafer, and M. Sebastiani. (2016). Reminders and Recidivism: Evidence From Tax Filing and EITC Participation Among Low-Income Nonfilers. NBER Working Paper No. 21904. Cambridge, MA: National Bureau of Economic Research, Public Economics Program.

Herlache, A.D., J. Millard, A.M. Miller, and M. Theel. (2018). Using Behavioral Insights in Notice Design To Improve Taxpayer Responses and Achieve Compliance Outcomes. *2018 IRS Research Bulletin* (Publication 1500), 49-82. Washington, DC: Internal Revenue Service, Statistics of Income Division.

Intille, S.S. (2002). Designing a Home of the Future. *IEEE Pervasive Computing*, 1(2), 76–82, April-June.

Kahneman, D., and A. Tversky. (1977). Intuitive Prediction: Biases and Corrective Procedures. *TIMS Studies in Management Science*, 12, 313–327.

Kettle, S., M. Hernandez, S. Ruda, and M. Sanders. (2016). Behavioral Interventions in Tax Compliance: Evidence from Guatemala. Policy Research Working Paper 7690. World Bank Group: Macroeconomics and Fiscal Management Global Practice Group.

Kruger, J., and M. Evans. (2004). If You Don't Want To Be Late, Enumerate: Unpacking Reduces the Planning Fallacy. *Journal of Experimental Social Psychology*, 40(5), 586–598. Retrieved from https://doi.org/10.1016/j.jesp.2003.11.001.

Nahum-Shani, S., S.N. Smith, A. Tewari, K. Witkiewitz, L.M. Collins, B. Spring, and S.A. Murphy. (2014). *Just-in-Time Adaptive Interventions (JITAIs): An Organizing Framework for Ongoing Health Behavior Support* (Technical Report No. 14-126). University Park, PA: The Methodology Center, The Pennsylvania State University.

Orlett, S., R. Javaid, V. Koranda, M. Muzikir, and A. Turk. (2017). Impact of Filing Reminder Outreach on Voluntary Filing Compliance for Taxpayers with a Prior Filing Delinquency, *2017 IRS Research Bulletin* (Publication 1500), 83–98. Washington, DC: Internal Revenue Service, Statistics of Income Division.

Rosage, L. (1995). Nonfiler Profiles, Fiscal Year 1993: A Focus on Repeaters. *SOI Bulletin,* Summer, 1–11. Washington, DC: Internal Revenue Service, Statistics of Income Division.

Soler, R.E., K.D. Leeks, L.R. Buchanan, R.C. Brownson, G.W. Heath, and D.H. Hopkins. (2010). Point-of-Decision Prompts To Increase Stair Use: A Systematic Review Update. *American Journal of Preventive Medicine*, 38(2), 292–300.

Appendix

TABLE 17: Nonfiler Regression Results With and Without Undeliverable Control for Wave 1, TY 2016

Parameters Dependent Variable: Filed return for TY 2016 after treatment (Note: outcomes observed for only 16 weeks after treatment)	Modeling With Undeliverable Control		Modeling Without Undeliverable Control	
	Parameter Estimates	Marginal Effect of Treatment	Parameter Estimates	Marginal Effect of Treatment
Intercept	0.6633*		-2.4233*	
	(0.1379)		(0.0776)	
Simple Letter	0.0663	0.006	0.0827	0.008
	(0.0569)		(0.0563)	
Complex Letter	0.0915	0.008	0.0974‡	0.009
	(0.0568)		(0.0562)	
Simple Postcard	-0.0975‡	-0.009	-0.0855	-0.007
	(0.0588)		(0.0583)	
Complex Postcard	0.0147	0.001	0.0109	0.001
	(0.0576)		(0.0570)	
TY 2016 Delinquent Return Notice Process	0.3505*	0.032	0.3410*	0.031
	(0.0623)		(0.0575)	
Secured Return Model Score	-1.4939*		3.0538*	
	(0.2149)		(0.1348)	
Balance Due Model Score	-1.8297*		-2.2448*	
	(0.2991)		(0.3033)	
Balance Due Model Score Squared	1.1580*		1.9700*	
	(0.3069)		(0.3127)	
Indicator Taxpayer Filed TY 2017 Return Prior to Outreach	0.6277*		0.7490*	
	(0.0630)		(0.0619)	
Probability of Undeliverable	-12.656*			
	(0.4839)			
Number of Observations	38,572		38,572	

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ‡indicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 18: Nonfiler Regression Results With and Without Undeliverable Control for Wave 1, TY 2017

Parameters Dependent Variable: Filed return or for an extension for TY 2017 after treatment (Note: outcomes observed for only 16 weeks after treatment)	Modeling With Undeliverable Control		Modeling Without Undeliverable Control	
	Parameter Estimates	Marginal Effect of Treatment	Parameter Estimates	Marginal Effect of Treatment
Intercept	0.2289*		-2.1567*	
	(0.0994)		(0.0584)	
Simple Letter	0.2784*	0.049	0.2805*	0.050
	(0.0402)		(0.0396)	
Complex Letter	0.1924*	0.034	0.1915*	0.034
	(0.0404)		(0.0399)	
Simple Postcard	0.0806*	0.014	0.0822*	0.015
	(0.0407)		(0.0401)	
Complex Postcard	0.0423	0.007	0.0340	0.006
	(0.0406)		(0.0401)	
TY 2016 Delinquent Return Notice Process	0.0575	0.010	0.0521	0.009
	(0.0479)		(0.0472)	
Secured Return Model Score	4.0373*		7.5670*	
	(0.1581)		(0.1121)	
Balance Due Model Score	-1.6554*		-1.9969*	
	(0.2234)		(0.2218)	
Balance Due Model Score Squared	1.4049*		2.0489*	
	(0.2253)		(0.2238)	
Indicator Taxpayer Filed TY 2016 Return Prior to Outreach	0.9473*		1.2452*	
	(0.0516)		(0.0524)	
Probability of Undeliverable	-9.5837*			
	(0.3299)			
Number of Observations	39,996		39,996	

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ǂindicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 19: Nonfiler Regression Results With and Without Undeliverable Control for Wave 2, TY 2016

Parameters Dependent Variable: Filed return for TY 2016 after treatment	Modeling With Undeliverable Control		Modeling Without Undeliverable Control	
	Parameter Estimates	Marginal Effect of Treatment	Parameter Estimates	Marginal Effect of Treatment
Intercept	1.0031*		-2.1621*	
	(0.1112)		(0.0636)	
Simple Letter	0.0672	0.009	0.0933	0.012
	(0.0765)		(0.0752)	
Complex Letter	0.0496	0.007	0.0658	0.009
	(0.0765)		(0.0754)	
Simple Postcard	-0.1024	-0.014	-0.0850	-0.011
	(0.0788)		(0.0777)	
Complex Postcard	-0.0384	-0.005	-0.0328	-0.004
	(0.0787)		(0.0775)	
Delinquent Return Notice Process	0.4370*	0.058	0.4171*	0.056
	(0.0519)		(0.0509)	
Secured Return Model Score	-1.0786*		3.5391*	
	(0.1701)		(0.1091)	
Balance Due Model Score	-0.7777*		-1.4184*	
	(0.2427)		(0.2487)	
Balance Due Model Score Squared	0.2824		1.3061*	
	(0.2471)		(0.2487)	
Indicator Taxpayer Filed TY 2017 Return Prior to Outreach	0.1728*		0.3135*	
	(0.0547)		(0.0562)	
Wave 2 Soft Notice only	0.0282	0.004	0.0332	0.004
	(0.0547)		(0.0539)	
Add - Soft Letter after Reminder Letter	-0.0012	-0.0002	-0.0162	-0.002
	(0.0721)		(0.0710)	
Add - Soft Letter after Reminder Postcard	0.0422	0.006	0.0312	0.004
	(0.0745)		(0.0735)	
Probability of Undeliverable	-13.239*			
	(0.3922)			
Number of Observations	42,376		42,376	

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ǂindicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 20: Nonfiler Regression Results With and Without Undeliverable Control for Wave 2, TY 2017

Parameters Dependent Variable: Filed return for TY 2017 after treatment	Modeling With Undeliverable Control		Modeling Without Undeliverable Control	
	Parameter Estimates	Marginal Effect of Treatment	Parameter Estimates	Marginal Effect of Treatment
Intercept	1.6083*		1.3338*	
	(0.0952)		(0.0546)	
Simple Letter	0.1908*	0.032	0.2016*	0.035
	(0.0656)		(0.0644)	
Complex Letter	0.1061‡	0.018	0.1119‡	0.019
	(0.0657)		(0.0645)	
Simple Postcard	0.0353	0.006	0.0514	0.009
	(0.0676)		(0.0664)	
Complex Postcard	0.0112	0.002	0.0162	0.003
	(0.0678)		(0.0664)	
Delinquent Return Notice Process	0.1174*	0.020	0.1031*	0.018
	(0.0482)		(0.0473)	
Secured Return Model Score	-1.8306*		2.2656*	
	(0.1404)		(0.0912)	
Balance Due Model Score	-0.6286*		-1.0349*	
	(0.2125)		(0.2119)	
Balance Due Model Score Squared	-0.4752*		0.2967	
	(0.2168)		(0.2170)	
Indicator Taxpayer Filed TY 2016 Return Prior to Outreach	1.6469*		1.9468*	
	(0.0448)		(0.044)	
Wave 2 Soft Notice Only	0.1377*	0.023	0.1367*	0.024
	(0.0470)		(0.0461)	
Add - Soft Letter After Reminder Letter	0.0269	0.005	0.0183‡	0.003
	(0.0619)		(0.0607)	
Add - Soft Letter After Reminder Postcard	0.0518	0.009	0.0338	0.006
	(0.0640)		(0.0628)	
Probability of Undeliverable	-12.211			
	(0.3288)			
Number of Observations	44,045		44,045	

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ‡indicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 21: Nonfiler Regression Results With and Without Undeliverable Control for Wave 3, TY 2016

Parameters Dependent Variable: Filed return for TY 2016 after treatment	Modeling With Undeliverable Control			Modeling Without Undeliverable Control		
	Parameter Estimates		Marginal Effect of Treatment	Parameter Estimates		Marginal Effect of Treatment
Intercept	0.9879*	(0.0893)		-1.8851*	(0.0547)	
Simple Letter	0.0958	(0.0703)	0.015	0.1196‡	(0.0689)	0.019
Complex Letter	0.0659	(0.0703)	0.010	0.0793	(0.0691)	0.013
Simple Postcard	-0.0369	(0.0719)	-0.006	-0.0219	(0.0705)	-0.004
Complex Postcard	-0.0199	(0.0787)	-0.003	-0.0142	(0.0706)	-0.002
Return Delinquency Notice Process	0.4381*	(0.0484)	0.069	0.4138*	(0.0474)	0.067
Secured Return Model Score	-0.7631*	(0.1399)		3.5816*	(0.0935)	
Balance Due Model Score	-0.6586*	(0.2067)		-1.2579*	(0.2065)	
Balance Due Model Score Squared	0.1891	(0.2103)		1.0715*	(0.2109)	
Indicator Taxpayer Filed TY 2017 Return Prior to Outreach	-0.0761	(0.0511)		0.0935‡	(0.0500)	
Wave 2 Soft Notice Only	0.0885‡	(0.0499)	0.014	0.0909‡	(0.0490)	0.015
Add - Wv2 Soft Letter After Reminder Letter	0.0867	(0.0857)	0.014	0.0708	(0.0840)	0.011
Add - Wv2 Soft Letter After Reminder Postcard	0.0267	(0.0879)	0.004	0.0072	(0.0864)	0.001
Add - Wv3 Soft Letter After Reminder Letter & Wv2 Soft Letter	-0.0782	(0.0744)	-0.012	-0.0752	(0.0729)	-0.012
Add - Wv3 Soft Letter After Reminder Postcard & Wv2 Soft Letter	0.0374	(0.0758)	0.006	0.0409	(0.0745)	0.007
Add - Wv3 RD Start After Reminder Letter & Wv2 Soft Letter	-0.0263	(0.0741)	-0.004	-0.0264	(0.0726)	-0.004
Add - Wv3 RD Start After Reminder Postcard & Wv2 Soft Letter	0.0407	(0.0758)	0.006	0.0540	(0.0745)	0.009
Wave 3 Soft Notice Only	0.0041	(0.0511)	0.0006	0.0208	(0.0502)	0.003
Wave 3 Delinquent Return Notice Process Only	0.0996‡	(0.0511)	0.016	0.0838‡	(0.0502)	0.013
Probability of Undeliverable	-11.9600	(0.3016)				
Number of Observations	49,974			49,974		

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ‡indicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 22: Nonfiler Regression Results With and Without Undeliverable Control for Wave 3, TY 2017

Parameters Dependent Variable: Filed return for TY 2017 after treatment	Modeling With Undeliverable Control			Modeling Without Undeliverable Control		
	Parameter Estimates		Marginal Effect of Treatment	Parameter Estimates		Marginal Effect of Treatment
Intercept	1.7651*	(0.0806)		-1.1542*	(0.0492)	
Simple Letter	0.1749*	(0.0636)	0.032	0.1872*	(0.0621)	0.036
Complex Letter	0.1125‡	(0.0636)	0.021	0.1165‡	(0.0621)	0.022
Simple Postcard	0.0719	(0.0648)	0.013	0.0844	(0.0633)	0.016
Complex Postcard	0.0220	(0.0650)	0.004	0.0266	(0.0633)	0.005
Delinquent Return Notice Process	0.1733*	(0.0462)	0.032	0.1534*	(0.0451)	0.029
Secured Return Model Score	-1.6947	(0.1227)		2.5690*	(0.0823)	
Balance Due Model Score	-0.6694	(0.1895)		-1.1132*	(0.1875)	
Balance Due Model Score Squared	-0.3560	(0.1926)		0.3937*	(0.1913)	
Indicator Taxpayer Filed TY 2016 Return Prior to Outreach	1.9440	(0.0475)		2.2753*	(0.0468)	
Wave 2 Soft Notice Only	0.1943*	(0.0452)	0.036	0.1890*	(0.0441)	0.036
Add - Wv2 Soft Letter After Reminder Letter	0.0989	(0.0777)	0.018	0.0825	(0.0758)	0.016
Add - Wv2 Soft Letter After Reminder Postcard	0.0381	(0.0795)	0.007	0.00768	(0.0776)	0.001
Add - Wv3 Soft Letter After Reminder Letter & Wv2 Soft Letter	-0.0306	(0.0674)	-0.006	-0.0229	(0.0657)	-0.004
Add - Wv3 Soft Letter After Reminder Postcard & Wv2 Soft Letter	0.0894	(0.0682)	0.017	0.0945	(0.0666)	0.018
Add - Wv3 RD Start After Reminder Letter & Wv2 Soft Letter	0.1074	(0.0669)	0.020	0.1056	(0.0652)	0.020
Add - Wv3 RD Start After Reminder Postcard & Wv2 Soft Letter	0.1335	(0.0681)	0.025	0.1485*	(0.0665)	0.029
Wave 3 Soft Notice Only	0.2170*	(0.0453)	0.040	0.2213*	(0.0443)	0.042
Wave 3 Delinquent Return Notice Process Only	0.1999*	(0.0462)	0.037	0.1764*	(0.0450)	0.034
Probability of Undeliverable	-12.033	(0.2658)				
Number of Observations	51,903			51,903		

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ‡indicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 23: Nonfiler Regression Results With and Without Undeliverable Control for Wave 3, TY 2018

Parameters Dependent Variable: Filed return or extension for TY 2018 after treatment	Modeling With Undeliverable Control			Modeling Without Undeliverable Control		
	Parameter Estimates		Marginal Effect of Treatment	Parameter Estimates		Marginal Effect of Treatment
Intercept	0.6069*	(0.0730)		-1.3691*	(0.0458)	
Simple Letter	0.1843*	(0.0583)	0.037	0.1985*	(0.0575)	0.041
Complex Letter	0.0916	(0.0584)	0.019	0.1022‡	(0.0576)	0.021
Simple Postcard	0.1129‡	(0.0588)	0.023	0.1203*	(0.0580)	0.025
Complex Postcard	0.0119	(0.0590)	0.002	0.0141	(0.0583)	0.003
Delinquent Return Notice Process	0.1136*	(0.0424)	0.023	0.1060*	(0.0418)	0.022
Secured Return Model Score	2.9396*	(0.1189)		5.9861*	(0.0856)	
Balance Due Model Score	-1.2469*	(0.1755)		-1.6957*	(0.1737)	
Balance Due Model Score Squared	0.6343*	(0.1788)		1.2731*	(0.1770)	
Wave 2 Soft Notice Only	0.1598*	(0.0413)	0.032	0.1615*	(0.0407)	0.033
Add - Wv2 Soft Letter After Reminder Letter	-0.0319	(0.0716)	-0.006	-0.0374	(0.0706)	-0.008
Add - Wv2 Soft Letter After Reminder Postcard	-0.0139	(0.0719)	-0.003	-0.0256	(0.0710)	-0.005
Add - Wv3 Soft Letter After Reminder Letter & Wv2 Soft Letter	0.0025	(0.0624)	0.0005	-0.0012	(0.0615)	-0.0002
Add - Wv3 Soft Letter After Reminder Postcard & Wv2 Soft Letter	0.0955	(0.0620)	0.019	0.0961	(0.0612)	0.020
Add - Wv3 RD Start After Reminder Letter & Wv2 Soft Letter	0.0717	(0.0623)	0.015	0.0688	(0.0614)	0.014
Add - Wv3 RD Start After Reminder Postcard & Wv2 Soft Letter	0.1039‡	(0.0620)	0.021	0.1111‡	(0.0612)	0.023
Wave 3 Soft Notice Only	0.1175*	(0.0419)	0.024	0.1258*	(0.0414)	0.026
Wave 3 Delinquent Return Notice Process Only	0.1216*	(0.0425)	0.025	0.1095*	(0.0419)	
Probability of Undeliverable	-8.1553	(0.0425)				
Number of Observations	55,721			55,271		

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ‡indicates significance at the 90-percent level.
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 24: Stopfiler Regression Results With and Without Undeliverable Control for Wave 1, TY 2017

Parameters Dependent Variable: Filed return or filed for an extension for TY 2017 after treatment (Note: outcomes observed for only 16 weeks after treatment)	Modeling With Undeliverable Control		Modeling Without Undeliverable Control	
	Parameter Estimates	Marginal Effect of Treatment	Parameter Estimates	Marginal Effect of Treatment
Intercept	3.3718*		1.5587*	
	(0.1028)		(0.0821)	
Simple Letter	0.1532	0.009	0.1684	0.008
	(0.1015)		(0.0967)	
Simple Postcard	0.1286	0.005	0.1388	0.003
	(0.1015)		(0.0967)	
Treatment Score	-0.8708		-1.0391	
	(0.7686)		(0.7288)	
Stopfiler Predictive Model Score	-0.6448		-2.8687*	
	(0.6733)		(0.6359)	
Probability of Undeliverable	-22.9358*			
	(0.6685)			
Number of Observations	17,538		17,538	

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ǂindicates significance at the 90-percent level. Significance of MEs determined by jointly testing the relevant main effect and interaction term (letter and interaction term: $\chi 2$ (2)=3.07, p=0.22; postcard and interaction term: $\chi 2$ (2)=2.19, p=0.33)
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

.

TABLE 25: Stopfiler Regression Results Without Undeliverable Control for Wave 2, TY 2017

Parameters Dependent Variable: Filed return for TY 2017 after treatment	Modeling Without Undeliverable Control	
	Parameter Estimates	Marginal Effect of Treatment
Intercept	1.9445*	
	(0.0759)	
Simple Letter (+Wave 2 Soft Letter)	0.1639ǂ	0.003
	(0.0892)	
Simple Postcard (+Wave 2 Soft Letter)	0.1996*	0.008
	(0.0895)	
Treatment Score	-1.2369ǂ	
	(0.6727)	
Stopfiler Predictive Model Score	-3.8273*	
	(0.5879)	
Number of Observations	24,753	

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ǂindicates significance at the 90-percent level. Significance of MEs determined by jointly testing the relevant main effect and interaction term (letter and interaction term: $\chi 2$ (2)=3.61, p=0.16; postcard and interaction term: $\chi 2$ (2)=5.03, p=0.08).
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 26: Stopfiler Regression Results Without Undeliverable Control for Wave 3, TY 2017

Parameters Dependent Variable: Filed return for TY 2017 after treatment	Modeling Without Undeliverable Control	
	Parameter Estimates	Marginal Effect of Treatment
Intercept	1.6667*	
	(0.0828)	
Simple Letter (+Wave 2 Soft Letter & Wave 3 RD Notice Start)	0.2566*	0.016
	(0.0980)	
Simple Postcard (+Wave 2 Soft Letter & Wave 3 RD Notice Start)	0.3116*	0.024
	(0.0983)	
Treatment Score	-1.3486‡	
	(0.7324)	
Stopfiler Predictive Model Score	-3.5212*	
	(0.6372)	
Number of Observations	17,658	

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ‡indicates significance at the 90-percent level. Significance of MEs determined by jointly testing the relevant main effect and interaction term (letter and interaction term: $\chi 2$ (2)=7.65, p=0.02; postcard and interaction term: $\chi 2$ (2)=13.45, p=0.001).
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

TABLE 27: Stopfiler Regression Results Without Undeliverable Control for Wave 3, TY 2018

Parameters Dependent Variable: Filed return for TY 2018 after treatment	Modeling Without Undeliverable Control	
	Parameter Estimates	Marginal Effect of Treatment
Intercept	1.2624*	
	(0.0772)	
Simple Letter (+Wave 2 Soft Letter & Wave 3 RD Notice Start)	0.0952	0.004
	(0.0907)	
Simple Postcard (+Wave 2 Soft Letter & Wave 3 RD Notice Start)	0.0922	0.003
	(0.0908)	
Treatment Score	-0.6173	
	(0.6959)	
Stopfiler Predictive Model Score	-2.8907*	
	(0.6069)	
Number of Observations	17,658	

NOTES: Standard errors are reported in parentheses. *Indicates significance at the 95-percent level; ‡indicates significance at the 90-percent level. Significance of MEs determined by jointly testing the relevant main effect and interaction term (letter and interaction term: $\chi 2$ (2)=1.10, p=0.58; postcard and interaction term: $\chi 2$ (2)=1.03, p=0.60).
SOURCE: IRS, Compliance Data Warehouse. Individual Return Transaction File, Individual Masterfile Status and Transaction History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2019.

Simple Letter (Wave 1 Reminder)

Department of the Treasury
Internal Revenue Service
[Address line 1]
[Address line 2]
[Address line 3]

Letter	5665
Date	[xx/xx/xxxx]
Website	[www.irs.gov/filing]
Contact telephone number	[xxx-xxx-xxxx]
Page 1 of 1	

[Taxpayer name]
[Address line 1]
[Address line 2]
[Address line 3]

REMINDER

This is a reminder to file your 2017 tax return.

What you should know

Scan this code with a QR app on your smartphone to go to IRS.gov/filing

If you're required to file:

- File by **Tuesday, April 17, 2018.**
- The average tax refund in 2016 was approximately **$2,800**. You could be eligible for valuable tax benefits, but you must file to receive them.
- For more information about filing, or getting an extension to file, go online to **IRS.gov/filing**.

If you've already filed this return, please disregard this reminder.

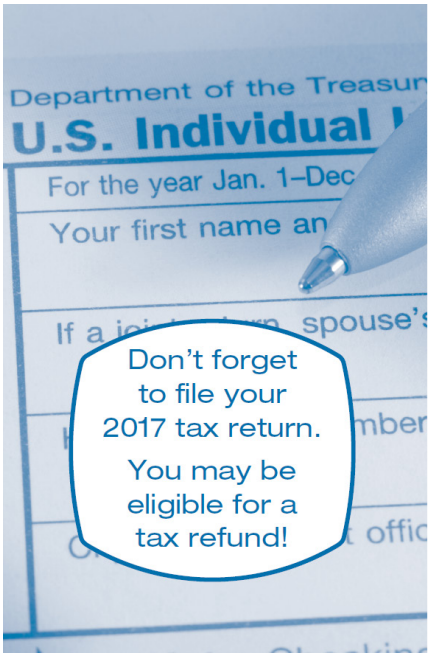
Additional information

For tax forms, instructions, and publications, visit IRS.gov/forms-pubs or call 800-TAX-FORM (800-829-3676).

Letter 5665 (Rev. 1-2018)
Catalog Number 68317C

Simple Postcard (Wave 1 Reminder)

Front



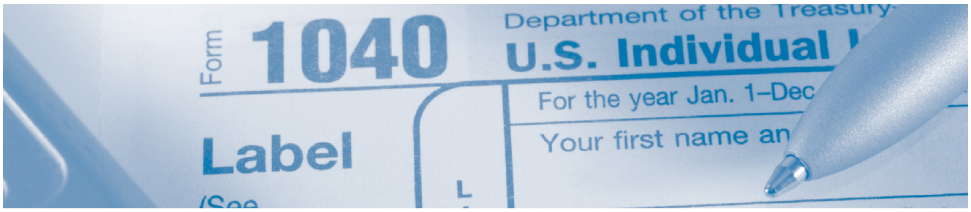
Internal Revenue Service
c/o Westat
1600 Research Blvd, RW2634
Rockville, MD 20850-3129
RETURN SERVICE REQUESTED

PRESORTED
FIRST-CLASS MAIL
POSTAGE & FEES PAID
IRS
PERMIT NO. G-48

Taxpayer's Address Here

Publication 5235 (12-2017) Catalog Number 68313K
Department of the Treasury **Internal Revenue Service** www.irs.gov

Back



If you have not already done so, remember to file your 2017 tax return by April 17th, 2018.

- Did you know the average tax refund in 2016 was approximately $2,800?
- You could be eligible for valuable tax benefits, but you must file to receive them.
- For more information about filing, or getting an extension to file, go online to www.irs.gov/filing.

Scan this code with a QR app on your smartphone to go to irs.gov/filing

Complex Letter (Wave 1 Reminder)

Department of the Treasury
Internal Revenue Service
[Address line 1]
[Address line 2]
[Address line 3]

Letter	5665-A
Date	[xx/xx/xxxx]
Website	[www.irs.gov/filing]
Contact telephone number	[xxx-xxx-xxxx]
Page 1 of 1	

[Taxpayer name]
[Address line 1]
[Address line 2]
[Address line 3]

REMINDER

This is a reminder to file your 2017 tax return

What you should know

Scan this code with a QR app on your smartphone to go to IRS.gov/filing

If you're required to file:

- File by **Tuesday, April 17, 2018.**
- The average tax refund in 2016 was approximately **$2,800**. You could be eligible for valuable tax benefits, but you must file to receive them.
- For more information about filing, or getting an extension to file, go online to **IRS.gov/filing**.

If you've already filed this return, please disregard this reminder.

Additional information

Scan this code with a QR app on your smartphone to go to IRS.gov/form4506t

It's not too late to file returns for previous years.

- You can file late tax returns at any time. You can generally claim refunds for up to 3 calendar years after the April filing deadline. For example, you can claim a tax refund for 2014, if you file that tax return by April 15, 2018. There are benefits to filing even if you don't claim a refund or if the three-year period has passed.
- To request information reported on prior year W-2s and other tax documents, submit Form 4506-T (from IRS.gov/form4506t).
- For tax forms, instructions, and publications, visit IRS.gov/forms-pubs or call 800-TAX-FORM (800-829-3676).

Letter 5665-A (1-2018)
Catalog Number 71058Y

Complex Postcard (Wave 1 Reminder)

Front

Internal Revenue Service
c/o Westat
1600 Research Blvd, RW2634
Rockville, MD 20850-3129
RETURN SERVICE REQUESTED

PRESORTED
FIRST-CLASS MAIL
POSTAGE & FEES PAID
IRS
PERMIT NO. G-48

Taxpayer's Address Here

Publication 5235-A (Rev. 2-2018) Catalog Number 68538H
Department of the Treasury **Internal Revenue Service** www.irs.gov

Back

If you have not already done so, remember to file your 2017 tax return by April 17th, 2018.

- Did you know the average tax refund in 2016 was approximately **$2,800?**
- You could be eligible for valuable tax benefits, but you must file to receive them.
- For more information about filing, or getting an extension to file, go online to www.irs.gov/filing.

Scan this code with a QR app on your smartphone to go to irs.gov/filing

It's not too late to file returns for prior tax years.

- You can file late tax returns at any time. You can generally claim refunds for up to 3 calendar years after the April filing deadline. For example, you can claim a tax refund for 2014, if you file that tax return by April 15, 2018. There are benefits to filing even if you don't claim a refund or if the three-year period has passed.
- To request information reported on prior year W-2s and other tax documents, submit Form 4506-T (from IRS.gov/form4506t).

Scan this code with a QR app on your smartphone to go to irs.gov/form4506t

Soft Letter (Waves 2 and 3)

Department of the Treasury
Internal Revenue Service
[Address line 1]
[Address line 2]
[Address line 3]

Letter	5938
Date	[xx/xx/xxxx]
Contact telephone number	800-829-1040
Page 1 of 2	

[Taxpayer name]
[Address line 1]
[Address line 2]
[Address line 3]

This is a reminder to file your [xxxx] Form 1040 tax return.

What you need to know

- The due date for filing your return has passed and we don't have record of receiving your [xxxx] federal income tax return. If you're required to file this tax return, please do so immediately.
- There are many options for electronically filing your return. Please visit www.irs.gov/efile for information on electronic filing.
- For information about mailing your return, visit www.irs.gov/filing.
- If you need wage or income information, visit www.irs.gov/transcript.
- If you've already filed this return, please disregard this reminder.

If you don't file your return

- If you expect a refund, you must file the return to claim the refund. If you wait too long to file, you may not receive what you're entitled to because the Internal Revenue Code sets strict time limits for claiming tax refunds.
- If you owe, the sooner you file and pay, the less interest or penalties you may owe. If you do owe, please see "Payment options" below.

Payment options

Pay now electronically
We offer free payment options to securely pay your tax bill directly from your checking or savings account. When you pay online or with your mobile device, you can:

- Receive instant confirmation of your payment
- Schedule payments in advance
- Modify or cancel a payment before the due date

You can also pay by debit or credit card for a small fee. To see all of our payment options, visit www.irs.gov/payments.

Letter 5938 (3-2018)
Catalog Number 69586B

Soft Letter (Waves 2 and 3)—Page 2

Letter	5938
Date	[xx/xx/xxxx]

Page 2 of 2

Payment options - **continued**

Payment plans
If you can't pay the full amount you owe, pay as much as you can now and make arrangements to pay your remaining balance. Visit www.irs.gov/paymentplan for more information on installment agreements and online payment agreements. You can also call us at 800-829-1040 to discuss your options.

Offer in Compromise
An offer in compromise allows you to settle your tax debt for less than the full amount you owe. If we accept your offer, you can pay with either a lump sum cash payment plan or periodic payment plan. To see if you qualify, use the Offer in Compromise Pre-Qualifier tool on our website. For more information, visit www.irs.gov/offers.

Additional information

- If you need additional information about filing, visit IRS.gov.
- If you need to request transcripts of your wage and income statements, visit IRS.gov/transcript. You can also request transcripts using Form 4506-T, Request for Transcript of Tax Return.
- For tax forms, instructions, and publications, visit IRS.gov/forms-pubs or call 800-TAX-FORM (800-829-3676).
- You may be able to file the return electronically within 2 years from the original due date of the return. However, you'll need an authorized e-file provider to submit your return on your behalf.
- Assistance can be obtained from individuals and organizations that are independent from the IRS. The Directory of Federal Tax Return Preparers with credentials recognized by the IRS can be found at http://irs.treasury.gov/rpo/rpo.jsf. IRS Publication 4134 provides a listing of Low Income Taxpayer Clinics (LITCs) and is available at www.irs.gov. Also, see the LITC page at www.taxpayeradvocate.irs.gov/litcmap. Assistance may also be available from a referral system operated by a state bar association, a state or local society of accountants or enrolled agents or another nonprofit tax professional organization. The decision to obtain assistance from any of these individuals and organizations will not result in the IRS giving preferential treatment in the handling of the issue, dispute or problem. You don't need to seek assistance to contact us. We will be pleased to deal with you directly and help you resolve your situation.
- Keep this letter for your records.

Example CP 59 (First Notice in the Return Delinquency Notice Process; Waves 1 and 3)

Department of the Treasury
Internal Revenue Service
Fresno, CA 93888-0025

Notice	CP59
Tax year	December 31, 2017
Notice date	January 28, 2019
Social Security number	nnn-nn-nnnn
To contact us	Phone nnn-nnn-nnnn
Your caller ID	Nnnn
Select code	nn

Page 1 of 5

TAXPAYER NAME
ADDRESS
CITY, STATE ZIP

Message about your 2017 Form 1040

You didn't file a Form 1040 tax return

Our records show that you haven't filed your tax return for the tax year ending on December 31, 2017.

What you need to do immediately
If you're required to file a tax return for 2017, please do so immediately.

- Using your current address, complete and sign your return, include a payment for any taxes due, and mail to us using the envelope provided.
- File electronically through an e-file provider if it's within 2 years from the original due date of the return.
- Pay online now at www.irs.gov/payments or mail a payment with your return.

If you don't think you had to file a tax return for 2017
Complete the enclosed Form 15103, Form 1040 Return Delinquency, to indicate whether any of the circumstances apply to you. Send us the form with the stub below in the enclosed envelope or fax it to xxx-xxx-xxxx
Indicate whether:

- You already filed a tax return for 2017 (if so, send us a signed and dated copy of the return along with your Form 15103)
- You don't think you are required to file for one of the reasons listed on the Form 15103

Example CP 59 (First Notice in the Return Delinquency Notice Process; Waves 1 and 3)—Continued

Notice	CP59
Notice date	January 28, 2019
Taxpayer ID number	Nnn-nn-nnnn
Page 2 of 5	

If we don't hear from you

- If you don't file a tax return, or dispute this notice if you feel you've received it in error, you may owe penalty and interest charges on the amount of tax due.
- You may continue to accrue penalties and interest charges on the amount of tax due.
- You risk losing your refund if you don't file your return. If you're due a refund for withholding or estimated taxes, you must file your return to claim it by April 15, 2021, plus any extension of time to file. The same rule applies to a right to claim refundable tax credits such as the Earned Income Credit.
- If you received interest or dividend income and you don't file your return or pay all taxes due, you could become subject to backup withholding. This means IRS will notify your payers (banks, etc.) to withhold a percentage of the payments you receive for dividends and interest. The payers will send the money to the IRS and you may claim it as a withholding credit on your federal income tax return.

Next steps

We'll contact you again if:

- We need additional information or clarification about your tax return
- We determine that you do need to file a tax return for 2017

Payment options

Pay now electronically

We offer free payment options to securely pay your tax bill directly from your checking or savings account. When you pay online or with your mobile device, you can:

• Receive instant confirmation of your payment

• Schedule payments in advance

• Reschedule or cancel a payment before the due date

You can also pay by debit or credit card for a small fee. To see all of our payment options, visit www.irs.gov/payments.

Payment plans

If you can't pay the full amount you owe, pay as much as you can now and make arrangements to pay your remaining balance. Visit www.irs.gov/paymentplan for more information on installment agreements and online payment agreements. You can also call us at xxx-xxx-xxxx to discuss your options.

Offer in Compromise

An offer in compromise allows you to settle your tax debt for less than the full amount you owe. If we accept your offer, you can pay with either a lump sum cash payment plan or periodic payment plan. To see if you qualify, use the Offer in Compromise Pre-Qualifier tool on our website. For more information, visit www.irs.gov/offers.

Example CP 59 (First Notice in the Return Delinquency Notice Process; Waves 1 and 3)—Continued

Notice	CP59
Notice date	January 28, 2019
Taxpayer ID number	Nnn-nn-nnnn
Page 3 of 5	

Payment options-**continued**

Account balance and payment history

For information on how to obtain your current account balance or payment history, go to www.irs.gov/balancedue.

If you already paid your balance in full within the past 21 days or made payment arrangements, please disregard this notice.

If you think we made a mistake, call xxx-xxx-xxxx to review your account.

Additional information

- Visit www.irs.gov/cp59.
- For tax forms, instructions and publications, visit www.irs.gov or call 800-TAX-FORM (800-829-3676).
- If you need wage and income information, you can request a transcript by visiting www.irs.gov/transcript.
- If you are outside the country and need assistance, please call +x-xxx-xxx-xxxx (not a toll free number), or visit www.irs.gov.
- If you had a mortgage interest debt reduced or discharged due to restructuring or foreclosure, you may qualify for tax relief under the Mortgage Forgiveness Debt Relief Act. For additional information, download Publication 4681, Canceled Debts, Foreclosures, Repossessions, and Abandonments.
- Keep this notice for your records.

If you need assistance, please don't hesitate to contact us.

Low Income Taxpayer Clinics (LITC)

Assistance can be obtained from individuals and organizations that are independent from the IRS. The Directory of Federal Tax Return Preparers with credentials recognized by the IRS can be found at http://irs.treasury.gov/rpo/rpo.jsf. IRS Publication 4134 provides a listing of Low Income Taxpayer Clinics (LITCs) and is available at www.irs.gov. Also, see the LITC page at www.taxpayeradvocate.irs.gov/litcmap. Assistance may also be available from a referral system operated by a state bar association, a state or local society of accountants or enrolled agents or another nonprofit tax professional organization. The decision to obtain assistance from any of these individuals and organizations will not result in the IRS giving preferential treatment in the handling of the issue, dispute or problem. You don't need to seek assistance to contact us. We will be pleased to deal with you directly and help you resolve your situation.

Example CP 59 (First Notice in the Return Delinquency Notice Process; Waves 1 and 3)—Continued

Notice	CP59
Notice date	January 28, 2019
Taxpayer ID number	Nnn-nn-nnnn

Page 4 of 5

This Page Intentionally Left Blank

Example CP 59 (First Notice in the Return Delinquency Notice Process; Waves 1 and 3)—Continued

Notice	CP59
Notice date	January 28, 2019
Taxpayer ID number	Nnn-nn-nnnn

Page 5 of 5

IRS

TAXPAYER NAME
ADDRESS
CITY, STATE ZIP

Notice	CP59
Social Security number	Nnn-nn-1234

Please detach and return this stub with your completed Form 15103

Internal Revenue Service
Fresno, CA 93888-0025

0000 0000000 0000000000 0000000 0000

Exchange of Information and Bank Deposits in International Financial Centres

Pierce O'Reilly and Michael A. Stemmer (OECD Centre for Tax Policy and Administration) and Kevin Parra Ramirez (Banque de France) [1]

1. Introduction

In 2009, in response to widespread international concern about tax evasion, the G20 (or Group of Twenty) declared that "the era of bank secrecy is over."[2] Since then there has been a dramatic expansion in tax transparency worldwide. At present, over 150 jurisdictions have committed to implementing the standard of Exchange of Information on Request (EOIR) and 130 jurisdictions now participate in the Convention on Mutual Administrative Assistance in Tax Matters (the MAC), which provides an international legal basis for all types of exchanges with more countries joining each year. More than 100 jurisdictions have committed to automatically exchanging information related to offshore accounts and over 90 jurisdictions have already commenced exchanges. These new initiatives have marked a step change in the global commitment to tax transparency.

These changes have brought with them significant interest among stakeholders in understanding the impact of the Exchange of Information (EOI) to both assess its effectiveness and identify strategies that could improve its functioning. These stakeholders include member jurisdictions of the Global Forum on Transparency and Exchange of Information for Tax Purposes (the Global Forum), the private sector, nongovernmental organisations, and the public. This paper provides results using cross-border banking statistics to provide an assessment of the impact of EOI.

There is a growing body of literature using international financial statistics to assess offshore activity and tax evasion, as well as stocks of potentially hidden wealth. There is also literature using international investment data to assess the impact of EOI. In many instances, however, studies have suffered from a lack of available data in both investments and treaties signed, as well as from challenges in accurately assessing whether changes, discrepancies, or asymmetries in these data reflect offshore activity and changes in this activity, or whether they represent measurement error or other factors. Using unique data, this paper provides a comprehensive assessment of the impact of EOI on cross-border bank deposits.

1.1 The literature on exchange of information and international financial centres

The literature on the impact of EOI is small but growing.[3] Using data on cross-border financial liabilities in International Financial Centres (IFCs) has been a key means of assessing the impact of EOI. In an earlier paper, Johannesen and Zucman (2014) showed that bank liabilities in IFCs had not declined significantly since the expansion of EOI in 2008, following the G20 declaration that the era of bank secrecy was over. While they did find evidence that some low-tax jurisdictions experienced a fall in bank deposits in the aftermath of the

[1] The authors would like to thank Pascal Saint-Amans for initiating and supporting the collaboration with the Banque de France. Moreover, the provision of data and valuable comments by staff at the Bank for International Settlements (BIS) during the writing process are gratefully acknowledged.

This paper has greatly benefited from support, comments, and suggestions provided by David Bradbury and Bert Brys at the OECD and by Delegates of Working Party No. 2 on Tax Policy Analysis and Tax Statistics of the OECD Committee on Fiscal Affairs. The authors wish to thank participants for comments at the 9th Annual IRS/TPC Joint Research Conference and the OECD CTPA Tax Policy Seminar.

The authors would also like to thank Marc-Alain Bahuchet, Sebastian Beer, Monica Bhatia, Maria Borga, Anzhela Cédelle, Ruud de Mooij, Donal Godfrey, Sarita Gomez, Peter Green, Niels Johannesen, Åsa Johansson, Philip Kerfs, Laurence Lelogeais, Giorgia Maffini, Zayda Manatta, Bethany Millar-Powell, Valentine Millot, Tom Neubig, Lyndsay Smyth, Stéphane Sorbe, Alastair Thomas, Hervé Thoumiand and Gabriel Zucman for highly valuable comments and input.

Thanks also to Violet Sochay for administrative support and Hazel Healy and Carrie Tyler for assistance with the publication.

[2] G20 Leaders Statement, London, UK. http://www.oecd.org/newsroom/44431965.pdf.

[3] This literature is more extensively summarised in Menkhoff and Miethe (2019) and as well as in OECD (2018).

signature of new EOIR agreements, the authors argued that the lack of a broad decline in deposits in IFCs suggested that taxpayers responded to EOIR by transferring deposits to other nonexchanging IFCs:

> *[… so far,] treaties have led to a relocation of bank deposits between tax havens but have not triggered significant repatriations of funds … A comprehensive network of treaties providing for automatic exchange of information would put an end to bank secrecy and could make tax evasion impossible (Johannesen and Zucman (2014)).*

At the time that Johannesen and Zucman's paper was published, the network of EOIR relationships was far from comprehensive (see Section 2 below). Since then, the tax transparency environment has continued to evolve, and several papers have used more up-to-date data to assess the impact of continuing developments. Each of these studies has found that EOIR and the Automatic Exchange of Information (AEOI) are, to varying degrees, associated with reductions in bank deposits in IFC jurisdictions. *Figure 1* compares the studies cited above and shows the different estimates of the extent to which EOIR and AEOI have led to a decline in IFC deposits. *Table 1* compares the above studies in terms of their varying sample sizes, time periods covered, and different jurisdictions defined as IFCs.

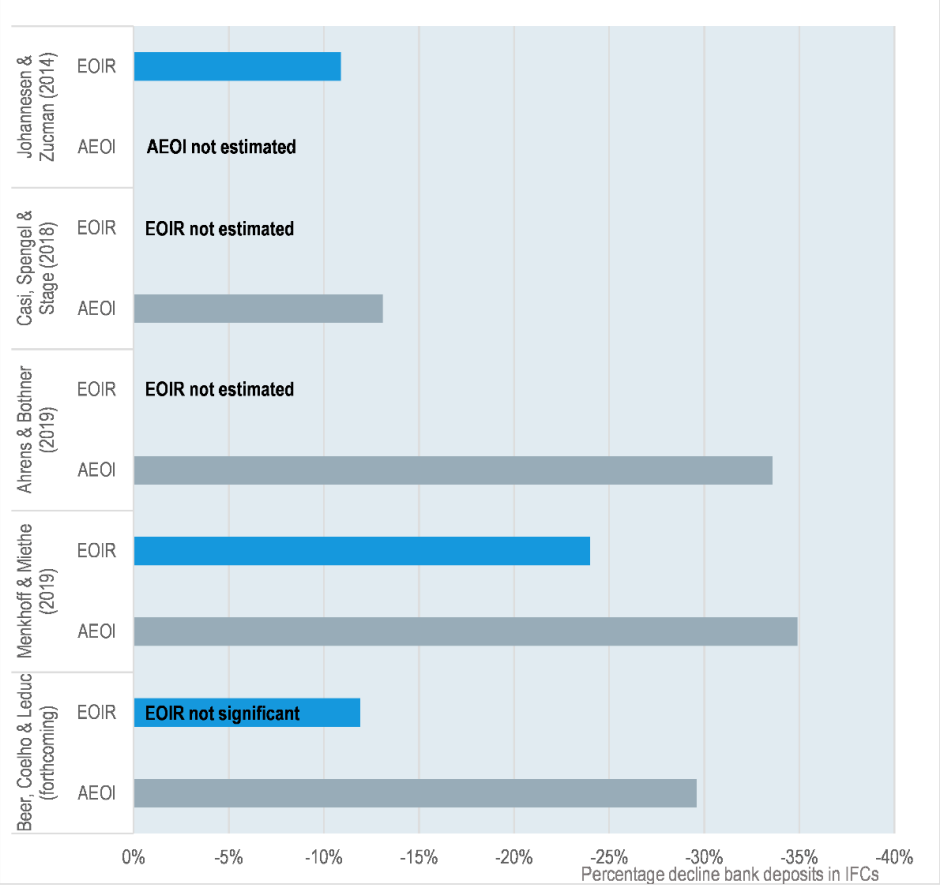
Menkhoff and Miethe (2019) repeat and extend the analysis of Johannesen and Zucman (2014) by analysing both inflow and outflow deposits held in the non-IFC and IFC jurisdictions. They find that EOIR is associated with a significant but declining impact on the bank deposits held in IFC jurisdictions from non-IFC jurisdictions after the signature of an EOIR agreement. They also find mirroring but lagged reactions to deposits in non-IFC jurisdictions from IFCs. Finally, they find a significant impact on IFC deposits from activation of AEOI agreements under the Common Reporting Standard (CRS). Casi, Spengel and Stage (2018) carry out a difference-in-differences analysis, with a sole focus on AEOI and a sample limited to the years from 2014 to 2017. They argue that this reduced sample allows them to better focus on the impact of AEOI and find that AEOI is associated with a statistically significant reduction in bank deposits in IFCs. Beer, Coelho and Leduc (2019) extend this analysis by assessing EOIR, AEOI, and Foreign Account Tax Compliance Act (FATCA) with a longer period covered and an increased IFC sample similar to Johannesen and Zucman (2014). Finally, Ahrends and Bothner (2019) employ a difference-in-differences model to estimate successfully the impact of AEOI on non-IFC deposits.

In addition to the papers focusing on the impact of EOI on bank deposits, several other papers in the literature analyse the effects of EOI on other forms of financial assets. Hanlon, Maydew and Thornock (2015) and De Simone, Lester and Markle (2019) focus on the response of portfolio holdings of IFCs in the United States in the aftermath of the implementation of FATCA and find that the implementation of FATCA agreements between the United States and IFCs is associated with reduced portfolio investment from those IFCs in the United States. Heckemeyer and Hemmerich (2018) assess the response of portfolio holdings of IFCs in securities markets in OECD countries. They find that EOIR is associated with reduced portfolio investment in securities markets in OECD countries by IFC jurisdictions participating in EOI. Kemme, Parikh and Steigner (2017) find similar results, albeit with more modest effects of the expansion of EOI on portfolio activity.

Other papers have analysed the impact of EOI in using other data. Omartian (2016) employs data from international data leaks to argue that EOI is associated with declines in company incorporations in IFCs. Johannesen *et al.* (2018) used the Internal Revenue Service's (IRS's) tax administration data and found that expanded enforcement initiatives in the United States have resulted in approximately 60,000 individuals disclosing offshore accounts with a combined value of around USD 120 billion, corresponding to around USD 0.7-1.0 billion in additional tax revenue.

Against this background, this paper makes several contributions to the literature. First, it expands on the work of Johannesen and Zucman (2014) by employing a larger sample in terms of time and country coverage than is available to other researchers. Second, unlike other papers in the literature, this paper accounts for the impact of the rapid expansion in EOI networks that has occurred through the signature of the MAC since 2010. By jointly testing for EOIR signatures, the impact of the announcement by jurisdictions of their commitment to implement AEOI and the commencement of exchanges under AEOI, it provides a comprehensive assessment of EOI impact and shows that the size of the banking sector in IFCs has been substantially limited by the expansion of the EOI network. Finally, it highlights the overall decrease in deposits in IFCs as evidence that EOI has improved tax compliance.

FIGURE 1. Estimates of the decrease in IFC deposits associated with EOI in the literature



NOTE: EOI impact expressed in percentage decline of deposits in IFCs based on baseline estimations in the respective articles. The EOI effects on deposits expressed in percentages have been recalculated where necessary based on the formula 100 * exp (estimated coefficient) - 1. This transformation accounts for log-linear specifications in the estimated models of the respective articles.
SOURCE: Authors' calculations based on the relevant literature cited.

TABLE 1. The existing literature employs different sample lengths and IFC lists

Articles	Sample length	IFC sample
Johannesen and Zucman (2014)	Q4 2003–Q2 2011	Austria; Belgium; Cayman Islands; Chile; Cyprus; Guernsey; Isle of Man; Jersey; Luxembourg; Macau, China; Malaysia; Panama; Switzerland
Casi, Spengel and Stage (2018)	Q4 2014–Q3 2017	Guernsey; Hong Kong, China; Isle of Man; Jersey; Macau, China
Ahrends and Bothner (2019)	Q1 2009–Q4 2017	Austria; Belgium; Chile; Guernsey; Hong Kong, China; Isle of Man; Jersey; Luxembourg; Macau, China; Switzerland
Menkhoff and Miethe (2019)	Q1 2003–Q4 2017	Belgium; Chile; Guernsey; Ireland; Isle of Man; Jersey; Luxembourg; Switzerland
Beer, Coelho and Leduc (2019)	Q1 1995–Q2 2018	Austria; Bahamas; Bahrain; Belgium; Bermuda; Chile; Netherlands Antilles/Curaçao[1]; Cyprus; Guernsey; Hong Kong, China; Isle of Man; Jersey; Luxembourg; Macau, China; Panama; Singapore; Switzerland

[1] In the BIS LBS, data for Netherlands Antilles are succeeded by data for Curaçao. See BIS (2017), https://www.bis.org/statistics/dsd_lbs.pdf.
SOURCE: Based on the relevant literature cited.

1.2 Paper outline

The remainder of this paper proceeds as follows:

Section 2 focuses on a specific data source of cross-border financial activity—locational banking statistics (LBS) available from the Bank for International Settlements (BIS). It provides some stylised facts about the data and notes the overall decline in deposits of banks in IFCs held by nonbank counterparties over the last 10 years. It also describes the expansion of EOI over this period.

Section 3 provides results of a regression analysis on the impact of EOI agreements between two jurisdictions on cross-border bank deposits.[4] The results suggest that when an IFC jurisdiction signs or commits to an EOI agreement with a non-IFC jurisdiction, the stock of bank deposits in that IFC with respect to counterparties in non-IFC jurisdictions decreases. Statistically significant results are found for the commencement of AEOI exchanges. The results also suggest that the impact of EOIR changes over time. Initial EOIR agreements signed in the aftermath of the commencement of a peer review in 2009 had a strong impact. However, the impact of each additional agreement has been more muted, potentially due to the increasingly multilateral nature of the EOIR network.

These results show the impact of tax transparency on bank deposits in IFCs, which suggests that secrecy is one of the features attracting wealth to these jurisdictions. Following an EOI agreement, tax authorities can obtain access to banking information. This means that the risks of engaging in tax evasion increase for holders of undisclosed bank deposits. The drop in offshore holdings in the aftermath of the EOI agreements suggests that EOI is successful in reducing bank deposits that were concealed from tax authorities in IFCs.

Section 4 concludes with a series of robustness checks that examine the results in Section 3 in more detail. It provides further evidence that the decline in bank deposits in IFCs associated with expanded EOI is linked to tax evasion by demonstrating that the decline is not present in non-IFCs. It assesses whether the results in Section 3 vary depending on the definition of IFCs and compares the impact of EOI with the impact of voluntary disclosure programmes implemented domestically in various jurisdictions.

Section 5 concludes the paper with suggestions for possible future research in this area.

2. Assessing changes in IFCs using cross-border banking statistics

Bank deposits are a key component of cross-border investment activity. The BIS publishes quarterly data on bank liabilities in the LBS, including both deposits and banks' other holdings of securities aggregated at the jurisdiction level. For example, in the case of France, it publishes total deposits held by French residents in foreign banks and total deposits held by foreign residents in French banks.

As discussed in Section 1.1, data on banking activity have been used repeatedly to study the impact of EOI (Johannesen and Zucman (2014); Huizinga and Nicodème (2004); Menkhoff and Miethe (2019)). There are several reasons for this. Access to banking information that is "foreseeably relevant" for tax purposes is specifically provided for under EOIR Agreements. Furthermore, information on bank deposits held abroad is one of the information categories covered by the AEOI Standard. This means that, to the extent that there are changes in cross-border investment activity because of EOI, bank deposits should be one of the assets most directly affected.

Moreover, banking data are among the best-quality data available on international financial activity. In recent years, the BIS has made substantial amounts of data publicly available to researchers. These data include bilateral information for reporting jurisdictions, which are data on assets held in the reporting jurisdiction by a resident of a counterparty jurisdiction. The coverage of the BIS data is further described in Box 1.

This paper, like others in the literature, focuses on bank deposits of nonbank actors and, in particular, on bank deposits in IFCs held by nonbank residents of non-IFCs (i.e., the category of loans and deposits in *Table 2*). This is discussed further in Box 2. Focusing on nonbank deposits involves excluding banks' deposits with respect to other banks and their own affiliates abroad, as banks' lending to each other on the interbank market is unlikely to be impacted substantially by EOI expansion.

[4] For the purposes of this paper, an EOI agreement includes all types of agreements enabling EOI, such as the MAC, bilateral tax treaties containing an article for exchange of information or bilateral tax information exchange agreements (TIEAs).

The nonbank category includes households, corporates, general government, noncorporate enterprises, such as trusts and other nonfinancial institutions (e.g., charities and foundations). Even though this is a narrower category than all bank liabilities, even this category is broad and presents several challenges from the perspective of accurately assessing the impact of EOI. While some entities may be used by individuals to evade taxes, others may be engaged in legitimate business purposes. An important caveat to the analysis is that various types of nonbank actors may respond to EOI differently, which influences the results presented in the analysis below.

A few additional limitations of the BIS LBS are noteworthy. The data are recorded as end-of-quarter observations and as such constitute stocks. These data thus provide a snapshot of deposits at a given point in time and cannot provide details of flows over periods compared to flow variables. Moreover, the deposit data are collected based on immediate rather than ultimate ownership.

Box 1: Coverage of the BIS data

The BIS public locational banking statistics

There are 47 reporting jurisdictions in the public BIS file. Of these, 29 jurisdictions have bilateral counterparty data in the public file, including: Australia; Austria; Belgium; Brazil; Canada; Chile; Chinese Taipei; Denmark; Finland; France; Greece; Guernsey; Hong Kong, China; Ireland; Isle of Man; Italy; Jersey; Korea; Luxembourg; Macau, China; Mexico; Netherlands; Philippines; South Africa; Spain; Sweden; Switzerland; the United Kingdom, and the United States.

Restricted BIS data provided to the Banque de France

Of the 29 jurisdictions reporting in the public file, seven provide time series extensions in the restricted sample of the BIS. Fourteen jurisdictions have further provided restricted but close to full bilateral data to the BIS for various periods. However, the data supplied pertain to varying dates. The confidential bilateral data reported to the BIS are not accessible; hence, they are not used in this paper.

Table 2. Variables available in locational banking statistics

Measure	Balance sheet position	Type of instrument	Currency denomination	Currency type of reporting country	Type of reporting institution	Counterparty sector	Position type
Stocks	**Total liabilities**	All instruments	**All currencies**	All currencies	All reporting banks	All sectors	Cross-border
Break adjusted changes	Total claims	Debt securities[1]	Swiss Franc	Foreign currency	Foreign branches	Banks, total[2]	All
		Debt securities, short-term	Euro	Domestic currency	Foreign subsidiaries	Banks, related offices	Local
		Loans and deposits[3]	Pound Sterling	Unclassified currency	Domestic banks	Banks, central banks	Unallocated
		Other instruments	Japanese Yen			Nonbanks, total[4]	
		Unallocated by instrument	U.S. Dollar			Nonbank financial institutions	
			All other currencies			Nonfinancial sectors	
			Unallocated currencies			Unallocated by sector	

1 Banks' holdings of debt securities are defined as comprising assets in all negotiable short and long-term debt instruments (see Box 2).
2 Generally defined as institutions whose business it is to receive deposits and/or close substitutes for deposits, and to grant credits or invest in securities on their own account. Within the scope of the BIS locational banking statistics only, official monetary authorities including the BIS and the ECB are also regarded as banks. Can refer to banks' head offices or affiliates. Money market funds, investment funds and pension funds are excluded from this category (BIS (2013)).
3 Deposits comprise all claims reflecting evidence of deposit, including non-negotiable certificates of deposit (CDs), which are not represented by negotiable securities (see Box 2).
4 All entities (including individuals but excluding official monetary authorities) other than those defined as "banks." General government and public corporations are part of the nonbank sector (BIS (2013)).
NOTE: Data series highlighted in bold are the focus of the analysis in this paper.
SOURCE: Authors' calculations based on BIS LBS.

Box 2: Bank deposits and bank liabilities

One issue in examining the impact of EOI on financial activity in IFCs is choosing the most appropriate outcome variable for the analysis. In assessing the impact of EOI on asset holdings in or through IFCs, a goal should be to analyse those financial assets that would be impacted by EOI, i.e., those financial assets that are likely to be held by potential tax evaders. Bank deposits held by individuals are one clear example of an asset class that may be impacted by EOI. This can be the case whether they are held directly or as part of structures designed to conceal beneficial ownership.

There is early evidence of the importance of bank deposits in the academic literature on tax evasion. Using data on Swiss bank liabilities, Zucman (2013) estimates that bank deposits form approximately 25 percent of global hidden wealth. Using data from Italy's Voluntary Disclosure Programme for Hidden Assets, Pellegrini, Sanelli, and Tosti (2016) found that while bank deposits were the most commonly repatriated asset class, they comprised 13.5 percent of total disclosed wealth. A more recent study by Alstadsaeter, Johannesen and Zucman (2018) allocated a wealth equivalent of about 10 percent of global gross domestic product (GDP) to IFCs.

The definition of bank deposits versus bank liabilities

While bank deposits are often an asset class discussed in the literature and media on tax evasion, it is important to understand how they are defined in the data used in this study. In the BIS LBS Reporting Guidelines, bank deposits are defined as "all claims reflecting evidence of deposit—including nonnegotiable certificates of deposit (CDs)—which are not represented by negotiable securities." This includes "[f]unds received by banks from nonresidents in any currency or from residents in foreign currency on a trust basis. The BIS reporting guidelines also note that "[f]unds lent or deposited on a trust basis in banks' own name, but on behalf of third parties, with nonresidents in any currency or with residents in foreign currency, represent international assets which also fall into the category of loans and deposits." Bank deposits also include working capital between related banks (BIS (2013)).

The other major component of banks' overall liabilities in the BIS LBS is banks' holdings of securities. Banks' holdings of securities are defined as "comprising assets in all negotiable short and long-term debt instruments (including negotiable CDs, but excluding equity shares, investment fund units and warrants) in domestic and foreign currency issued by nonresidents and all such instruments in foreign currency issued by residents. Banks' holdings of debt securities should include those held in their own name and those held on behalf of third parties as part of trustee business" (BIS (2013)).

Bank deposits and bank liabilities in international financial centres

Data are available in the BIS LBS for both bank liabilities and bank deposits, with differing degrees of detail. This study focuses on bank deposits alone, omitting securities from consideration. There are several reasons for this. First, the reporting quality of information for securities is uneven. Some countries do not collect or report high-quality data on securities as a part of bank liabilities in the BIS data, which means that it is difficult to compare those countries that do include securities in their overall figures of bank liabilities with those countries that do not. A key reason for this is that it is challenging for banks to know the counterparty country and sector of the holders for tradable securities.

Second, bank deposits (as opposed to broader bank liabilities) may offer a better proxy of the taxpayer activity that EOI tries to address. This is because securities held in banks in IFCs may be held there not on behalf of individual households who may be hiding wealth, but on behalf of mutual funds or other asset management companies who locate in IFCs due to regulatory or other considerations. Several of the IFCs with large bank liabilities in the BIS data, such as Bermuda, Luxembourg or the Cayman Islands, are well-known centres for asset management activity.[1] Where mutual funds or hedge funds buy and sell assets, they may hold these assets on deposit with banks, who act as custodians on behalf of the funds. It is likely that these kinds of bank liabilities will be less responsive to the expansion of EOI when compared to bank deposits held by individual taxpayers or held by these taxpayers through companies. This means that a broader definition of bank liabilities inclusive of securities may function less well as a proxy for overall assets being hidden in IFCs relative to focusing on bank deposits alone.

[1] These stylised facts are discussed further in Section 2.1.

2.1 Stylised facts of deposits in BIS reporting countries

Zucman and Johannsen (2014) highlighted the lack of decline in IFC deposits relative to non-IFC deposits in the aftermath of the financial crisis as evidence of the limited impact of EOI. However, as the sample period used in their paper concluded in 2011, it did not take into account the significant further development of the network of exchange relationships after 2011, nor did it consider the widespread adoption of the AEOI Standard. Since 2011, there has been a change in the overall trend of IFC deposits as compared to non-IFC deposits. While both IFC and non-IFC deposits declined in the years after the financial crisis, non-IFC deposits have since surpassed precrisis levels, while IFC deposits have continued to decline.[5] This could suggest that the immediate postcrisis contraction in bank deposits, which affected both IFCs and non-IFCs, was a result of the crisis itself. However, the contraction in IFC deposits (especially those in European and Caribbean IFCs) in more recent years, while there has been an expansion in non-IFC deposits, points to the potential impact of EOI.

Figure 2 shows bank deposits aggregated across IFCs and non-IFCs (in USD millions). Whereas the upper panel displays foreign-owned deposits in all IFCs, the lower panel presents IFC cross-border deposits excluding the Cayman Islands; Hong Kong, China; and Macau, China, as discussed below.

The broad trends in the data are similar in both charts. Following a peak in 2008, the level of bank deposits declined in both IFCs and non-IFCs. Bank deposits in non-IFCs began a return to precrisis levels from 2010 onwards and have recently even surpassed the 2008 peak. However, they continued to decline steadily in IFCs, albeit more gradually when excluding the Cayman Islands.

Deposits including all reporting IFC jurisdictions rose substantially in the period since the early 2000s and rose even faster in the period immediately before the global financial crisis, reaching a peak in the second quarter (Q2) of 2008 (USD 2.5 trillion).[6] Since then, deposits of banks in IFCs in respect of nonbanks have fallen substantially, by USD 1.055 billion or 42 percent. Amid an overall declining trend, however, periods of stronger decreases appear. A large part of the total reduction occurred during and in the immediate aftermath of the global financial crisis, where deposits fell by 14 percent between the Q2 of 2008 and Q2 of 2010. During the subsequent 2 years, IFC deposits experienced an even steeper decline of about 12 percent (from Q2 of 2010 to Q2 of 2012) and suffered from another decrease of around 17 percent during Q2 of 2013 and Q4 of 2015. This decrease has continued in recent years by a further 18 percent since the first quarter (Q1) of 2016.

Figure 2 also presents results with the Cayman Islands; Hong Kong, China; and Macau, China omitted from the set of IFCs. This is because there is a particularly strong reduction in bank deposits in the Cayman Islands. Bank deposits in the Cayman Islands have historically been driven by a strong share of bank deposits from financial institutions in the United States.[7] It is likely that domestic regulatory changes in the United States (e.g., the Dodd-Frank Act), have led U.S. financial institutions to significantly reduce bank account activity in the Cayman Islands. Given that this reduction may be driven, at least in significant part, by factors other than changes in the tax transparency environment, separate results are presented for the rest of the sample as well. When excluding the Cayman Islands, the overall downward trend of IFC deposits is more modest. After the peak in Q1 of 2008 (USD 1.7 billion), deposits fell by USD 410 billion, an equivalent of 24 percent. However, the overall decline also disguises periods of stronger and weaker declines. During and directly after the financial crisis, IFC deposits decreased strongly by 23 percent between Q1 of 2008 and Q2 of 2010. The period between Q2 of 2011 and Q2 of 2012 was marked by another decrease of around 8 percent. Since 2013, deposits have dwindled rather slowly but steadily by around 11 percent.

While cross-border deposits have been stable over time in some IFC jurisdictions, others have experienced an increase around the time of the global financial crisis and a subsequent decrease, which has continued through to the present. Since Q1 of 2008, declines have been evident in Guernsey, the Isle of Man, and Jersey as well as in the Bahamas and the Cayman Islands. By contrast, Bahrain; Hong Kong, China; Macau,

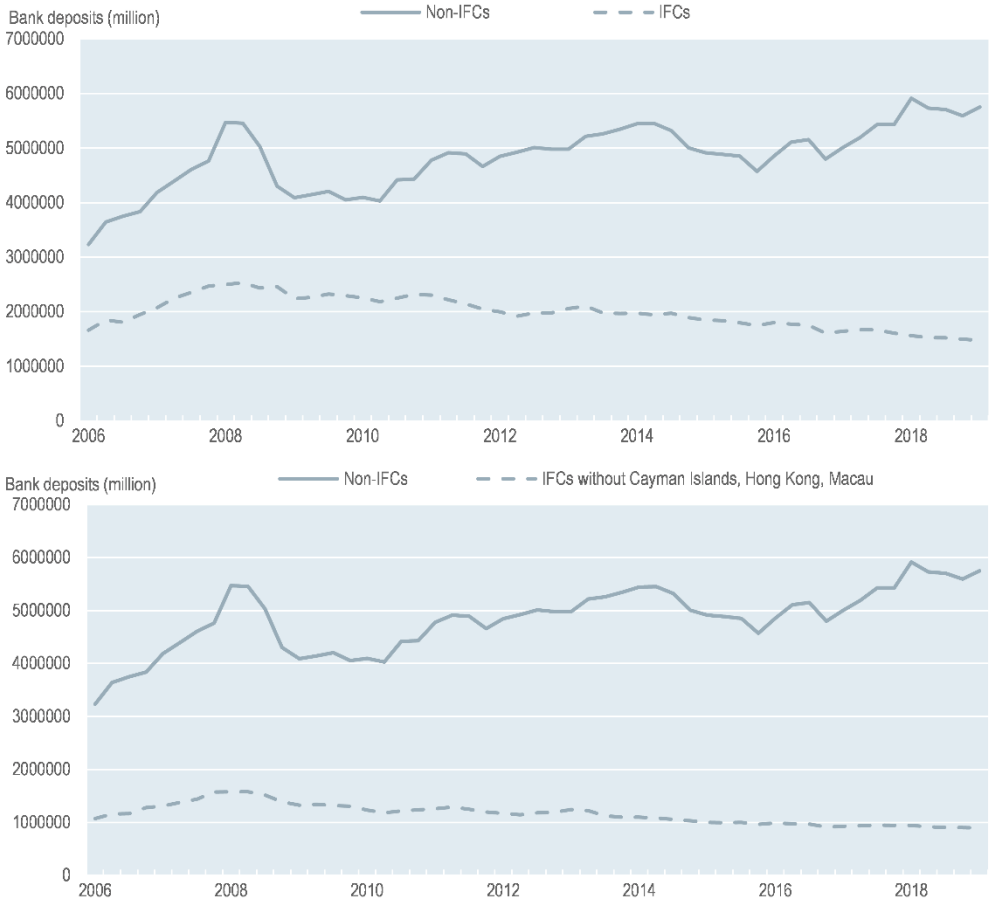
[5] The definition of IFCs is discussed further in Box 3.

[6] Reporting IFC jurisdictions are discussed in Box 3 and are highlighted in bold.

[7] "Historically, overnight sweep accounts in OFCs such as the Cayman Islands developed because Regulation Q prohibited U.S. banks from paying interest on demand deposit accounts. Regulation Q was repealed in 2011 and this may partly explain the drop in Cayman LBS from USD 1800 billion in 2011 to about USD 1400 billion" (Fichtner (2016)).

China; Panama, and Singapore have experienced an increase in cross-border deposits over time, though in the case of Macau, China, and Panama, this increase has levelled off in recent years. In Switzerland, a sharp decline in deposits (of just over USD 100 billion) can be noted between June and September of 2013 (the G20 endorsed the AEOI Standard in September 2013 and Switzerland announces the U.S.-Swiss Bank Program in August 2013).[8, 9]

FIGURE 2. Changes in cross-border bank deposits (2006–2019)



NOTE: The upper panel shows cross-border deposits in non-IFCs and IFCs, the lower panel cross-border deposits in non-IFCs and IFCs excluding the Cayman Islands; Hong Kong, China; and Macau, China. Data are provided for nonbank counterparties only. Data are aggregated across currencies, sectors, reporting institutions, and instrument type.
SOURCE: Authors' calculations based on BIS LBS.

[8] The U.S.-Swiss Bank Program was announced jointly by U.S. and Swiss authorities on August 29, 2013, to resolve potential criminal liabilities of Swiss banks in the United States. Eligible Swiss banks had to advise U.S. authorities of suspected tax-related criminal offenses linked to undeclared U.S.-related accounts. To date, 82 Swiss banks benefit from nonprosecution agreements (https://www.justice.gov/tax/swiss-bank-program).

[9] This shift may also have been driven by changes in the reporting of trustee deposits.

Box 3: Definitions of international financial centres

The definition of what constitutes an international financial centre is a controversial and challenging subject. In the academic literature, a wide variety of lists have been used, based on a wide variety of criteria. These criteria are often subjective. From the perspective of the assessment of EOI on bank deposits, the ideal focus would be on those jurisdictions that specialise in international banking. This presents an important caveat, as different IFCs may have different specialisations. For example, some IFCs may specialise in insurance activity, some as a centre for hedge fund and mutual fund activity, some in banking activity, some in trust activity, and so on. Assessing the impact of EOI requires a nuanced understanding of the differences across IFC profiles, and therefore of the varying ways the expansion of EOI will affect different IFCs.

The list of IFCs used in this study is based on a list of 46 jurisdictions defined by the IMF (2000). This IMF report defines an offshore financial centre (OFC) as follows:

"[A] centre where the bulk of financial sector activity is offshore on both sides of the balance sheet (i.e., the counterparties of the majority of financial institutions' liabilities and assets are nonresidents), where the transactions are initiated elsewhere, and where the majority of the institutions involved are controlled by nonresidents. OFCs are usually referred to as:

- Jurisdictions that have relatively large numbers of financial institutions engaged primarily in business with nonresidents;
- Financial systems with external assets and liabilities out of proportion to domestic financial intermediation designed to finance domestic economies; and
- More popularly, centres which provide some or all of the following services: low or zero taxation; moderate or light financial regulation; banking secrecy and anonymity."

Of the jurisdictions on this IMF list, many smaller centres do not report bank liability data to the BIS. Those who do report to the BIS are the Bahamas; Bahrain; Bermuda; the Cayman Islands; the Netherlands Antilles/Curaçao; Cyprus; Guernsey; Hong Kong, China; Ireland; the Isle of Man; Jersey; Luxembourg; Macau, China; Malaysia; Panama; Singapore; and Switzerland. Reporting of bilateral liability and deposit information is even more patchy and has been discussed in Section 2.1.

In this paper, the analysis relies on an amended list of IFCs based on the IMF OFC definition. Countries in bold are those reporting to the BIS. The full list is as follows: Andorra; Anguilla; Antigua and Barbuda; Aruba; **Bahamas**; **Bahrain**; Barbados; Belize; **Bermuda**; British Virgin Islands; **Cayman Islands**; Cook Islands; Costa Rica; **Netherlands Antilles/Curaçao**; **Cyprus**; Dominica; Gibraltar; Grenada; Guatemala; **Guernsey**; **Hong Kong,** China; **Isle of Man**; **Jersey**; Lebanon; Liechtenstein; **Luxembourg**; **Macau, China**; **Malaysia**; Malta; Marshall Islands; Mauritius; Monaco; Montserrat; Nauru; Niue; Palau; **Panama**; Saint Kitts and Nevis; Saint Lucia; Saint Vincent and the Grenadines; American Samoa; San Marino; Seychelles; **Singapore**; **Switzerland**; Turks and Caicos Islands; United Arab Emirates; Uruguay; and Vanuatu.[10]

In the headline results in Section 2.1, the analysis focuses on a decline in deposits in those IFCs from the list above that report to the BIS since 2006, in order to work with a balanced panel and avoid the effect of new reporting countries. The headline results are reported as declines in IFC deposits from nonbank counterparties in all countries including all IFCs. In the headline results, the sample excludes the Cayman Islands, based on the particular nature of the U.S.-Cayman Islands relationship outlined in Section 2.1. For confidentiality reasons, it is not possible to report the overall aggregated decline in deposits with just the Cayman Islands-U.S. series removed, so the entire Cayman Islands series is removed together with Hong Kong, China, and Macau, China in *Figure 2*.

In the regression analysis in Section 3, the sample is different, as not all jurisdictions that provide aggregated data provide bilateral data that can be used in the regression analysis. The panel used in the regression analysis is unbalanced. The analysis relies on a regression for all available country-pairs where there are sufficient quarters with and without EOI to estimate the effects. One exception is that in this sample, the U.S.-Cayman Islands series is removed, but the series between Cayman Islands and other jurisdictions are kept in the sample. This means that the sample underlying the headline decline of USD 410 billion reported in Section 2.1 and the sample underlying the association with EOIR and AEOI are slightly different.

Section 4.3 contains a robustness analysis of the main results in the paper to the inclusion of different IFCs subject to data availability.

[10] In the BIS LBS, the following jurisdictions report on an aggregated basis or as part of other reporting jurisdictions: Anguilla, Antigua and Barbuda, British Virgin Islands, Cook Islands, Monaco, Montserrat, Niue, Saint Kitts and Nevis, and American Samoa. Given this aggregation, these IFC jurisdictions cannot be analysed separately. For further information, see BIS (2017), https://www.bis.org/statistics/dsd_lbs.pdf.

2.2 Stylised facts on the expansion of the EOI network

Figure 3 shows the expansion of EOI of various forms over the course of the last 10 years (see Box 4 for further discussion). There is a steady increase in the global number of bilateral EOIR relationships from 2009 to 2018 (the blue dashed line). However, more striking than the increase in total EOIR relationships is the extent to which this increase is driven by MAC signatures. The number of global MAC-based EOIR relationships expands dramatically post-2012. The chart also shows the dramatic expansion in AEOI—first following the commitment of G20 countries to exchange information automatically in September of 2014, with increasing commitments over the course of 2014.

Figure 4 shows the expansion of EOIR in IFCs over the period from 2008 to 2018. The figure shows, for each jurisdiction, the number of EOI relationships of all kinds (under Tax Information Exchange Agreements (TIEAs), Double Tax Conventions (DTCs), European Union Directives, the MAC, or any other relevant transparency agreements). The blue line shows the number of EOI relationships that existed for each jurisdiction under the MAC. The flat blue lines in many jurisdictions, followed by sharp rises, serve to highlight the date of MAC signature. It is important to highlight that in some countries, a MAC signature comprises a larger share of the total EOIR relationships than in others. It is clear, for example, that Switzerland had a large EOIR network prior to a MAC signature. This means that many of the EOIR relationships established by Switzerland under the MAC already existed under other agreements. However, for other jurisdictions, such as Montserrat, for example, it can be noted that agreements under the MAC constitute the vast majority of the EOIR relationships in which the jurisdiction participates.

Consideration of the impact of the MAC is particularly important, as this has not been taken into account by previous studies. To our knowledge, none of the major studies in the literature on the impact of EOI have accounted for the relationships generated by the MAC signature in the analysis. Johannesen and Zucman (2014) write that a "comprehensive multilateral agreement would prevent tax evaders from transferring their funds from haven to haven." The MAC performs exactly this function.

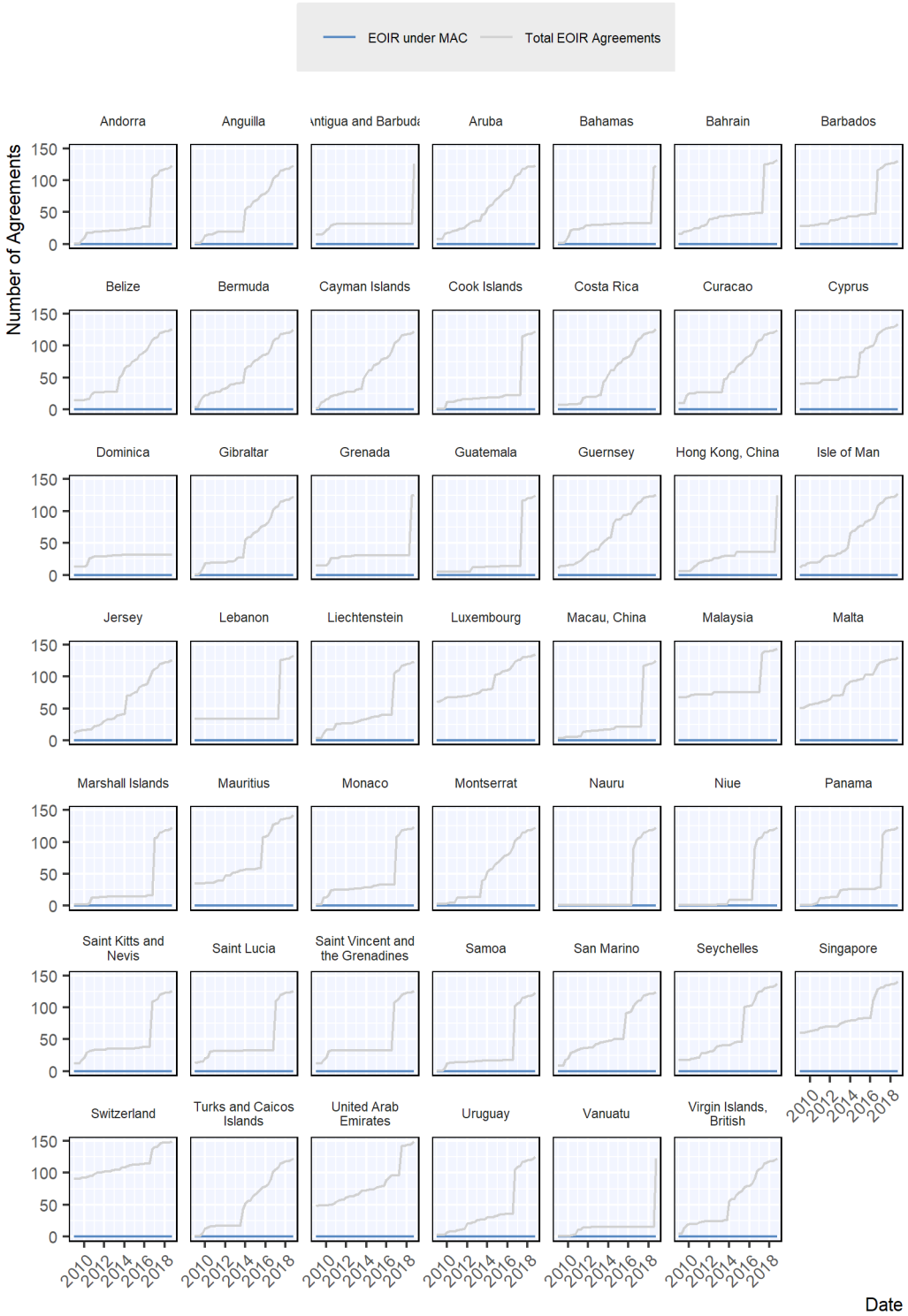
FIGURE 3. Number of bilateral EOI relationships

NOTE: Data on bilateral EOIR agreements post-2017 are preliminary and subject to revision. "EOIR agreement signed" refers to the signature of any agreement that establishes an EOIR relationship, including TIEAs, DTCs, and the MAC itself. To avoid double-counting, agreements that establish an EOIR relationship where one was already in place are not included (e.g., instances where two countries sign a DTC that provides for EOIR where a TIEA already provided for EOIR between the two countries).
SOURCE: Data on information exchange agreements provided by the Global Forum.

Box 4: A timeline of the expansion of tax transparency

- *April 2009*: The London G20 summit Communique states that the G20 agree, "to take action against noncooperative jurisdictions, including tax havens. We stand ready to deploy sanctions to protect our public finances and financial systems. The era of banking secrecy is over."
- *September 2009*: Global Forum on Transparency and Exchange of Information for Tax Purposes, which earlier comprised OECD countries working with Financial Centres, restructured, and membership opened up to all those who commit to the Standard of Exchange of Information on Request (EOIR) and agree to undergo a peer review to assess its implementation.
- *2010–2011*: The Convention on Mutual Administrative Assistance in Tax Matters developed jointly by the OECD and the Council of Europe in 1988 and amended by Protocol in 2010. Opened for signature to non-OECD and non-Council of Europe countries in 2011.
- *2010*: The U.S. Foreign Account Tax Compliance Act enters into law.
- *September 2013*: The G20 Leaders endorse the OECD proposal for a global model of AEOI and invite the OECD, working with G20 countries, to present such a new single standard for AEOI.
- *August 2014*: The Global Forum puts in place a commitment process to enable its members to commit publicly to a timetable to implement the new AEOI Standard. All Global Forum members, other than developing countries that do not house a financial centre, are asked to commit to begin automatically exchanging information in accordance with the Standard, reciprocally and with appropriate partners, by 2017 or 2018.
- *September 2014*: The full AEOI Standard is endorsed by the G20 Finance Ministers at their meeting in Cairns.
- *October 2014*: The Global Forum announces commitments to implementation of AEOI, with exchanges to commence by September 2017 or September 2018.
- *October 2014*: The first jurisdictions sign the Common Reporting Standard Multilateral Competent Authority Agreement at the sidelines of the Global Forum Plenary Meeting in Berlin.
- *2015*: Commitments to the new AEOI Standard announced by Bahrain, Panama, Cook Islands, Nauru, and Vanuatu, subsequent to the Panama Papers Data Leaks.
- *September 2017*: The first exchanges take place under the new AEOI Standard.
- *End of 2018*: Exchanges taking place under the AEOI Standard now cover 90 jurisdictions.

FIGURE 4. EOIR agreements and MAC agreements over time. (Total number of EOIR agreements signed by each jurisdiction)



NOTE: The list of IFCs is based on IMF (2000).
SOURCE: Data on information exchange agreements provided by the Global Forum.

3. Investigating the impact of EOI on cross-border bank deposit holdings

The previous section highlighted that there have been substantial reductions in the size of bank deposits in certain IFCs reported to the BIS. A challenge in assessing the impact of EOI is attempting to identify the extent to which these reductions are a result of EOI or of other factors. There are several aspects to consider. First, changes in bank deposits will be impacted by nontax factors such as the attractiveness of a jurisdiction's investment and legal environments, its overall economic growth, and recent or impending regulatory changes. Second, even if changes in bank deposits are tax-driven, they may not be due to changes in EOI. For example, in as much as deposits in IFCs potentially represent hidden wealth, it is possible that these deposits have been reduced as a result of other forms of tax enforcement such as targeted audits. In addition, the major data leaks in recent years may have provided information to tax authorities to address tax evasion.

In addition, the extent to which offshore bank deposits represent hidden wealth is by no means clear. From a tax perspective, assets held offshore may be fully compliant with tax rules. Where this is the case, these deposits would be expected to be unresponsive to EOI. Changes in IFC bank deposits may also respond to other contemporaneous tax factors including changes in the tax environment of the IFC and the home jurisdiction of the capital owner. These could include changes in statutory rates or changes in tax rules, such as those that might result from implementing the OECD/G20 Base Erosion and Profit Shifting (BEPS) package, or temporary voluntary disclosure programmes to incentivise disclosure of funds hidden abroad. Disentangling these various effects constitutes a significant challenge.

There is complementary evidence that there has been significant disclosure of previously undisclosed assets, discussed further in Section 4.2. Since the widespread adoption of EOI, an estimated 500,000 individuals have disclosed offshore assets through voluntary disclosure programmes and around EUR 95 billion in additional tax revenue has been identified as a result of voluntary compliance mechanisms and offshore investigations.[11] The fact that these sums were in large part disclosed through voluntary disclosure programs (see Section 4.2) set up in advance of the commencement of AEOI in 2017, points to a relationship between taxpayer behaviour and EOI. However, it is important to recognise that the assets held in foreign jurisdictions that were disclosed may not have been repatriated. These assets could have stayed in foreign jurisdictions. This means that, while quantitative evidence of the link between EOI and reductions in cross-border bank deposits in IFCs is important, it is only one part of the overall picture.

3.1 Key hypotheses and methodological approach

While the decline in overall bank deposits in IFCs provides some suggestive evidence of the impact of EOI, it does not fully analyse the impact of EOI at a bilateral level. It is useful to turn to regression analysis to investigate further, whether the advent of EOI can be associated with changes in bank deposits. The key expectation is that, to the extent that some fraction of deposits of banks in IFCs have historically existed for the purposes of tax evasion, the expansion of EOIR and the introduction of AEOI should have made holding assets in EOI jurisdictions riskier.[12] The expected response is that taxpayers would remove their assets from IFCs that commit to, sign, or implement EOI agreements with non-IFCs.[13] This leads to the following hypothesis:

H1: An EOIR agreement between a given IFC and a given non-IFC triggers a reduction in bank deposits held in the IFC by residents of the non-IFC.

This hypothesis is tested using the following general regression equation:

$$log(Deposits_{ijq}) = \alpha + \beta\, EOI_{ijq} + \epsilon_{ijq}\,, \tag{1}$$

where $Deposits_{ijq}$ denotes the bank deposits held in jurisdiction i by residents of jurisdiction j in quarter q. This paper focuses on deposits in countries that are IFCs.[14, 15] It relies on an unbalanced panel of 16 IFCs based on

[11] OECD (2019), http://www.oecd.org/tax/oecd-secretary-general-tax-report-g20-leaders-june-2019.pdf.

[12] The approach here follows closely that of Johannesen and Zucman (2014) as well as subsequent examinations of this issue by Casi, Spengel and Stage (2018) and Menkhoff and Miethe (2019). These papers, in turn, build on Huizinga and Nicodème (2004), who used only one year of data as opposed to a panel approach.

[13] This may occur at the time of announcement, signature, ratification or entry into force.

[14] This excludes confidential bilateral data that are not available.

[15] This is not to discount the fact that deposits in non-IFC jurisdictions could respond to EOI as well. Section 4 examines potential deposit reactions between non-IFCs and other non-IFCs as well as between IFCs and other IFCs. The issue of "inward" deposit flows is explored further in Menkhoff and Miethe (2019).

the earlier list with sufficient bilateral deposit relations available.[16] The IFCs included are Bahrain; Bahamas; Bermuda; Netherlands Antilles/Curaçao; Cayman Islands; Cyprus; Guernsey; Hong Kong, China; Isle of Man; Jersey; Luxembourg; Macau, China; Malaysia; Panama; Singapore; and Switzerland. EOI_{ijq} is a dummy variable that denotes whether any kind of EOI relationship exists in quarter q between jurisdictions i and j.

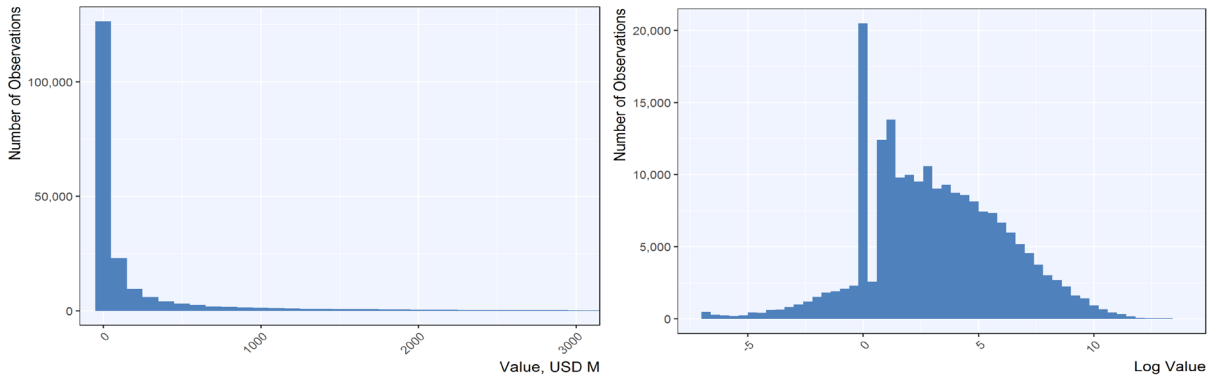
This paper examines the two main forms of EOI that have expanded in recent years: EOIR and AEOI. The independent variable for EOIR is the signature of a bilateral or multilateral agreement providing for EOIR. Such an agreement could be a bilateral agreement such as a DTC, a TIEA, or any other relevant multilateral transparency agreement, such as when two jurisdictions sign the MAC. As stated above, signatures of the MAC have particularly expanded during the post-2012 period and have accounted for the majority of EOIR relationships since then.

The independent variable for AEOI is either a public commitment to exchange information automatically or the commencement of AEOI under the CRS, or signature of a FATCA Intergovernmental Agreement (IGA).[17] All three different approaches to assessing the impact of AEOI are tested below. Taxpayers may have responded to such agreements with varying speeds. Some taxpayers may have responded at the earliest possible date, declaring deposits to tax authorities or shifting them out of IFC jurisdictions with the advent of expanded EOI, or they may have waited until the last possible moment before EOI would come about. This means that it is useful to examine separately both the announcement of commitment to AEOI as well as the commencement of exchange under AEOI agreements to capture behavioural responses of taxpayers, who may change their behaviours either upon announcement of the upcoming changes in the EOI environment, or at the time of the actual commencement of AEOI exchanges.

According to hypothesis H1 above, it is expected that the sign of the coefficient will be negative for deposits of IFCs with respect to non-IFCs. The regression approach uses log deposits as a dependent variable to account for the substantial skewness of bank deposits in the BIS dataset.

Figure 5 shows the distribution of the BIS values (left panel), and the distribution of the logged BIS values (right panel). The distribution of the logged values more closely approximates the normality assumption. This means that the regression results should be interpreted as percentage changes in bank deposits.[18] Moreover, in equation (1) denotes a jurisdiction-pair-year-quarter specific error term that is modelled in various ways as discussed below.

FIGURE 5. Distribution of BIS data



NOTE: Data are provided for nonbank counterparties only. Data are aggregated across currencies, sectors, reporting institutions, and instrument type.
SOURCE: Authors' calculations based on the BIS LBS.

[16] Each IFC has on average 74 different bilateral deposit relations per year-quarter. To profit most from the data available, an IFC-non-IFC pair has been included when at least four quarters of data were available either side of the relevant EOI independent variable. While earlier studies such as Johannesen and Zucman (2014), Casi, Spengel and Stage (2018) or Beer, Coelho and Leduc (2019) also used unbalanced panels, others like Ahrends and Bothner (2019) or Menkhoff and Miethe (2019) employed balanced panels largely at the expense of IFC coverage.

[17] Dates for the commencement of AEOI are taken from Automatic Exchange of Information Implementation Report 2018 (Global Forum on Transparency and Exchange of Information for Tax Purposes (2018)) page 3. AEOI agreements are activated on a bilateral basis and exchanges are also bilateral, which is not taken into account in this analysis. Jurisdiction-pairs are coded 1 if both jurisdictions have begun exchanging information under the CRS or under FATCA, and zero otherwise. However, this does not necessarily mean that they are exchanging with each other. The details of which jurisdictions have exchanged with each other are not public at this stage. Incorporating actual activated bilateral agreements could be an avenue for future enhancement of this work.

[18] To obtain estimates of the percentage impacts of EOI, the following transformation is applied to the estimated coefficients: 100 * (exp(β)–1).

3.2 Main results for liabilities of IFCs with respect to non-IFCs

The results of the analysis are presented first with jurisdiction-pair fixed effects, and then with both jurisdiction-pair fixed effects and time-fixed effects. Time-fixed effects work to account for many nontax factors (e.g., declines in interest rates) that could have also impacted bank deposits over this period. As will be discussed below, the presence of time-fixed effects complicates the interpretation of the results, because many significant changes in the EOI environment have proceeded quickly across all IFCs. It is thus challenging to separate the impact of EOI from the broader time-trends of IFC deposits that can be discerned in the data.

Omitting time-fixed effects

The first set of key regression results from the above equation are presented in *Table 3*. The dependent variable in this analysis is bank deposits in IFCs held by counterparties in non-IFCs. As discussed above, these models have jurisdiction-pair fixed effects, but omit year-quarter fixed effects, in contrast to much of the literature. In each instance, the results are presented with clustered standard errors at the jurisdiction-pair level. The regression equation is as follows:

$$log(Deposits_{ijq}) = \alpha + \beta_1 \ EOIR\ Signature_{ijq} + \mu_{ij} + \epsilon_{ijq}, \tag{2}$$

where μ_{ij} represents the dummy variable for the jurisdiction-pair *ij*. This means that the estimation of the impact of EOIR is averaging out the impact of a specific jurisdiction-pair relationship on cross-border bank deposits. This takes account of, for example, the fact that France and Switzerland may have higher expected cross-border bank deposits owing to their geographical proximity compared to, for example, Switzerland and Australia.[19]

The first column presents the specification with EOIR signature as the only independent variable. The coefficient on EOIR signature (β_1) is negative and statistically significant at the 1-percent level, suggesting that without controlling for either AEOI or time characteristics, EOIR signature is associated with a reduction in bank deposits held in IFCs of about 20 percent. In the following specifications, the other EOI variables are gradually added to control simultaneously for the different forms of EOI, to avoid omitted variable bias and to account for potential endogeneity in treaty adoption among jurisdictions.

The second set of results adds a dummy variable for the announcement of a commitment to AEOI commencement. This regression specification is as follows:

$$log(Deposits_{ijq}) = \alpha + \beta_1 \ EOIR\ Signature_{ijq} + \beta_2\ AEOI\ Announcement_{ijq} + \mu_{ij} + \epsilon_{ijq}. \tag{3}$$

In this specification, the coefficients on both β_1 and β_2 are negative and statistically significant. AEOI announcement is associated with an 18.6-percent reduction in bank deposits over and above EOIR signature, the coefficient for which falls from 20 percent to 12 percent.

The third set of results does not consider the impact of AEOI announcement but rather the impact of the commencement of automatic information exchange mechanisms, i.e., AEOI operational and FATCA in place. As discussed above, this helps to assess whether taxpayers respond to the announcement of AEOI or the commencement of AEOI. The regression specification is as follows:

$$log(Deposits_{ijq}) = \alpha + \beta_1\ EOIR\ Signature_{ijq} + \beta_2\ AEOI\ Commencement_{ijq} + \mu_{ij} + \epsilon_{ijq}. \tag{4}$$

The regression results show a negative association between EOIR (associated with a 14.7-percent decrease in bank deposits) as well as a larger negative association between AEOI commencement and bank deposits. The coefficient suggests a reduction of 31 percent in expected bank deposits in the aftermath of AEOI commencement.

[19] A jurisdiction-pair dummy facilitates controlling for all such invariant jurisdiction-pair specific effects without the loss of degrees of freedom that would come with separately controlling for distance, common language, common legal system, contiguous borders, and other jurisdiction-pair effects typically used in some cross-jurisdiction data analysis.

The subsequent set of results incorporate both AEOI announcement and commencement. The regression equation is as follows:

$$log(Deposits_{ijq}) = \alpha + \beta_1 EOIR\ Signature_{ijq} + \beta_2 AEOI\ Announcement_{ijq} + \beta_3 AEOI\ Commencement_{ijq} + \mu_{ij} + \epsilon_{ijq}. \quad (5)$$

The results in this instance are broadly consistent with the effect found in the previous two models, with EOIR signature being associated with a roughly 11-percent reduction in bank deposits in IFCs, AEOI announcement also being associated with an approximately 11-percent reduction, and AEOI commencement being associated with a -28 percent impact.

The last specification in Column 5 is similar to equation 5 as it includes again EOIR signature and AEOI announcement. In addition, it adds a variable on AEOI commencement without accounting for established FATCA-IGA relationships to test both impacts separately. Both EOIR and AEOI announcements exert the same negative 10.6-percent effect on IFC deposits. The coefficient from the AEOI commencement variable excluding FATCA is, with an estimated impact of -30 percent, significantly higher. It has a slightly larger impact than the previous combined AEOI commencement variable. It is important to note that the sample size in this specification is relatively small, as there are only five quarters after September 2017 (when AEOI was widely implemented) in the dataset. This suggests that a longer time series may give further support to this result.

TABLE 3. The effect of EOI on foreign-owned deposits in IFCs, with jurisdiction-pair fixed effects

Item	EOIR Only	EOIR and AEOI Announcement	EOIR and AEOI (incl. FATCA) Commencement	EOIR and AEOI (include FATCA) Announcement and Commencement	EOIR and AEOI Announcement and Commencement
EOIR Signature	-0.219***	-0.128***	-0.159***	-0.116***	-0.112***
	(0.046)	(0.042)	(0.044)	(0.042)	(0.042)
AEOI Announcement		-0.206***		-0.115**	-0.112**
		(0.053)		(0.052)	(0.051)
AEOI Commencement					-0.354***
					(0.044)
AEOI (incl. FATCA) Commencement			-0.374***	-0.322***	
			(0.050)	(0.045)	
R^2	0.011	0.017	0.023	0.025	0.026
Number of Observations	29,461	29,461	29,461	29,461	29,461
Jurisdiction-Pair FEs	Yes	Yes	Yes	Yes	Yes
Year-Quarter FEs	No	No	No	No	No

NOTE: Regression of foreign-owned bank deposits in IFCs on EOI dummy variables. The dependent variable is the stock of deposits held by savers of jurisdiction *i* in banks of IFC *j* at the end of quarter q. The unit of observation is the jurisdiction-pair (*i, j*) and the sample period goes from Q1 of 2006 to Q4 of 2018. Data are provided for nonbank counterparties only.

*** and ** represent statistical significance levels of 1 percent and 5 percent respectively.

The countries used as reporting IFCs in this regression are: Bahrain, Bermuda, Bahamas, Cayman Islands, Cyprus, Netherlands Antilles/Curaçao, Guernsey, Hong-Kong, Isle of Man, Jersey, Luxembourg, Macau (China), Malaysia, Panama, Singapore, and Switzerland. The Cayman-U.S. series has been removed from the regression as outlined in Section 2.1

SOURCE: Authors' calculations based on BIS LBS, and data on information exchange agreements provided by the Global Forum.

[20] Huizinga and Nicodème (2004) do not use time-fixed effects as they have only 1 year of data. However, all other papers looking at this issue follow this approach.

Including time-fixed effects

In *Table 4*, the approach follows the literature and includes year-quarter fixed effects.[20] Time fixed effects factor out events at specific times that may have affected all IFCs in a similar way, such as the financial crisis or global regulatory changes.

The regression equation becomes as follows:

$$log(Deposits_{ijq}) = \alpha + \beta_1 EOIR\ Signature_{ijq} + \mu_{ij} + \theta_q + \epsilon_{ijq}, \quad (6)$$

where the term θ_q represents the specific time effect of each year-quarter q on log-bank deposits.

When year-quarter fixed effects are accounted for, the size of many coefficients in the regressions shrinks substantially or becomes nonsignificant. EOIR signature is now associated with a small and not-statistically significant decrease in IFC bank deposits of between 2 percent and 4 percent. AEOI announcement is also no longer significant despite the expected sign on the coefficient. Both AEOI commencement variables, however, continue to be associated with a strong decrease in deposits. While the AEOI and FATCA combined variable exerts an impact of between -17 percent and -18 percent, the AEOI-only dummy indicates again an even higher negative effect of around 22 percent. All AEOI commencement variables are significant at the 1-percent level.

TABLE 4. The effect of EOI on foreign-owned deposits in IFCs, with jurisdiction-pair and year-quarter fixed effects

Regression of foreign-owned bank deposits in IFCs on EOI dummy variables

Item	EOIR Only	EOIR and AEOI Announcement	EOIR and AEOI (including FATCA) Commencement	EOIR and AEOI (including FATCA) Announcement and Commencement	EOIR and AEOI Announcement and Commencement
EOIR Signature	-0.024	-0.028	-0.041	-0.042	-0.043
	(0.044)	(0.044)	(0.045)	(0.045)	(0.044)
AEOI Announcement		-0.074		-0.041	-0.033
		(0.066)		(0.064)	(0.064)
AEOI Commencement					-0.249***
					(0.062)
AEOI (including FATCA) Commencement			-0.199***	-0.185***	
			(0.068)	(0.062)	
R2	0.0001	0.0005	0.002	0.002	0.003
Number of Obs.	29,461	29,461	29,461	29,461	29,461
Jurisdiction-Pair FEs	Yes	Yes	Yes	Yes	Yes
Year-Quarter FEs	Yes	Yes	Yes	Yes	Yes

NOTE: The dependent variable is the stock of deposits held by savers of jurisdiction i in banks of IFC j at the end of quarter q. The unit of observation is the jurisdiction-pair (i, j) and the sample period goes from Q1 of 2006 to Q4 of 2018. Data are provided for nonbank counterparties only.

*** represents a statistical significance level of 1 percent.

The countries used as reporting IFCs in this regression are: Bahrain, Bermuda, Bahamas, Cayman Islands, Cyprus, Netherlands Antilles/Curaçao, Guernsey, Hong-Kong, Isle of Man, Jersey, Luxembourg, Macau (China), Malaysia, Panama, Singapore, and Switzerland. The Cayman-U.S. series has been removed from the regression as outlined in Section 2.1

SOURCE: Authors' calculations based on BIS LBS, and data on information exchange agreements provided by the Global Forum.

Consistent with the literature on this topic, these results continue to show the robust negative association of AEOI implementation on bank deposits in IFCs. Compared to other relevant studies in the field, estimates in this paper end up in the middle of an AEOI impact range of between -3.1 percent and -34.9 percent (see

[21] Bilicka and Fuest (2014) also find that jurisdictions are more likely to initially sign EOI agreements with jurisdictions with which they have stronger economic ties. This may be a partial explanation as to why EOIR agreements signed earlier may exert a stronger impact on deposit flows between jurisdictions.

Figure 1). The findings come closest to Beer, Coelho and Leduc (2019), who use an unbalanced sample with a slightly reduced coverage of IFCs and sample length. They report an average effect of about -25 percent exerted by AEOI commencement on IFC deposits.

The null results with respect to EOIR in *Table 4* stand in contrast to work by Johannesen and Zucman (2014) as well as Menkhoff and Miethe (2019), who demonstrate statistically significant negative results of -11 percent and -24 percent respectively. To examine this further, *Table 5* re-estimates the model specification of *Table 4* for EOIR only. As in Johannesen and Zucman (2014), the beginning of the sample period considered in the analysis is Q4 of 2003 and the end of the sample period varies from Q4 of 2011 up to Q4 of 2014. This facilitates the examination of whether the impact of EOIR signature has varied over time.

TABLE 5. The impact of EOIR over time

Item	EOIR only Sample length: Q1 2006–Q4 2018	EOIR only Sample length: Q4 2003–Q4 2011	EOIR only Sample length: Q4 2003–Q4 2012	EOIR only Sample length: Q4 2003–Q4 2013	EOIR only Sample length: Q4 2003–Q4 2014
EOIR Signature	-0.024	-0.066	-0.106*	-0.095*	-0.093*
	(0.044)	(0.056)	(0.055)	(0.051)	(0.049)
R2	0.0001	0.001	0.002	0.001	0.001
Number of Obs.	29,461	16,169	18,585	21,065	23,834
Jurisdiction-Pair FEs	Yes	Yes	Yes	Yes	Yes
Year-Quarter FEs	Yes	Yes	Yes	Yes	Yes

NOTE: Regression of foreign-owned bank deposits in IFCs on EOIR signature for varying sample lengths. The dependent variable is the stock of deposits held by savers of jurisdiction i in banks of IFC j at the end of quarter q. The unit of observation is the jurisdiction-pair (i, j) and the maximum sample period goes from Q4 of 2003 to Q4 of 2014. Data are provided for nonbank counterparties only. Data are aggregated across currencies, sectors, reporting institutions, and instrument type.

* represents a statistical significance level of 10 percent.

SOURCE: Authors' calculations based on LBS, BIS, and data on information exchange agreements provided by the Global Forum.

Table 5 demonstrates the impact of expanded EOIR agreements during the early years of the EOIR Standard and confirms previous results in the literature. Whereas in Column 1 the original sample does not yield significant results with respect to EOIR impact, subsequent estimates show some significant results at the 10 percent levels that are decreasing in size with the lengthening time series. Column 3 reports an effect on IFC deposits of about -10 percent during the period from Q4 of 2003 to Q4 of 2012, which is close to the estimate reported by Zucman and Johannesen (2014), in spite of the different cross-country sample. Adding additional years up to Q4 of 2014, however, decreases the impact to about 8.5 percent. Menkhoff and Miethe (2019) document a similar weakening effect of EOIR over time.

This change in impact could be explained by the nature of the country-pairs experiencing changes in EOI relationships over this period. As more and more countries signed the MAC, more and more EOI relationships were coming into place (see *Figure 3* above). As MAC coverage became close to comprehensive, the multilateral nature of the MAC meant many of these relationships were among countries that had little or no bilateral cross-border financial activity that might be impacted by the MAC.[21] Countries signing the MAC established potential EOIR relationships with every other signatory, whether there was substantial volumes of cross-border banking activity or not. This may account for the relative decline in the size of the impact of EOIR over time.

3.3 Accounting for multicollinearity

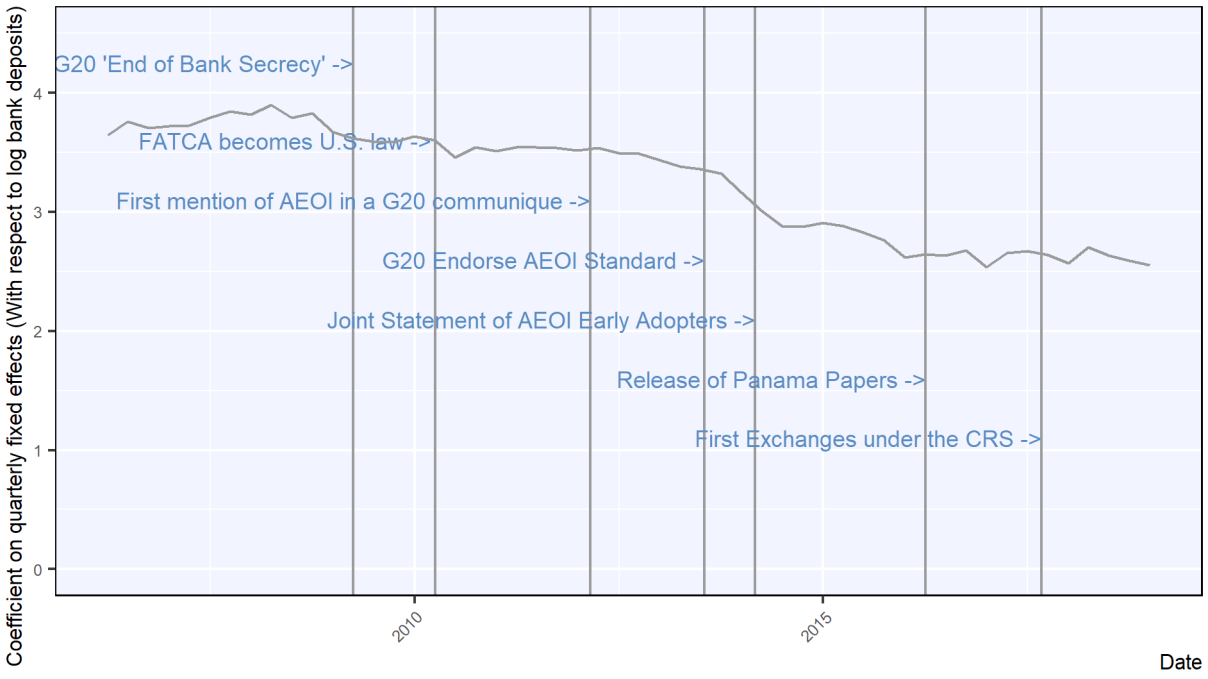
The reduction in the size and significance of the coefficients suggests that time-fixed effects explain some of the effects previously attributed to EOI.[22] This is complicated by the fact that changes in several of the independent variables are concentrated in certain periods. This suggests that there is some multicollinearity between

[22] This is also evidenced by the notable decline in the R2 statistics between Tables 3 and 4 due to the time-fixed effects absorbing some of the variation in the data.

specific events factored out by time-fixed effects and the EOI variables, which may imply that the time-fixed effects capture some of the impact of the changes in the EOI environment found in *Table 4*. To see this, it is useful to examine the fixed effects as well as the time trends in the independent variables themselves.

Figure 6 shows these fixed effects over time. There is an overall decline in bank deposits in IFCs being captured by the quarterly fixed effects. Several of these periods of substantial declines coincide with changes in the EOI environment, either through substantial increases in the expansion of both EOIR (i.e., through the expansion of the MAC) and through public commitments to AEOI, most notably in the period from the end of 2013 to the end of 2014.

FIGURE 6. Year-quarter fixed effects over time



NOTE: Based on the regression of EOIR and AEOI commencement with jurisdiction-pair and year-quarter fixed effects The dependent variable is the stock of deposits held by savers of jurisdiction i in banks of IFC j at the end of quarter q.

SOURCE: Authors' calculations based on BIS LBS.

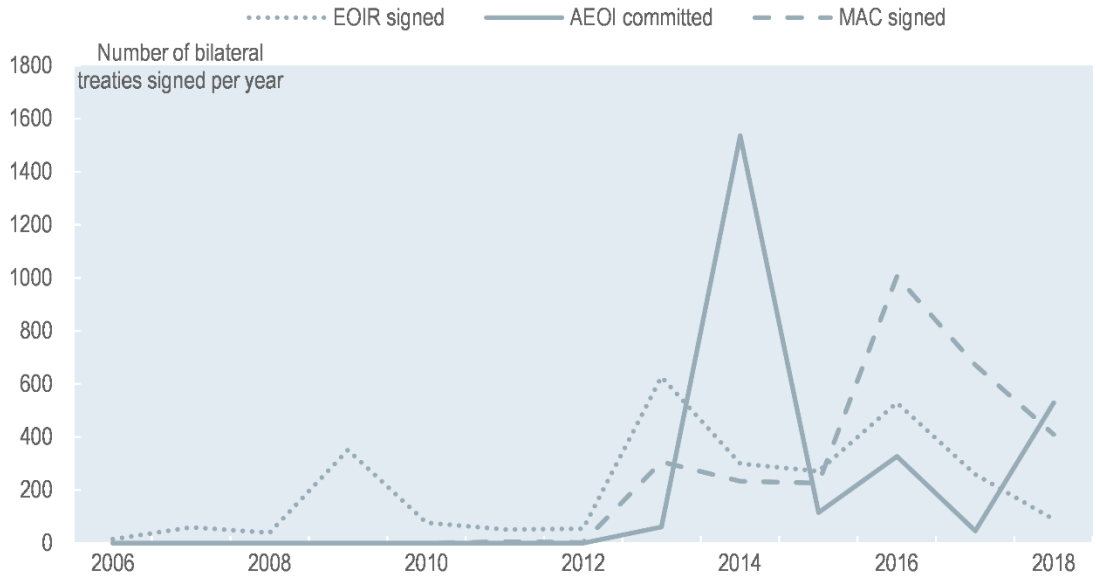
Figure 7 shows that most IFC jurisdictions declared commitments to AEOI over this period. The periods of highest new signature levels are also the periods of the sharpest declines in the fixed effects.

Around the same time, other countries such as, for instance, Switzerland, entered into economically important bilateral treaties (such as the U.S.-Swiss Bank Program in August 2013) and then experienced significant declines of foreign-owned deposits (*Figure 8*). Over the course of the quarters covered, the trend effect shows several reductions (albeit of varying sizes) that coincide with key events in the tax transparency timeline. This includes after FATCA became law in the United States as well as in the aftermath of early signals that AEOI would expand beyond the United States' FATCA legislation (e.g., the first time AEOI is mentioned in a G20 Communique).

This, in turn, suggests that certain events in the timeline of the expansion of tax transparency are associated with decreases in bank deposits in IFCs. However, the fact that these events are collinear with AEOI announcement dates makes this effect difficult to conclusively associate with AEOI in the regression specification.[23]

[23] Some mild multicollinearity between the time dummies and the AEOI announcement variables has also been detected based on a somewhat elevated variance inflation factor (VIF) and the Farrar-Glauber test.

FIGURE 7. Changes to the EOI environment over time



NOTE: The figure shows the number of bilateral treaties signed in each year.
SOURCE: Data on information exchange agreements provided by the Global Forum.

Quantifying the impact of the AEOI Joint Announcement

In March 2014, some 44 jurisdictions jointly announced their commitment to AEOI at the same time (referred to hereinafter as the Joint Announcement).[24] This substantial number of jurisdictions participating in the Joint Announcement provides the opportunity to analyse the potential impact of EOI on a subsample of IFCs in more detail, allowing us to check for a diluting effect of multicollinearity and establish further the robustness of the results presented in Section 3.2.

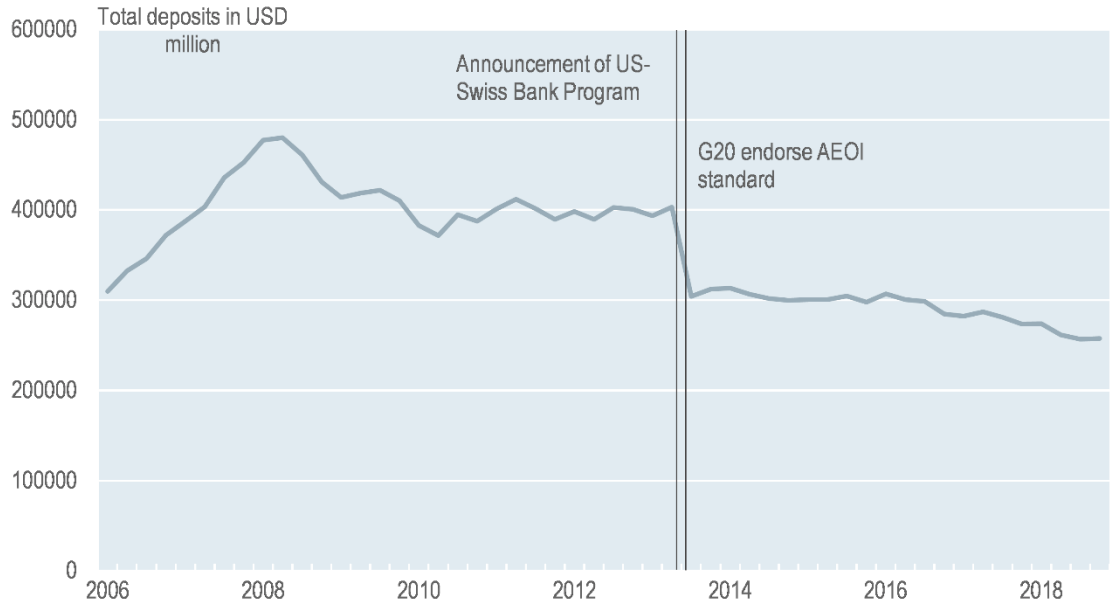
Among those jurisdictions that were part of the Joint Announcement, six IFCs provide bilateral data in the sample available from the BIS.[25] Combining the data for these IFCs with other early-adopting non-IFC jurisdictions allows the examination of their bank deposits relative to those of other jurisdictions that did not participate in the Joint Announcement.[26] The analysis relies on a subsample of the bilateral deposit database, which is composed of two different jurisdiction pairs, namely those that announced early and others that did not. *Figure 9* illustrates this, whereby the IFC-non-IFC pairs that both participated in the Joint Announcement can be compared to those IFC-non-IFC country pairs that did not. This allows the examination of the impact of many jurisdictions publicly committing to implementing AEOI at the same time and addresses the issue of multicollinearity that makes it difficult to assess this through the regression specification above. This is because for a short period, a set of IFCs and non-IFCs had publicly committed to AEOI while another set had not. By comparing these two groups, it is possible to assess the impact of public commitment on bank deposits.

[24] The joint announcement jurisdictions are Anguilla, Argentina, Belgium, Bermuda, British Virgin Islands, Bulgaria, Cayman Islands, Colombia, Croatia, Cyprus, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Gibraltar, Greece, Guernsey, Hungary, Iceland, Isle of Man, India, Ireland, Italy, Jersey, Latvia, Liechtenstein, Lithuania, Malta, Mexico, Montserrat, Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, South Africa, Spain, Sweden, Turks and Caicos Islands, and the United Kingdom.

[25] These are Bermuda, Cayman Islands, Cyprus, Guernsey, Isle of Man, and Jersey.

[26] There are twelve other jurisdictions in the sample for which there are bilateral data available. These twelve other jurisdictions committed later: Bahamas; Netherlands Antilles/Curaçao; Hong Kong, China; Luxembourg; Macau, China; Malaysia; Singapore, and Switzerland in October 2014; Bahrain and Panama in May 2016.

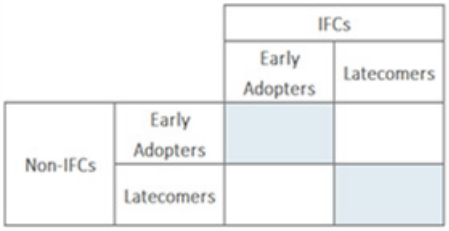
FIGURE 8. Foreign-owned deposits in Switzerland



NOTE: The variable is the stock of deposits held by foreign savers in Swiss banks at the end of each quarter. The vertical lines indicate respectively the joint announcement of the U.S.-Swiss Bank Program by U.S. and Swiss authorities on August 29, 2013, and the G20 endorsement of the AEOI Standard on September 5-6, 2013, at the G20 St. Petersburg summit, held at the Constantine Palace in St. Petersburg, Russia.

SOURCE: Authors' calculations based on BIS LBS, and data on information exchange agreements provided by the Global Forum.

FIGURE 9. Composition of different subsamples for the difference-in-differences estimation



NOTE: Both panels in blue can be compared to each other.

It is assumed that responses to the AEOI Joint Announcement in the form of reductions in bank deposits in IFCs should occur between early-adopting jurisdiction pairs and leave the jurisdictions that commit at a later stage relatively unaffected. An approach similar to Johannesen (2014) is used to test this assumption, estimating using OLS an extended version of a regular two-period difference-in-differences model such as below:

$$log(deposits)_{ijt} = \alpha + \mu_{ij} + \gamma_t \theta_t + \delta_t \theta_t * EA_{ij} + \varepsilon_{ijt}, \quad (7)$$

where μ_{ij} is a set of jurisdiction-pair dummies, θ_t is a set of year-quarter fixed effects and EA_{ij} is an indicator variable coded as one whenever a jurisdiction pair belongs to the group of early adopters and zero otherwise. As the joint announcement of jurisdictions to adopt AEOI happened in March 2014, the first quarter (Q1) of 2014 becomes our reference quarter in the regression and consequently remains omitted.

The model estimates time trends in foreign-owned deposits among early-announcing jurisdiction pairs, (the treatment group), and those that commit at a later stage, the control group. Any significant divergence in trends around the time of the Joint Announcement, the first quarter of 2014, is interpreted as a causal effect of early AEOI commitment on bank deposits. Due to the inclusion of various fixed effects, results are

reported conditional on time-invariant jurisdiction-pair effects, accounting for gravity factors such as common language or geographical distance, and common time-varying year-quarter effects accounting for instance for global regulatory changes or financial crises (this approach is similar to that in Section 3.2 above). Estimated standard errors are robust and clustered at the jurisdiction-pair level, following the recommendation of Bertrand, Duflo and Mullainathan (2004).

The estimated treatment effect for a given post-announcement quarter t is captured by (δ_t). This parameter represents the difference in growth of deposits in early-adopting IFCs held by other early-adopting non-IFCs over deposit growth in the control group (the later-committing jurisdiction pairs) in every year-quarter as of 2014 Q1. The causal interpretation of the treatment effect relies on the strict assumption that only the IFCs within the treatment group encounter withdrawals of deposits upon early announcement.[27] The deposit time trend of early-announcing IFCs should thus follow a significantly different trajectory after 2014 Q1. In the absence of the Joint Announcement, both trends would follow roughly identical paths prior and post Joint Announcement. This implies that pretreatment trend differentials should be relatively negligible, with the coefficients of being relatively small and statistically insignificant.

Figure 10 shows the main results of this analysis: the estimated aggregated time trends for early-adopting jurisdiction pairs relative to non-early-adopting jurisdiction pairs.[28]

The two lines represent respectively the treatment and control group in the difference-in-differences estimation. The dotted line is the estimated time trends of foreign deposits in early-adopting IFCs held by early-adopting non-IFC counterparties. The solid line is the estimated time trend of foreign deposits held between jurisdiction pairs that committed later. The columns indicate the statistical significance of the interaction terms, the combined impact of being an early adopting IFC jurisdiction compared to non-early-adopting IFCs.

The results point to a notable common trend in both series of about 10 quarters preceding the Joint Announcement, which is followed by an increasing divergence of both trends after the first quarter of 2014. The estimated trend line of the treatment group declines considerably more than the control group trend amid an overall fall in IFC deposits. This is particularly the case in the first four post-announcement quarters.

The statistical significance of this diverging trend trajectory is confirmed by the bars on the bottom of the figure, which indicate rising significance directly following the Joint Announcement, surpassing the 5-percent level around the third quarter of 2014. The bars represent the p-values in the regression, so lower bars point to evidence of a statistically significant difference between the early-adopters and non-early-adopters. The very low bars after the Joint Announcement point to a statistically significant difference between those jurisdictions that announced and those that did not. Moreover, both trend lines fail to converge and continue their constant earlier decline after the Joint Announcement. This suggests that early AEOI announcement seems to trigger a permanent shift in the level of bank deposits of the six IFC treatment groups.

A comparison of average growth rates in deposits between late 2012 until the Joint Announcement and the third and fourth quarter of 2014 (i.e., deposits measured on 30 September and 31 December) provide further evidence for this divergence. While prior to the Joint Announcement growth rates move synchronically at around -1 percent, they drop by about 5 percent and 10 percent for the control and the treatment group respectively. These developments are mirrored by the similarity in the calculated treatment effect on the trend of the treatment group, which amounts to about -15 percent during the same period.[29] Averaged over the four post-announcement quarters (i.e. until 31 March 2015), the analysis suggests that the impact of AEOI joint announcement has a treatment effect on the early-announcing IFC jurisdictions of about -11 percent.[30]

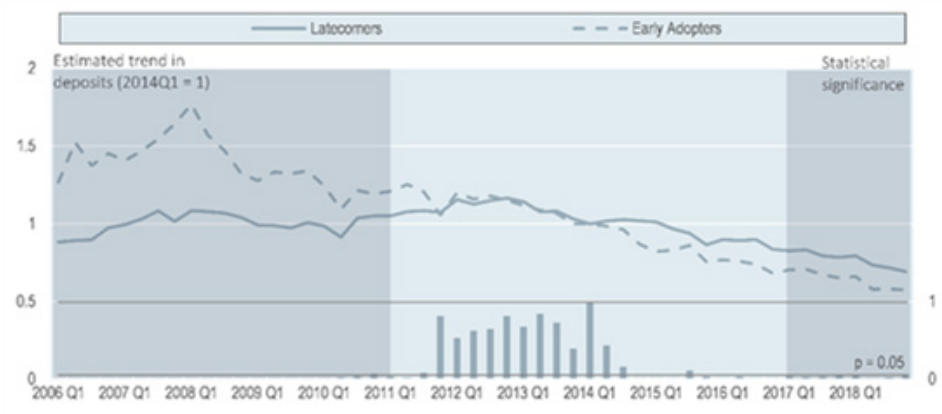
[27] It is important to note that a potential confounding weakness of this approach is whether jurisdictions that did not participate in the Joint Announcement were interpreted as committing the AEOI (e.g., if taxpayers suspected that even if they had not committed via the joint announcement, they would commit eventually).

[28] Detailed regression results are contained in Table A1 in the Annex.

[29] The treatment effect for a period t is calculated as exp(0)–exp(δ_t) for post-treatment values of t, where 0 under the identifying assumption is the expected, counterfactual value of (δ_t) without the treatment.

[30] The average over four quarters provides a more robust estimate as it smoothes the impact over the different quarters and accounts for seasonality and random variation in deposit series. The analysis of trend reactions beyond the four-quarter window does not seem to be reasonable because after this period more countries had committed to AEOI. This means that the difference between the treatment and control group declined over the course of 2014 and 2015.

FIGURE 10. The impact of AEOI Joint Announcement on time trends in IFC deposits



NOTE: Deposit trends based on difference-in-differences estimation Lines indicate trends in deposits as captured by coefficients on time dummies θ_t and the interaction terms $\theta_t * EA_{ij}$, that is $\exp(\gamma_t)$ for non-Early Adopters and $\exp(\gamma_t + \delta_t)$ for Early Adopters. Columns indicate statistical significance levels of interaction terms $\theta_t * EA_{ij}$. Areas shaded on both ends of the sample range direct the reader to a time window of analysis relevant for inspection due to being less likely influenced by other events than the Joint Announcement.
SOURCE: Authors' calculations based on BIS LBS data.

The impact on individual IFCs

To analyse further the impact of EOI on individual jurisdictions, it is useful to disaggregate the impacts of the Joint Announcement by country. Aggregating six different IFCs from across regions risks grouping underlying heterogeneity in the impact of AEOI on different jurisdictions. To examine further, the same difference-in-differences specification in Equation 7 is estimated one-by-one for each early-adopting IFC for which data are available from the BIS. The counterparty non-IFCs are split up again into early adopters (the treatment group) and those that announced later on (the control group). The respective figures are contained in Annex A.

As the estimations in Figures A1, A2, and A3 in the Annex show, there is substantial heterogeneity in the impact of the AEOI Joint Announcement on deposits in each country. The results suggest some signs of trend divergence for Guernsey and the Cayman Islands. As depicted by the bars in the individual figures, the interaction terms turn significant after the reference period 2014 Q1, with partial effects only during the four post-announcement quarters. Estimated effects over the four post-announcement quarters of the same period are estimated at around -53 percent for Guernsey and -27 percent for the Cayman Islands. The results suggest that Jersey was affected to a much lesser extent, as a very slight trend divergence and the barely significant interaction terms demonstrate. Cyprus shows a parallel decrease in both trends after the reference quarter with no significant drop for the early-adopting jurisdictions, suggesting a very modest impact of the announcement in that jurisdiction.

The trend results for Bermuda and the Isle of Man point to further heterogeneity and suggest that the AEOI Joint Announcement has increased deposits from early-adopting non-IFCs during some periods (Figure A3). This finding is counter-intuitive. However, most of the interaction terms in the four post-announcement quarters for Bermuda and the Isle of Man are not statistically significant, suggesting that the estimated effects are weak.

Overall, the difference-in-differences estimations indicate that the AEOI Joint Announcement in March 2014 had a small and relatively mild significant one-off impact on deposits across the six IFCs covered that were early adopters and for which detailed data are available. The effect on the individual IFCs varies considerably in size and statistical significance, pointing to a heterogeneous impact of EOI on different IFCs. These results contrast the regression results obtained earlier, which do not show a statistically significant impact of AEOI announcement. These results strengthen the findings in two ways. The statistically significant difference between early adopters and non-earlier adopters points to some degree to multicollinearity that is driving the statistical insignificance in *Table 4*. The underlying heterogeneous country effects are masked by estimated

average responses, which are picked up by the previous regressions with the larger sample. This raises an important qualifier to the headline result in this paper—the average effects of AEOI reported conceal important variation, with larger impacts in some countries and smaller impacts elsewhere.

4. Robustness checks

This section presents the analysis and results for establishing robustness of the main findings from the regression analysis above. These robustness checks are organised along three topics. First, the analysis considers whether the impacts of EOI changes are confined to IFC-non-IFC country pairs and examines the impact of EOI on deposits between non-IFCs and between IFCs. Second, the analysis incorporates into the main model a variable on voluntary disclosure and amnesty programmes to check whether these programmes, often implemented in jurisdictions around the same time as EOI initiatives, alter the main results. Third, the headline regression analysis is re-run on different samples of IFCs to ensure that the results are not driven by the specific list of IFCs used in the paper.

4.1 The effect of EOI across jurisdiction pairs

The results in Section 3 have shown a strong negative impact of AEOI commencement on bank deposits in IFCs owned by non-IFC jurisdictions. This is in line with expectations that the impact of EOI through potential noncompliant taxpayers would be concentrated in IFCs. However, the impact of EOI outcomes is strengthened if it is possible to highlight that this impact is confined to IFCs, and that non-IFC jurisdictions did not experience the same impacts as IFCs. For example, AEOI commencement should not trigger any significant reduction effect among deposits between non-IFCs and deposits with IFC counterparties only.

Table 6 shows the main regressions for deposits between non-IFC-IFC jurisdiction pairs from Section 3 estimated again, this time for non-IFC-non-IFC pairs (Columns 1 and 2) and for IFC-IFC jurisdiction pairs (Columns 3 and 4). The reported coefficients across all four columns on the AEOI commencements confirm the intuition. They do not exhibit significant negative effects on deposits held in the respective jurisdictions. The negative impact of EOI changes on cross-border bank deposits appears confined to deposits from non-IFCs into IFCs. Deposits between IFCs themselves are not affected in a significant way. Deposits from non-IFCs in other non-IFCs are also not affected in a statistically significant way.

In contrast, the results suggest that AEOI commitments had a positive impact on non-IFC deposits between each other. This can be interpreted as additional evidence of the impact of AEOI, suggesting that AEOI commitments appear to have spurred banking activity between non-IFC jurisdictions and point to an increasing shift in cross-border banking activity away from IFCs.

4.2 The potential impact of voluntary disclosure and amnesty programmes

The signature of EOIR treaties or AEOI commencement has in the past often coincided with the domestic implementation of Voluntary Disclosure and Amnesty Programmes (VDPs). Because these VDPs may have incentivised taxpayers with offshore deposits to declare or repatriate hidden assets, the presence of these VDPs may act as a confounding variable in the analysis above. That is, it is possible that the impacts found in the analysis of EOI are not actually results of EOI but rather of the VDPs that coincided with the expansion in EOI. This section assesses whether this is the case.

Table 7 assesses the impact of VDPs and shows results from the previous regression specification from *Table 4*, accounting for these programmes. To do this, a list of 38 VDPs in 27 countries is compiled. Some of these have been implemented since 2009 and some are still ongoing and are added as dummy variables to the regression specification.[31] An important caveat to these dummy variables is that the specifics of VDPs can differ significantly by jurisdiction in terms of length and legal consequences of disclosure. These different characteristics may result in varying impacts of the programmes and may influence the findings below.

[31] This list has been compiled based on sources from the OECD (2015), public notes from global audit firms such as PwC, Deloitte or KPMG as well as information scraped from national tax authority or finance ministry websites.

TABLE 6. The effect of EOI on foreign-owned deposits in different jurisdiction pairs

Item	EOIR and AEOI (including FATCA) Announcement and Commencement	EOIR and AEOI Announcement and Commencement	EOIR and AEOI (including FATCA) Announcement and Commencement	EOIR and AEOI Announcement and Commencement
	Non-IFC from Non-IFC	Non-IFC from Non-IFC	IFC from IFC	IFC from IFC
EOIR Signature	-0.033	-0.034	-0.065	-0.065
	(0.059)	(0.059)	(0.069)	(0.069)
AEOI Announcement	0.272**	0.275**	-0.14	-0.14
	(0.111)	(0.111)	(0.121)	(0.064)
AEOI Commencement		-0.030		-0.133
		(0.073)		(0.106)
AEOI (incl. FATCA) Commencement	-0.014		-0.133	
	(0.070)		(0.106)	
R2	0.004	0.004	0.002	0.002
Number of observations	23,860	23,860	15,645	15,645
Jurisdiction-pair FEs	Yes	Yes	Yes	Yes
Year-Quarter FEs	Yes	Yes	Yes	Yes

NOTE: Regression of foreign-owned bank deposits in different jurisdiction pairs on EOI implementation. The dependent variable is the stock of deposits held by savers of jurisdiction i in banks of either non-IFCs or IFC j at the end of quarter q. The unit of observation is the jurisdiction-pair (i, j) and the sample period goes from Q1 of 2006 to Q4 of 2018. Data are provided for nonbank counterparties only. Data are aggregated across currencies, sectors, reporting institutions, and instrument type.

** represents a statistical significance level of 5 percent.

SOURCE: Authors' calculations based on LBS, BIS, and data on information exchange agreements provided by the Global Forum.

The estimated models confirm the findings in *Table 4* of a statistically significant negative impact of both AEOI commencement variables on IFC deposits, albeit with the size of the coefficients slightly reduced. The coefficients on the VDP variable exhibit positive signs and are significant at the 1-percent level. These results contrast, for instance, with Menkhoff and Miethe (2019), who find no significant impact of VDPs on IFC deposits, based on a considerably smaller list of VDPs.

Several reasons may explain the estimated size and direction of the coefficients on the VDP variables. One possibility is that the existence of VDPs is endogenous to the size of bank deposits in IFCs; that jurisdictions that felt they had a large tax compliance challenge with respect to bank deposits implemented a VDP for this purpose.

Other explanations are possible, including the possibility that VDPs may reduce tax compliance by inducing some taxpayers to increase noncompliance afterwards or disclose outside of VDPs.[32] Finally, the fact that several of the VDPs in the list are still active may bias the results. Self-declarations may peak towards the end of VDP eligibility periods. Although conclusive evidence on the effect of VDPs is still subject to further research, the evidence presented shows that accounting for disclosure programmes does not seem to invalidate the expected negative impact of AEOI on foreign bank deposits.

[32] Langenmayr (2017), conducting a study on the 2009 VDP in the U.S., finds that the programme increased the number of individuals who evade taxes. She argued that voluntary disclosure allows individuals to better differentiate their actions according to the probability of detection, potentially resulting in more taxes evaded by low risk-averse taxpayers. Analysing the 2009, 2011, and 2012 VDPs in the U.S., Johannesen, et al. (2019) find that VDPs are not necessarily conducive to disclosures. Their results suggest that most disclosures happened outside of VDPs by individuals who never admitted prior noncompliance.

TABLE 7. Testing for the impact of voluntary disclosure programmes on IFC deposits

Item	EOIR and AEOI (including FATCA) Announcement and Commencement	EOIR and AEOI Announcement and Commencement
	IFC from Non-IFC	IFC from Non-IFC
EOIR Signature	-0.043	-0.044
	0.044	0.044
AEOI Announcement	-0.510	-0.044
	0.064	0.064
AEOI Commencement		-0.230***
		0.062
AEOI (including FATCA) Commencement	-0.172***	
	0.061	
Voluntary Disclosure/ Amnesty	0.227***	0.219***
	0.064	0.064
R2	0.004	0.005
Number of observations	29,461	29,461
Jurisdiction-pair FEs	Yes	Yes
Year-Quarter FEs	Yes	Yes

NOTE: Regression of foreign-owned bank deposits in IFCs on EOI and VDP dummy variables. The dependent variable is the stock of deposits held by savers of non-IFC jurisdiction i in banks of IFC j at the end of quarter q. The unit of observation is the jurisdiction-pair (i, j) and the sample period goes from Q1 of 2006 to Q4 of 2018. Data are provided for nonbank counterparties only. Data are aggregated across currencies, sectors, reporting institutions, and instrument type.

*** represents a statistical significance level of 1 percent.

SOURCE: Authors' calculations based on LBS, BIS, and data on information exchange agreements provided by the Global Forum

4.3 Differing definitions of international financial centres

The regressions in this paper use a list of IFCs based on that outlined in IMF (2000) (see Box 3). However, there are many definitions of what constitutes an IFC, with differing lists having been developed by many different authors (see for example, Johannesen and Zucman (2014) or Gravelle (2015)). To ensure that the results in the regression analysis are not being driven by the selective use of different jurisdictions, this section examines the results with different IFCs omitted from the analysis.

Changing the IFC list also changes the sample of counterparty countries. Following the literature, the analysis in Section 3 focuses on deposits in IFCs held by non-IFC residents. This means that for each of the IFC jurisdictions in the sample, those countries that are not on a given IFC list are added to the list of potential counterparties.

Table 8 reproduces the tests carried out in *Table 4* but removes each IFC one by one from the analysis. This shows the impact that each IFC has had on the main result. The focus here is on the specification with only EOIR and AEOI commencement as the independent variables of interest. Both models are shown: with jurisdiction-pair fixed effects (left panel) and both jurisdiction-pair and year-quarter fixed effects (right panel).

Table 8 mirrors the results from the regression analysis above, where most results remain significant at the 1-percent level. The impact of the changes in the sample and the composition of the data used is clear.

For those IFCs that are BIS reporters in the analysis, the exclusion from the list of IFCs affects the results only marginally and the coefficient size of the highly significant AEOI commencement variable varies only slightly across the IFC jurisdictions. This result points to a rather homogeneous impact of AEOI commencement on cross-border deposits in IFCs.

TABLE 8. Robustness checks of IFC list

Jurisdiction excluded	Coefficient for EOIR signature	Coefficient for AEOI commencement
Bahrain	-0.039	-0.254***
Bahamas	-0.029	-0.277***
Bermuda	-0.054	-0.240***
Cayman Islands	-0.034	-0.197***
Netherlands Antilles/Curaçao	-0.029	-0.256***
Cyprus	-0.035	-0.240***
Guernsey	-0.024	-0.267***
Hong Kong, China	-0.052	-0.256***
Isle of Man	-0.058	-0.281***
Jersey	-0.055	-0.292***
Luxembourg	-0.043	-0.276***
Macau, China	-0.047	-0.243***
Malaysia	-0.043	-0.266***
Panama	-0.047	-0.267***
Singapore	-0.043	-0.262***
Switzerland	-0.049	-0.277***

NOTES: Coefficient on EOIR signature and AEOI commencement including jurisdiction-pair and year-quarter fixed effects. The dependent variable is the stock of deposits held by savers of jurisdiction i in banks of IFC j at the end of quarter q.

*** represents a statistical significance level of 1 percent.

SOURCE: Authors' calculations based on BIS LBS, and data on information exchange agreements provided by the Global Forum.

5. Conclusion and future research

This paper examines the overall impact of EOI on foreign-owned bank deposits in IFCs. The key contributions of the paper include a more detailed dataset on bank deposits than has been used elsewhere in the literature, a more accurate dataset of information agreements, and a more granular examination of key events in the EOI timeline. The results suggest that the expansion of EOI in many jurisdictions around the world is having a positive impact on tax compliance and is reducing offshore bank deposits that, at least to some extent, represent hidden wealth. These findings accord with a fast-growing body of literature in this area.

The BIS data show a strong decline in bank deposits in IFCs in a period of expanded tax transparency. The results point to a decline of over USD 400 billion in these deposits, a significant reduction in the scale of offshore banking in IFCs. Using a panel regression model following the approach of Johannesen and Zucman (2014), the results show that AEOI commencement is associated with a significant, 22-percent decrease in foreign-owned IFC deposits. The results on EOIR, based on a shorter sample, suggest that the impact of EOIR has changed over time. Initial EOIR agreements signed in the aftermath of the commencement of peer review in 2009 had a strong impact; however, the impact of each additional agreement has been more muted, potentially due to the increasingly multilateral nature of the EOIR network.

There are important future areas of research to better understand the impact of EOI and hidden wealth. For instance, the impact of EOI on other asset classes (e.g., portfolio holdings) is not considered in this paper. The use of other assets not covered under EOI agreements to hide wealth (such as art or real property), is also an important area of study for detailed analysis (see e.g., De Simone, Lester, and Markle (2019)). Moreover, a departure from the predominantly macroeconomic, cross-country perspective of analysis can provide important insights into country-specific dynamics of tax and hidden wealth (Cassetta *et al.* (2014)).

References

Ahrends, L. and F. Bothner. (2019). The Big Bang: Tax Evasion After Automatic Exchange of Information Under FATCA and CRS. *New Political Economy* 24: 605–622.

Alstadsaeter, A., N. Johannesen, and G. Zucman. (2018). Who owns the wealth in tax havens? Macro evidence and implications for global inequality. *Journal of Public Economics* 162: 89–100.

Bank for International Settlements (BIS). (2013). *Reporting Requirements for the Locational Banking Statistics.* Basel, Switzerland.

Bank for International Settlements (BIS). (2017). *BIS Location Banking Statistics: Explanation of the Data Structure Definitions.* Version 201712. Basel, Switzerland. Retrieved from: https://www.bis.org/statistics/dsd_lbs.pdf.

Beer, S., M. Coelho, and S. Léduc. (2019). *Hidden Treasures: The Impact of Automatic Exchange of Information on Cross-Border Tax Evasion.* International Monetary Fund (IMF) Working Paper.

Bertrand, M., E. Duflo, and S. Mullainathan. (2004). How Much Should We Trust Differences-In-Differences Estimates? *Quarterly Journal of Economics* 119(1): 249–275.

Bilicka, K. and C. Fues. (2014). With Which Countries Do Tax Havens Share Information? *International Tax and Public Finance* 21(2): pp. 175–197.

Bouvatier, V., G. Capelle-Blancard, and A. L. Delatte. (2018). *Banks Defy Gravity in Tax Havens.* CEPR Discussion Paper 12222.

Casi, E., C. Spengel, and B. Stage. (2018). *Cross-Border Tax Evasion After the Common Reporting Standard: Game Over?* ZEW Discussion Paper 8–036.

Cassetta, A., C. Pauselli, L. Rizzica, and M. Tonello. (2014). Financial Flows To Tax Havens: Determinants and Anomalies. *Quaderni dell'antiriciclaggio* 1.

Collin, M., S. Cook, S. and K. Soramaki. (2016). The Impact of Anti-Money Laundering Regulation on Payment Flows: Evidence From SWIFT Data. *SSRN Electronic Journal.*

De Simone, L., R. Lester, and K. Markle. (2019). *Transparency and Tax Evasion: Evidence From the Foreign Account Tax Compliance Act (FATCA).* Stanford Graduate School of Business Working Paper No 3744.

Fichtner, J. (2016). The Anatomy of the Cayman Islands Offshore Financial Center: Anglo-America, Japan, and the Role of Hedge Funds. *Review of International Political Economy* 23(6): 1034–1063.

Global Forum on Transparency and Exchange of Information for Tax Purposes. (2018). *Automatic Exchange of Information Implementation Report.* OECD Publishing.

Gravelle, J. G. (2015). *Tax Havens: International Tax Avoidance and Evasion.* Washington, DC: Congressional Research Service 7–5700 (R40623)

Hanlon, M., E.L. Maydew, and J. R. Thornock. (2015). Taking the Long Way Home: U.S. Tax Evasion and Offshore Investments in U.S. Equity and Debt Markets. *Journal of Finance* 70(1): 257–287.

Heckemeyer, J. and A.K. Hemmerich. (2018). Information Exchange and Tax Haven Investment in OECD Securities Markets. *SSRN Electronic Journal.*

Huizinga, H. and G. Nicodème. (2004). Are International Deposits Tax-Driven? *Journal of Public Economics* 88(6): 1093–1118.

International Monetary Fund (IMF). (2000). *Offshore Financial Centers.* IMF Background Paper.

Johannesen, N. (2014). Tax Evasion and Swiss Bank Deposits. *Journal of Public Economics* 111: 46–62.

Johannesen, N., P. Langetieg, D. Reck, M. Risch, and J. Slemrod. (2018). *Taxing Hidden Wealth: The Consequences of U.S. Enforcement Initiatives on Evasive Foreign Accounts.* Cambridge, MA: National Bureau of Economic Research (NBER) Working Paper No. 24366.

Johannesen, N. P. Langetieg, D. Reck, M. Risch, and J. Slemrod. (2019). *Taxing Hidden Wealth: The Consequences of U.S. Enforcement Initiatives on Evasive Foreign Accounts.* Washington, DC: Internal Revenue Service, Statistics of Income Division Working Paper.

Johannesen, N. and G. Zucman. (2014). The End of Bank Secrecy? An Evaluation of the G20 Tax Haven Crackdown. *American Economic Journal: Economic Policy* 6(1B): 65–91.

Kemme, D. M., B. Parikh, and T. Steigner. (2017). Tax Havens, Tax Evasion and Tax Information Exchange Agreements in the OECD. *European Financial Management* 23(3): 519–542.

Langenmayr, D. (2017). Voluntary Disclosure of Evaded Taxes—Increasing Revenue, or Increasing Incentives to Evade? *Journal of Public Economics* 151: 110–125.

Menkhoff, L. and J. Miethe. (2019). Tax Evasion in New Disguise? Examining Tax Haven's International Bank Deposits. *Journal of Public Economics* 176: 53–78.

OECD. (2000). *Towards Global Tax Co-operation: Progress in Identifying and Eliminating Harmful Tax Practices.* Paris, France: Organisation for Economic Co-operation and Development (OECD). Report to the 2000 Ministerial Council Meeting and Recommendations by the Committee on Fiscal Affairs.

OECD. (2015). *Update on Voluntary Disclosure Programmes: A Pathway to Tax Compliance.* Paris, France: Organisation for Economic Co-operation and Development (OECD) Publishing.

OECD. (2018). *Taxation of Household Savings.* Paris, France: Organisation for Economic Co-operation and Development (OECD) Publishing, Tax Policy Studies.

OECD, (2019). OECD Secretary-General Report to G20 Leaders. Paris, France: Organisation for Economic Co-operation and Development (OECD).

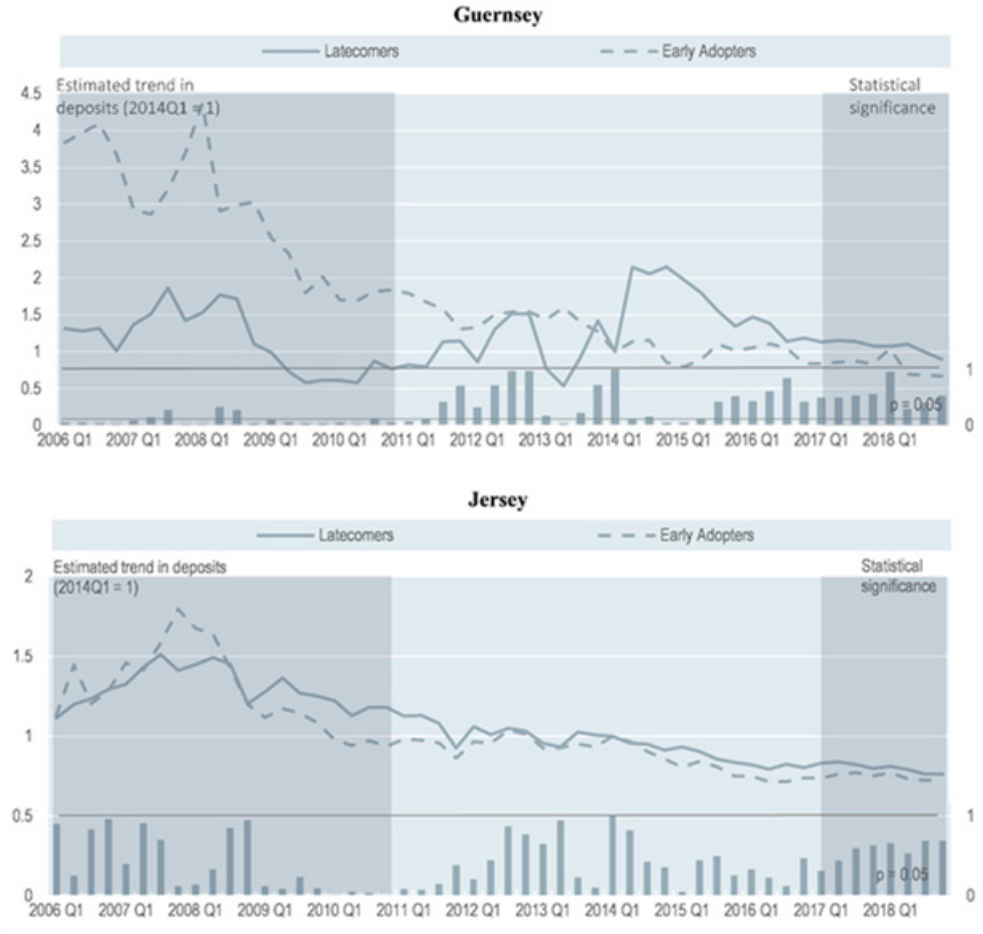
Omartian, J. (2016). Tax Information Exchange and Offshore Entities: Evidence From the Panama Papers. *SSRN Electronic Journal.*

Pellegrini, V., A. Sanelli, and E. Tosti. (2016). *What do External Statistics tell us About Undeclared Assets held Abroad and Tax Evasion?* Bank of Italy Occasional Paper No. 367.

Zucman, G. (2013). The Missing Wealth of Nations: Are Europe and the U.S. Net Debtors or Net Creditors? *Quarterly Journal of Economics* 128(3): 1321–1364.

Appendix

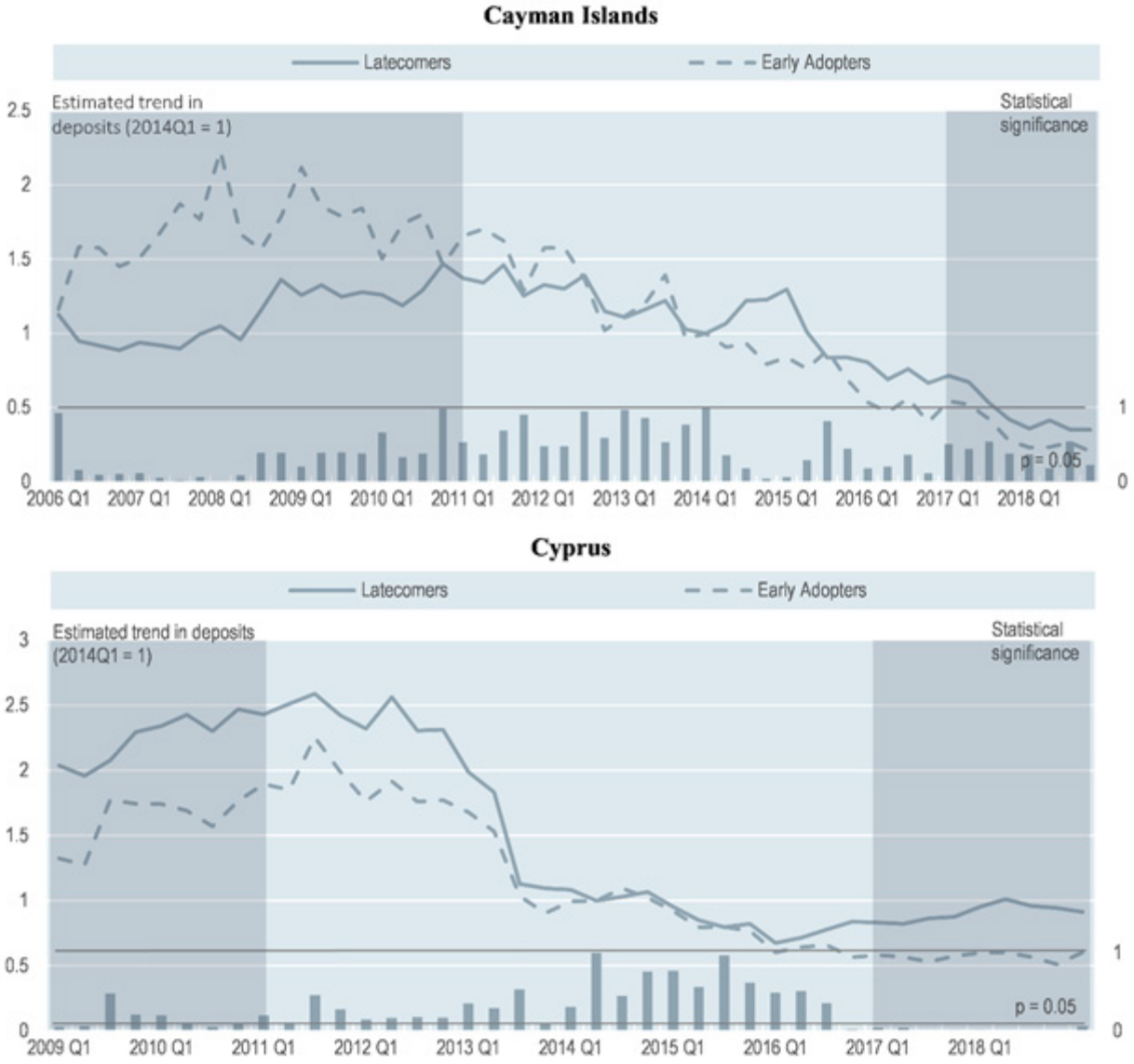
FIGURE A1. Difference-in-differences analysis of AEOI commitment by Guernsey and Jersey



NOTE: Deposit trends based on difference-in-differences estimation Lines indicate trends in deposits as captured by coefficients on time dummies θ_t and the interaction terms $\theta_t * EA_{ij}$, that is $\exp(\gamma_t)$ for non-Early Adopters and $\exp(\gamma_t + \delta_t)$ for Early Adopters. Columns indicate statistical significance levels of interaction terms $\theta_\tau * EA_{ij}$. Areas shaded on both ends of the sample range direct the reader to a time window of analysis relevant for inspection due to being less likely influenced by other events than the joint announcement.

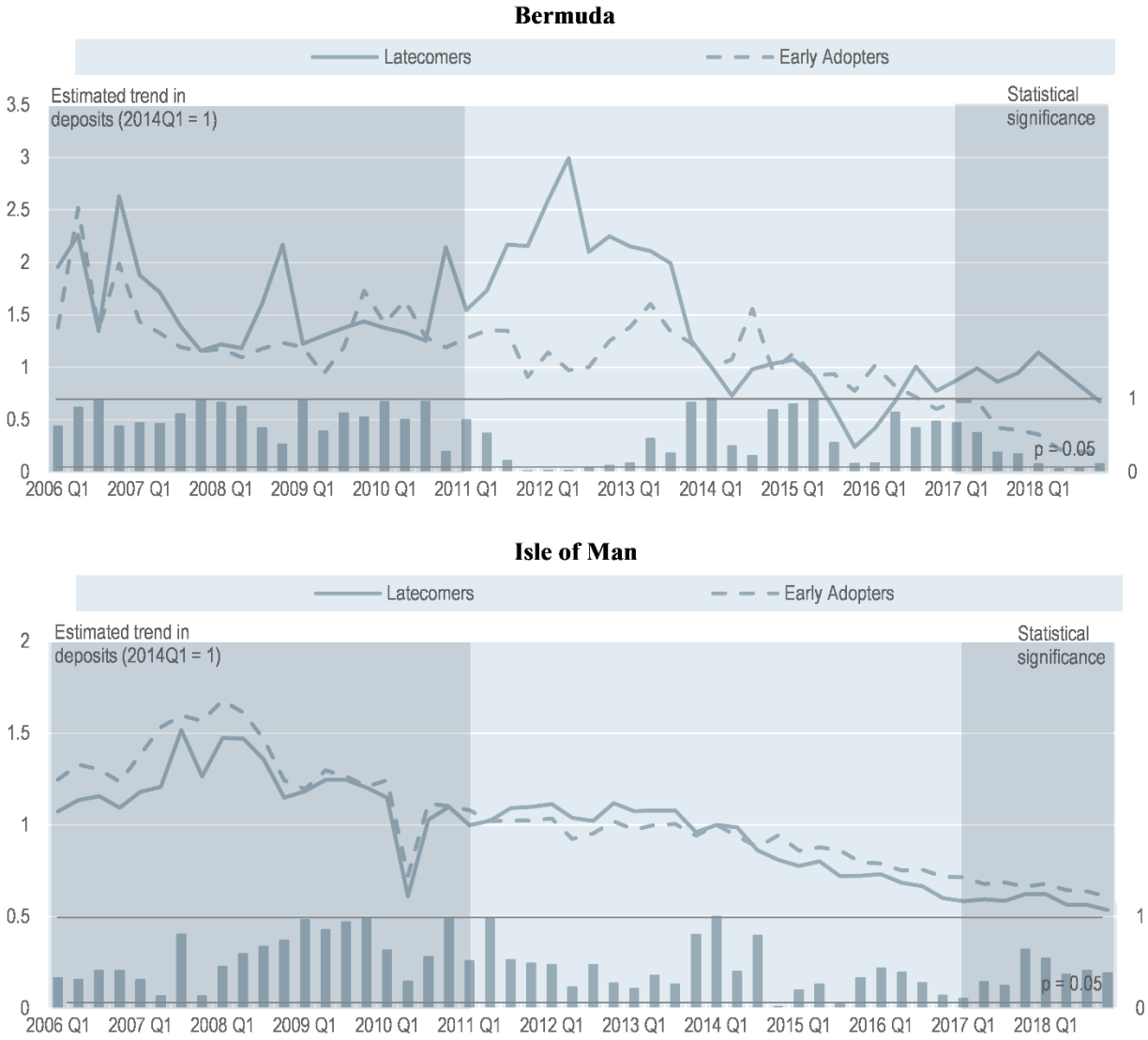
SOURCE: Authors' calculations based on BIS LBS data.

FIGURE A2. Difference-in-differences analysis of AEOI commitment by Cayman Islands and Cyprus



NOTE: Deposit trends based on difference-in-differences estimation Lines indicate trends in deposits as captured by coefficients on time dummies θ_t and the interaction terms $\theta_t * EA_{ij}$, that is $\exp(\gamma_t)$ for non-Early Adopters and $\exp(\gamma_t + \delta_t)$ for Early Adopters. Columns indicate statistical significance levels of interaction terms $\theta_t * EA_{ij}$. Areas shaded on both ends of the sample range direct the reader to a time window of analysis relevant for inspection due to being less likely influenced by other events than the joint announcement.
SOURCE: Authors' calculations based on BIS LBS data.

FIGURE A3. Difference-in-differences analysis of AEOI commitment by Bermuda and Isle of Man



NOTE: Deposit trends based on difference-in-differences estimation Lines indicate trends in deposits as captured by coefficients on time dummies θ_t and the interaction terms $\theta_t * EA_{ij}$, that is $\exp(\gamma_t)$ for non-Early Adopters and $\exp(\gamma_t + \delta_t)$ for Early Adopters. Columns indicate statistical significance levels of interaction terms $\theta_t * EA_{ij}$. Areas shaded on both ends of the sample range direct the reader to a time window of analysis relevant for inspection due to being less likely influenced by other events than the joint announcement.
SOURCE: Authors' calculations based on BIS LBS data.

2

The Influence of External Factors on Compliance

Lin

Holtzblatt ◆ McGuire

Recent Changes in the Paid Return Preparer Industry and EITC Compliance

Emily Y. Lin (Office of Tax Analysis, U.S. Department of the Treasury)[1]

1. Introduction

The U.S. tax system operates on voluntary compliance, with enforcement function, such as audit, penalty, or criminal investigation, serving as a deterrent to noncompliance. The most recent estimate from the Internal Revenue Service (IRS), covering Tax Years 2008–2010, shows a level of voluntary compliance at 81.7 percent of the total Federal tax liability. The annual gross tax gap—the tax liability that is not paid voluntarily and timely—associated with this level of voluntary compliance is estimated to be $458 billion for 2008–2010.[2]

The standard economic model predicts that voluntary tax compliance decreases when the benefit for evasion is high while the cost for evasion is low (Allingham and Sandmo (1972)). The Earned Income Tax Credit (EITC) provides a large tax benefit with an extended list of eligibility requirements, many of which are not readily verifiable by the tax authority under self-certification of individual tax filing. The U.S. Department of the Treasury (2018) has identified inability by the IRS to authenticate eligibility before tax refunds are paid as the main root cause for EITC improper payments. The EITC improper payment rate, measured as the ratio of EITC overpayments (net of revenue protected through IRS pre-refund enforcement activity) to total EITC claims, is estimated to be 25 percent, amounting to $18.4 billion in Fiscal Year 2018 (U.S. Department of the Treasury (2018)).

With this high error rate, EITC returns had an audit coverage rate of 1.2 percent in Fiscal Year 2017, about twice the audit coverage rate for all individual income tax returns.[3] When EITC errors are detected, not only must the taxpayer pay back the amount claimed in error plus interest, but she may also be subject to penalties, a re-certification requirement upon the subsequent EITC claim, or even a ban from claiming the EITC for the next 2 or 10 years. In addition to audits, the IRS also conducts computer matching of information provided on tax returns with third-party information reports to detect income misreporting and potential tax adjustments. Moreover, in limited circumstances, the IRS can make an immediate tax assessment to correct a mathematical or clerical error, e.g., a missing or an incorrect taxpayer identification number, on an EITC claim.

Since over 50 percent of EITC returns use a paid preparer,[4] one intriguing issue surrounding EITC compliance is the role played by paid tax return preparers in affecting the compliance outcome of EITC claims. Paid preparers can assist taxpayers in understanding and complying with tax law, but incompetent or unethical preparers can add to tax noncompliance, consequently contributing to the tax gap and undermining the tax system. Two limited sample studies (GAO (2006); Treasury Inspector General for Tax Administration (2008)) reveal that returns prepared by unlicensed and unenrolled preparers contained significant errors. Moreover, Jones (2017) finds evidence that EITC compliance is compromised when a preparer has a strong incentive to sell refund anticipation products to taxpayers.

Between 2009 and 2012, paid tax return preparers saw a series of changes in legislation, regulation, and tax administration that affected their profession. The goals of these changes were to strengthen preparer compliance and competence and thereby reduce tax return errors, often with a focus on addressing EITC overclaims. In 2009, Congress enacted an electronic-filing mandate on individual income tax return preparers, and the

[1] I am grateful to Adam Cole, Kara Leibel, Alan Plumley, and participants at the 2019 IRS-TPC Research Conference on Tax Administration for their helpful comments and discussions. The views expressed in this article are those of the author and not necessarily those of the U.S. Department of the Treasury.

[2] Internal Revenue Service (2016).

[3] Author calculation of the statistics provided in the 2017 Internal Revenue Service Data Book (Internal Revenue Service (2018)).

[4] Author calculation of tax data for Tax Year 2016.

IRS announced plans to increase oversight of paid return preparers. Beginning in January 2011, tax preparers were required to register with the IRS and to furnish the IRS-issued preparer tax identification number (PTIN) on returns they helped prepare. Also in 2011, IRS issued a regulation to require standards for all paid preparers, including those without license or credential. In the same year, Congress raised the EITC due diligence penalty on paid preparers, and the IRS subsequently intensified the preparer due diligence requirements through regulation. In 2012, the IRS implemented an EITC Return Preparer Study to heighten preparer enforcement activities through phone calls, letters, and preparer audits. Finally, the IRS took additional administrative steps beginning in Tax Year 2012 to further enhance preparer EITC due diligence requirements.

In this paper, I investigate the extent to which these legislative, regulatory and tax administrative efforts have improved EITC compliance through a reduction in EITC errors on paid-preparer returns. To monitor individual income tax compliance, the IRS conducts audits on a random sample of individual income tax returns each year as part of its National Research Program (NRP). The NRP 1040 study provides the underlying data for the estimates of EITC improper payments. Due to the sample size, several years of NRP 1040 studies are required to produce reliable estimates of different types of EITC errors. The latest NRP 1040 study available for analysis is from Tax Year 2013. When more years of post-reform data become available, NRP 1040 studies will be the ideal data source for estimating the compliance effect of the preparer industry reform.

In the absence of recent data on random audit results, I use information reported on a tax return as well as the scoring result from an IRS compliance system to construct four indicators that would predict the occurrence of EITC claim errors. Although showing these indicators on a tax return is not necessarily noncompliant, such returns are more likely to contain EITC claim errors than others. The indicators are: (1) claiming the head-of-household filing status; (2) claiming an EITC qualifying child to whom the taxpayer is not the parent; (3) being identified as violating at least one rule in the IRS compliance scoring system for child-related tax benefits; and (4) reporting income around the first "kink" of the EITC schedule.[5] These indicators of potential EITC errors form the measures of compliance outcomes analyzed in the paper.

Using the difference-in-difference estimation, I find evidence of improved EITC compliance as a result of the 2009–2012 paid preparer industry reform. Specifically, the reform has lowered the frequency that the EITC error indicators appear on paid-preparer EITC returns. Under plausible assumptions, the estimates suggest modest and statistically significant reductions in the share of EITC returns containing filing status, qualifying child, or self-employment income errors, with the effects ranging from 0.34 to 0.35 percentage points for the filing status error, 0.68 to 0.95 percentage points for qualifying child errors, and 0.33 to 0.70 percentage points for misreporting of self-employment income.

The paper is organized as follows. Section 2 provides background on EITC eligibility rules and the paid preparer industry changes as well as a review of the literature about sources of EITC errors. Section 3 describes the administrative tax data used in the analysis and provides summary statistics. It also discusses the estimation strategies. Section 4 presents estimation results. Section 5 concludes and suggests future directions of research.

2. Background and Previous Studies

2.1 EITC Eligibility Rules

Taxpayers must meet a host of rules to qualify for the EITC. First, they must have earned income (including wage and self-employment income) and the earned income, along with adjusted gross income (AGI) and investment income, must be below certain limits. The fully refundable credit initially phases in with earned income, reaches the maximum level over a range of earned income, and then phases out as income (measured as the larger of earned income or AGI) further increases. The phase-in and phase-out rates for the credit vary with the number of qualifying children claimed, resulting in different income limits for families of different sizes shown in Table 1.[6] Moreover, married taxpayers filing jointly benefit from a marriage penalty relief, under

[5] The amount of Earned Income Tax Credit increases as a linear function of income until reaching the maximum credit amount; the income at which the credit reaches the maximum credit amount is the first "kink" in the EITC schedule. It "kinks" again at a higher income level to begin the phaseout of the credit at higher incomes.

[6] The third-child EITC, which provides a larger credit for families with three or more children, was enacted in the 2009 American Recovery and Reinvestment Act (ARRA). The ARRA also expanded the EITC marriage penalty relief enacted in 2001.

which the credit starts to phase out at a higher income level for joint-filing taxpayers than for single or head-of-household taxpayers.

These rules lead the amount of EITC to vary with income, filing status, and the number of qualifying children claimed. As illustrated in Table 1, the maximum credits and income limits are much higher for taxpayers claiming qualifying children than those without children, and the more qualifying children are claimed, the higher the credit amount and income limit will be. In 2018, the maximum credit was $6,431 for families with three or more children, $5,716 for those with two children, $3,461 for those with one child, and $519 for those without children.

Individuals who are married at the end of the year must file a joint return to qualify for the credit. Married individuals who file as married-filing-separately are not eligible. Even when living apart from their spouses, married individuals are treated as married for filing status purpose unless they are legally separated under a decree of divorce or separate maintenance. Exceptions however are permitted for married individuals who live apart from their spouses for the last 6 months of the year, maintain a home with a child for more than one half of the year, and provide over one half of the cost of maintaining the household. In this case, the individual may use the head-of-household status and claim the EITC accordingly.

TABLE 1. EITC Parameters, 2018

Item	Number of Qualifying Children			
	Zero	One	Two	Three or More
Phase-in rate (%)	7.65	34	40	45
Minimum income for maximum credit ($), first EITC kink point	6,780	10,180	14,290	14,290
Maximum credit ($)	519	3,461	5,716	6,431
Phase-out rate (%)	7.65	15.98	21.06	21.06
Income at which phase-out begins ($)	8,490	18,660	18,660	18,660
	(14,170 if joint)	(24,350 if joint)	(24,350 if joint)	(24,350 if joint)
Income at which phase-out ends ($)	15,270	40,320	45,802	49,194
	(20,950 if joint)	(46,010 if joint)	(51,492 if joint)	(54,884 if joint)

SOURCE: IRS Revenue Procedure 2018–18.

If any child is claimed for the credit, each child must pass a number of qualifying child tests—age, relationship, residency, joint return—in order to qualify. For the age test, the child must be younger than age 19, or 24 if a full-time student, at the end of the year. The child, however, can be any age if disabled. Beginning in 2009, qualifying children who are not disabled must be younger than the taxpayer. For the relationship test, the child must be the son or daughter, an adopted child, a stepchild, a foster child, a sibling, or a descendent of any of them. For example, a grandchild or a niece would meet the relationship test. For the residency test, the child must reside with the taxpayer in the United States for more than one-half of the year. Last, for the joint return test, the qualifying child cannot file a joint return except for claiming a refund.

It is possible for a child to meet the qualifying child tests for more than one taxpayer (e.g., a child in a three-generation household), but only one taxpayer can claim the child for the benefits. There are particular rules, the so-called tiebreakers, taxpayers need to follow in order to determine the right person to claim the child for the EITC. Under current law, parents have priority to claim a qualifying child over nonparents.[7] If no parent can claim the child, and more than one nonparent taxpayer can claim the child, the child is treated as the qualifying child of the person with the highest AGI. If parents can claim the child but do not do so, an eligible nonparent can claim the child only if her AGI is higher than the highest AGI of any parent of the child.

[7] If parents are on separate returns and both claim a child who qualifies, the child is treated as the qualifying child of the parent with whom the child lives for a longer period. If the child lives with the parents for the same amount of time, the child is treated as the qualifying child of the parent with a higher AGI.

Both the taxpayer and qualifying children must live in the United States for more than half of the year, and each must have a Social Security number (SSN) that is valid for employment and issued by the return due date. In addition, a taxpayer who may be claimed as a qualifying child for the EITC by another taxpayer may not claim the EITC. Finally, to claim the credit without a qualifying child, a taxpayer must be 25 to 64 years old as of the end of the year and cannot be a dependent of another taxpayer.

2.2 Sources of EITC Errors

Results from random audits of individual income tax returns have been used to investigate sources of EITC errors. Early studies based on tax data in the 1990s (McCubbin (2000); IRS (2002)) identify violation of the qualifying child rules, stemming mostly from failure of the residency test, as the primary source of EITC errors. Since then, qualifying child errors have remained the principal factors for EITC noncompliance. Studying EITC errors detected in random audits for Tax Years 2006–2008, IRS (2014) reveals that qualifying child errors accounted for 52 percent of the total EITC overclaim dollars, followed by misreporting of self-employment income (23 percent) and filing status errors (16 percent). These errors are associated with the types of EITC eligibility criteria about which the IRS has little third-party information, such as taxpayers' living arrangements, relationship with the child, self-employment income, and marital status. Leibel (2014) estimates that about 75 percent and 20 percent of all children claimed in error failed, respectively, the residency and relationship tests.

Leibel, *et al.* (2017) explore the social welfare implication of EITC qualifying child errors, using random audit data from Tax Years 2006–2011. They study the intensity of the familial relationship between the child and the wrong taxpayer whose EITC claim was denied by the audit, and then examine why the right taxpayer did not claim the child.[8] The authors find that children claimed with qualifying child errors often have a less intense, but eligible, familial relationship with their claimants than children who meet all of the eligibility rules. Specifically, less than 50 percent of the children claimed with qualifying child errors were the son or daughter of the taxpayers, compared to over 90 percent of the children who met all of the eligibility rules.

Also using the IRS's random audit data, Chetty, *et al.* (2012) conclude that income misreporting is more prevalent among EITC claimants who report self-employment income than those reporting only wage income. The authors find sharp bunching in the reported income at the first kink of the EITC schedule among EITC claimants with self-employment income, an income point that maximizes tax refunds. They also find a substantial reduction in such bunching in these persons' true income as determined by audits. For wage earners, the distributions of the reported and detected income are similar.

Unlike wage income, which employers are required to report to the IRS annually on an employee's Form W-2, self-employment income is subject to little third-party information reporting, making it susceptible to reporting manipulation at filing. IRS estimates that self-employment income has a very high misreporting ratio, which consequently contributes to a significant share of the gross tax gap (IRS (2016)). It is worth noting that, despite the high error rate of EITC claims, EITC noncompliance makes up a comparatively small share of the total tax gap. Misreporting of all individual income tax credits, EITC included, accounted for 9 percent of the gross tax gap in Tax Years 2008–2010, whereas underreporting of business income accounted for 27 percent (IRS (2016)).

The benefit structure of the EITC leads to marriage penalties or bonuses for low-income couples, depending on the couple's income (Acs and Maag (2005); Lin and Tong (2012); Lin and Tong (2014); Maag and Acs (2015)). Marriage penalties (or bonuses) are the additional (or lesser) tax liability faced by a jointly filing couple when comparing their tax liability to the tax they would owe if they were to claim the single or head-of-household status on two separate returns. For some married couples, by incorrectly filing as single or head-of-household, the two spouses together can overclaim the EITC based on individual, instead of family, income. Furthermore, as previously described, only in very limited circumstances can married individuals who live apart from their spouses file as unmarried, and these rules may be confusing to some taxpayers. Intentional or not, filing status errors come mainly from married individuals incorrectly filing as unmarried, mostly as head-of-household, and overclaiming the EITC as a result (Leibel (2014)).

[8] The sample consists of children who were claimed with qualifying child errors, as determined by audits, but were potentially eligible for being claimed by another taxpayer because the children met the SSN, age, and joint filing eligibility rules.

Using random audit data from Tax Years 2006–2008, IRS (2014) finds little difference in the dollar overclaim rate between self-prepared and paid-preparer EITC returns but a wide range of dollar overclaim rates across returns prepared by different types of paid preparers. In particular, self-prepared and paid-preparer EITC returns both had a dollar overclaim rate ranging from 28 to 39 percent. In contrast, EITC returns prepared by unenrolled tax return preparers, constituting 43 percent of paid-preparer EITC returns, had the highest error rate; about 33 to 40 percent of EITC dollars on these returns were claimed in error. EITC returns prepared by national tax return preparation firms, another type of paid preparers frequently used by EITC claimants, had a dollar overclaim rate of 20 to 30 percent.

2.3 Paid Return Preparers

Researchers have long studied the effect of tax return preparers on individual compliance with the tax law. Evidence exists that return preparers play dual roles as enforcers of compliance and enablers of noncompliance (Klepper and Nagin (1989); Klepper, *et al.* (1991)). These two conflicting roles of return preparers arise because, while helping taxpayers follow their obligations under the tax law, preparers also attempt to assist taxpayers in minimizing tax liability. Specific to the EITC, Book (2007) outlines scenarios in which claim errors may occur when a taxpayer uses an uncredentialed preparer. The author concludes that an erroneous claim can contain inadvertent or intentional errors resulting from neglect or noncompliance of the taxpayer, the preparer, or both.

To enhance the positive role of paid preparers, a number of key changes took place in the legislative, regulatory, and tax administrative aspects affecting the profession in the late 2000s and the early 2010s. From strengthening the EITC preparer due diligence to applying standards to all paid preparers, these changes were expected to improve preparer compliance and competence, consequently increasing the accuracy of tax returns they helped prepare.

Paid tax return preparers are subject to penalties for violation of a number of preparer standards and obligations specified in the Internal Revenue Code (IRC).[9] The EITC due diligence penalty was codified in 1997. The penalty is assessed on paid preparers for each tax return with which the preparers fail to exercise due diligence in determining taxpayer eligibility for, and the amount of, the EITC.[10] Under this authority, preparers are required by regulation to complete Form 8867, *Paid Preparer's Due Diligence Checklist*, to assist them in determining and documenting taxpayer eligibility.[11] Before Tax Year 2011, preparers were not required to attach Form 8867 to the return; they instead kept Form 8867 in their files, along with the EITC worksheets and a record documenting information used to complete the form and worksheets, for 3 years.

To make EITC preparers more aware of their responsibility, a legislative change in 2011 increased the preparer due diligence penalty from $100 (unindexed) to $500 (indexed). The subsequent regulations prescribed additional due diligence requirements on paid preparers, including submitting to the IRS the completed Form 8867 along with the tax return. After a series of educational outreaches to preparers about the changes in the due diligence requirements, the IRS started in Tax Year 2012 to summarily impose the penalty on preparers with a missing Form 8867 for an EITC claim.

Although the questions and wording of Form 8867 have been revised several times since its inception in 1998, until 2016 it always contained a checklist of EITC eligibility rules for preparers to complete, effectively making preparers as a gatekeeper to EITC payments. Noticeable changes were made to Form 8867 in Tax Year 2006 to highlight preparers' due diligence responsibilities. Specifically, a new Part IV about the due diligence checklist was added, turning the description of a list of due diligence requirements into a set of questions for preparers to answer. For example, preparers were asked to answer—yes or no—if they completed Form 8867 based on information provided by the taxpayer or obtained by themselves.

[9] A summary of preparer penalties is available on https://www.irs.gov/tax-professionals/summary-of-preparer-penalties-under-title-26.

[10] The due diligence requirements were extended to returns claiming the child tax credit and the American Opportunity Tax Credit beginning in Tax Year 2016 and to returns claiming the head-of-household filing status beginning in Tax Year 2018.

[11] The form was vastly redesigned, and renamed to *Paid Preparer's Due Diligence Checklist,* beginning in Tax Year 2016, when the requirements were extended to other child-related tax benefits.

Major revisions to Form 8867 were made in Tax Year 2012 to address the less compliant EITC rules and key error sources, including tiebreakers, qualifying child tests, business income, etc. Additional due diligence questions were asked in Part IV of the form and a new Part V was added, in which preparers were asked to check from a list of taxpayer documents used by the preparer to substantiate the qualifying child's residency arrangement (e.g., school records), disability status of the child (e.g., doctor statements), and the taxpayer's reported business income (e.g., bank statements). However, in Tax Year 2016, the IRS significantly redesigned the form, making it substantially shorter, when the due diligence requirements were extended to other child-related tax credits.

Another legislative change, which occurred in 2009, involved mandating electronic-filing of individual income tax returns prepared by paid preparers, except for preparers expecting to file 10 or fewer returns during the calendar year. IRS phased in this mandate, setting the e-file requirement at 100 or more returns for Calendar Year 2011 (generally for Tax Year 2010 returns) and 11 or more for Calendar Year 2012 (generally for Tax Year 2011 returns). Langetieg, *et al*. (2013) estimate that, out of the approximately 81 million individual income tax returns prepared by paid preparers, about 6 million and another 7 million additional returns were e-filed, respectively, for Tax Years 2010 and 2011. Furthermore, the authors show the association between this higher e-filing rate and fewer math errors on tax returns.

In addition to the IRC, Title 31 of the U.S. Code gives the Secretary authority to regulate practice before the Treasury Department. Regulations have long been issued under this authority by the Treasury Department, referred to as Circular 230, to oversee the practice of licensed attorneys, certified public accountants, enrolled agents and actuaries, and enrolled retirement plan agents—individuals who may practice before the IRS. Practitioners are required to demonstrate good character and reputation as well as necessary qualification and competence as misconducts may lead to censure, suspension, or disbarment from practice. Dubin, *et al*. (1992) find that return audit rates and IRS penalties play a role in explaining taxpayer demand for paid return preparers who are practitioners. Of all types of paid preparers, left out from this regulation were those who were unlicensed and unenrolled—the group of preparers who regularly filed most EITC returns—due to their nonpractitioner status.[12]

In 2009, the IRS released the *Return Preparer Review* report (IRS (2009)) after conducting a comprehensive review of its paid preparer strategies. In the report, the IRS recommended an increased oversight of paid preparers and identified implementation plans to achieve the goal. First off, beginning in January 2011, all paid preparers, irrespective of practitioner status, were required to register with the IRS and receive a preparer tax identification number (PTIN) to furnish on tax returns they prepared as well as on Form 8867 for returns claiming the EITC. A subsequent regulation in 2011 created a new category of practitioners, registered tax return preparers (RTRPs), to formally incorporate unenrolled and unlicensed paid preparers into the regulatory system under Circular 230. RTRPs could practice before the IRS in limited circumstances and would be subject to minimum education and competency requirements. The education and competency requirements imposed on unlicensed or uncredentialed preparers, however, were later ruled by the U.S. Court of Appeals for the D.C. Circuit to be exceeding the IRS's authority to regulate practice before the IRS.[13]

Langetieg, *et al*. (2013) found that, coincident with the several developments in the paid preparer industry environment, the number of paid preparers dropped significantly beginning in Processing Year 2010 (generally for Tax Year 2009 returns) and the preparer attrition rate escalated in Processing Year 2011 (generally for Tax Year 2010 returns). The number of preparer-assisted returns, however, remained relatively stable in these years, indicating increased consolidation in the preparer industry. The share of tax returns prepared by large-volume preparers—those preparing 100 or more returns—rose as a result. A descriptive analysis in the same study further shows that the share of returns containing mismatches with third-party information was lower for returns prepared by PTIN holders than returns prepared by non-PTIN holders.

Administratively, in Fiscal Year 2012, the IRS began a multi-year initiative aimed at improving accuracy of paid-preparer EITC returns (U.S. Department of the Treasury (2016)). In this EITC Return Preparer Study,

[12] Some states impose their own requirements to regulate this type of preparers.

[13] Loving v. IRS, 742 F.3d 1013 (D.C. Cir. 2014).

IRS tested different outreach and enforcement strategies on paid preparers of EITC returns and then adopted those strategies identified as effective treatments in its routine operation. To maximize the effectiveness, these educational and enforcement techniques, including phone calls, letters, preparer audits, etc., were targeted to paid preparers who failed to meet the EITC due diligence requirements or who filed a large number of EITC returns with potential errors (U.S. Department of the Treasury (2016)).

In summary, between 2009 and 2012, the paid preparer environment was significantly changed by a number of developments concerning the profession. In 2009, the e-file mandate was enacted, and the IRS released the *Return Preparer Review* in which it announced new requirements for unenrolled and unlicensed paid preparers. In 2011, when preparers filed returns for Tax Year 2010, the IRS implemented the first phase of the e-file mandate and began to require PTINs. Also in 2011, the IRS released the revised Circular 230 to require standards for all paid preparers, including those without license or credential. In the same year, Congress raised the EITC due diligence penalty and the IRS intensified the due diligence requirements in the proposed regulation, both becoming effective in Tax Year 2011. In 2012, when preparers filed returns for Tax Year 2011, the IRS implemented the EITC Return Preparer Study as well as the second phase of the e-file mandate. Additional IRS administrative changes—a revised Form 8867 and broad applications of due diligence penalties—became effective in Tax Year 2012.

3. Data and Estimation

3.1 Data

For the paper, I drew a 1-percent EITC sample from the population files of Federal individual income tax returns for Tax Years 2004 through 2016, covering the period of paid preparer industry reform during Tax Years 2009–2012.[14] Due to lack of random audit data for detected EITC noncompliance, I used a number of tax return indicators known to be linked with a high likelihood of EITC errors to predict noncompliance of an EITC claim. The four tax return indicators are: (1) claiming the head-of-household status; (2) claiming an EITC qualifying child to whom the taxpayer is not the parent; (3) being identified by the IRS as violating at least one rule specified in its Dependent Database Scoring System; and (4) reporting an amount of earned income around the first EITC kink point if the taxpayer is self-employed.

The first noncompliance predictor is the use of the head-of-household filing status. As mentioned earlier, a common reason for EITC overclaim is misreporting of filing status when married individuals file as unmarried, often using the head-of-household status. Table 2 shows that over one-half of EITC claimants used this filing status before 2009, with the ratio falling to 48 percent afterwards. The dip observed in 2009 was likely related to the expansion of the credit for married couples.

The second noncompliance predictor is claiming the EITC with qualifying child when the taxpayer is not the child's parent. This indicator is related to qualifying child errors because children claimed with errors are much less likely to be the son or daughter of the taxpayer than children meeting the eligibility tests. Specifically, a return takes the value of one (1) for this dummy indicator if the taxpayer claimed at least one child to whom she is not the parent. I use a database the IRS received from the Social Security Administration, known as Kidlink, to determine whether a taxpayer is the parent of the child she claimed for the EITC. Kidlink contains the Social Security number of a child and the Social Security number(s) of the parent(s) listed on the child's application for Social Security number. The information allows me to determine the parent-child relationship. As shown in Table 2, the proportion of EITC returns making a nonparent claim increased slightly over the period, rising from about 20 percent in the earlier years to as high as 24 percent in 2013 before tapering down to 22 percent in 2016.

[14] Taxpayers younger than 15 or older than 64 were dropped.

TABLE 2. Trends in Frequency of Potential EITC Error Indicators

Tax Year	Head-of-Household Filing Status		Nonparent Claim for EITC with Qualifying Child		DDB Rule Violation		Income at the First EITC Kink, if Self-Employed		Number of Observations
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	
2004	0.53	(0.50)	0.21	(0.41)	0.12	(0.32)	0.15	(0.35)	219,426
2005	0.53	(0.50)	0.19	(0.39)	0.12	(0.32)	0.15	(0.36)	222,242
2006	0.54	(0.50)	0.19	(0.39)	0.12	(0.32)	0.14	(0.35)	225,124
2007	0.52	(0.50)	0.21	(0.41)	0.12	(0.33)	0.15	(0.36)	236,238
2008	0.51	(0.50)	0.22	(0.41)	0.12	(0.33)	0.14	(0.35)	240,022
2009	0.48	(0.50)	0.22	(0.42)	0.17	(0.37)	0.14	(0.35)	263,980
2010	0.48	(0.50)	0.23	(0.42)	0.18	(0.38)	0.16	(0.37)	265,189
2011	0.48	(0.50)	0.23	(0.42)	0.20	(0.40)	0.16	(0.37)	268,292
2012	0.48	(0.50)	0.23	(0.42)	0.20	(0.40)	0.16	(0.37)	266,605
2013	0.48	(0.50)	0.24	(0.43)	0.21	(0.41)	0.18	(0.38)	271,879
2014	0.48	(0.50)	0.23	(0.42)	0.17	(0.38)	0.19	(0.39)	269,250
2015	0.49	(0.50)	0.23	(0.42)	0.22	(0.41)	0.19	(0.39)	266,230
2016	0.49	(0.50)	0.22	(0.42)	0.21	(0.41)	0.18	(0.38)	257,925

SOURCE: Individual Returns Transaction File and the IRS Dependent Database.

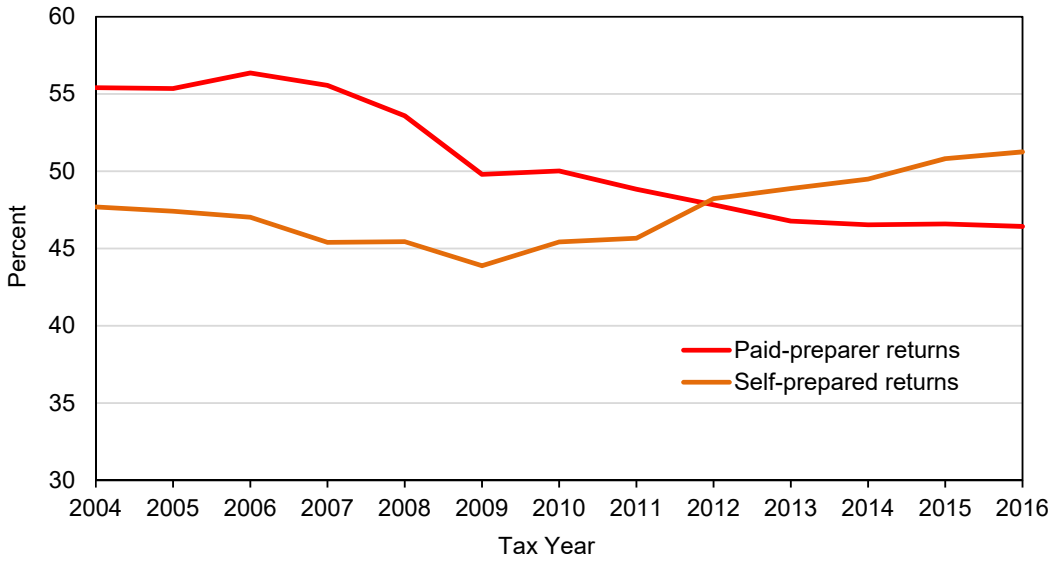
The third indicator is drawn from the IRS's Dependent Database Scoring System (DDB) for potential qualifying child errors. DDB is an IRS database that applies a large number of decision rules—rules that are connected to the eligibility criteria for child-related tax benefits—to tax returns claiming the EITC and other child-related tax benefits. It generates DDB scores for tax returns to help the IRS detect return errors and is one of the scoring systems used by the IRS for selecting returns for audit (GAO (2016)). I observe from the DDB scoring database if a return is identified as violating the rules specified in the system, and use this rule violation as an indicator of potential noncompliance with EITC eligibility criteria. Table 2 shows that the rule-breaking rate jumped from 12 percent in 2008 to 17 percent in 2009, followed by gradual increases to over 20 percent at the end of the period. Some of the large year-to-year swings likely stemmed from updates of the IRS scoring algorithm because the estimated EITC improper payment rates based on random audits did not show comparable variation.

The last noncompliance predictor is reporting income around the first kink of the EITC schedule if the taxpayer is self-employed. Studies (Saez (2010); Chetty, *et al.* (2012)) show that income bunching at the first EITC kink point is more prevalent among the self-employed than wage earners and, for the self-employed, some of this income bunching is driven by income misreporting. I identify self-employed EITC returns with earned income that falls in the $500 bin associated with the taxpayer's first EITC kink, according to the number of qualifying children claimed, or the next $500 bin above. As listed in Table 2, the bunching rate among self-employed EITC claimants increased moderately from 15 percent to 18 percent over the period.

Figures 1 through 4 provide visual diagnosis of the relationship between the 2009–2012 preparer industry changes and the potential compliance outcomes of paid-preparer EITC returns, using self-prepared returns as a comparison group.[15] In Figure 1, the share of returns claiming the head-of-household status dropped for paid-preparer returns in 2009 and continued to decline after 2010. This trend was in contrast to the steady climb in the share of self-prepared returns claiming the head-of-household status beginning in 2010. The head-of-household rate for self-prepared returns surpassed the head-of-household rate for preparer-assisted returns for the first time in Tax Year 2012, and the discrepancy widened afterwards.

[15] A return is self-prepared if it does not use a preparer regardless of whether it uses a software. About 70 percent of self-prepared EITC returns in the sample used a software in 2004, and the ratio steadily rose to 94 percent by 2013 and then plateaued. Returns that use the Volunteer Income Tax Assistance (VITA), Tax Counseling for the Elderly (TCE), or IRS assistance are excluded from the analysis.

FIGURE 1. Share Using the Head-of-Household Filing Status

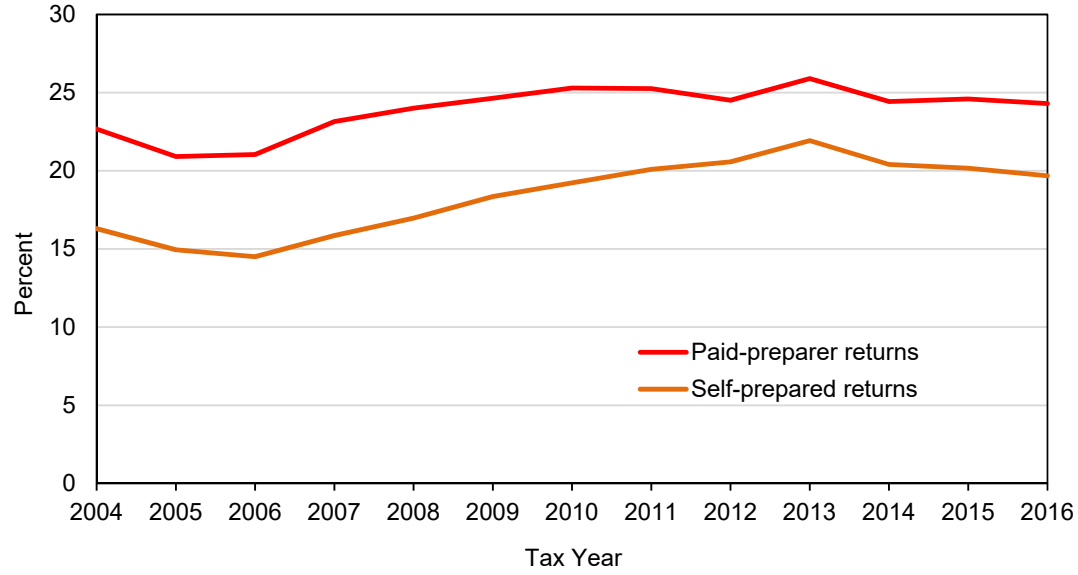


SOURCE: Individual Returns Transaction File.

Furthermore, as illustrated in Figures 2 and 3, the post-2008 period saw an increase in the frequency of both nonparent claims of qualifying children and DDB rule violation among EITC returns. However, self-prepared EITC returns had a slightly more rapid increase in these potential EITC noncompliance rates than did paid-preparer returns.

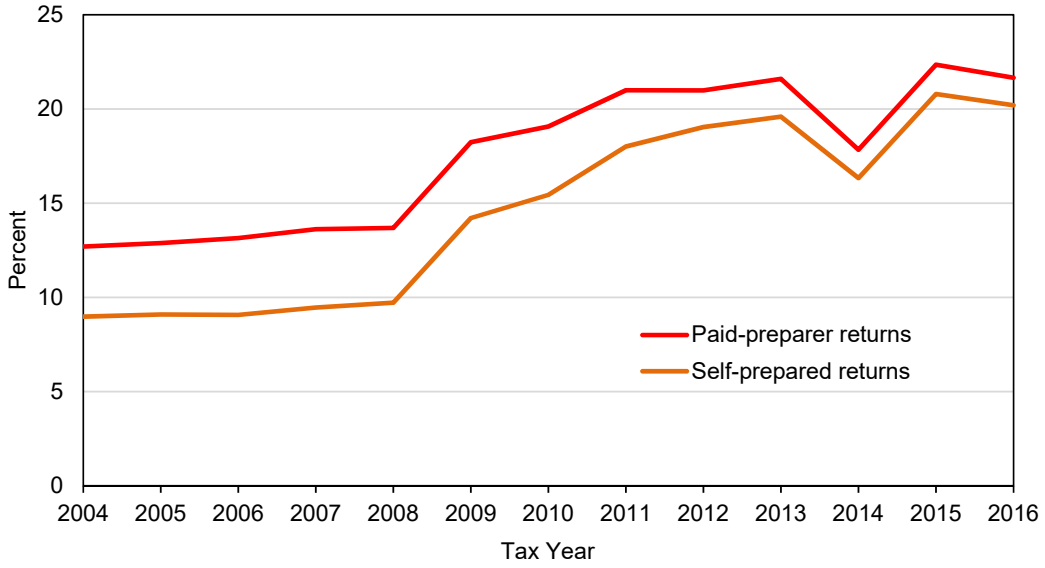
Finally, in Figure 4, among EITC returns reporting self-employment income, bunching of earned income around the first EITC kink point became noticeably more prevalent for self-prepared returns than for preparer-assisted returns after 2008. Continuing this trend, the disparity in the bunching rates between EITC returns prepared under the two methods reached the highest level in 2014.

FIGURE 2. Share Claiming Qualifying Children as a Nonparent



SOURCES: Individual Returns Transaction File and the IRS Dependent Database.

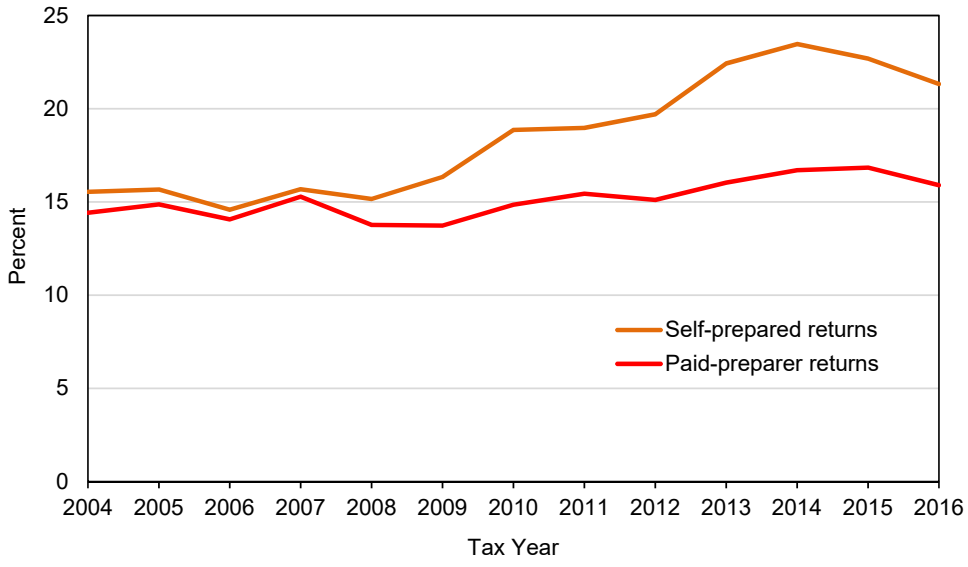
FIGURE 3. Share Violating DDB Rules



SOURCES: Individual Returns Transaction File and the IRS Dependent Database.

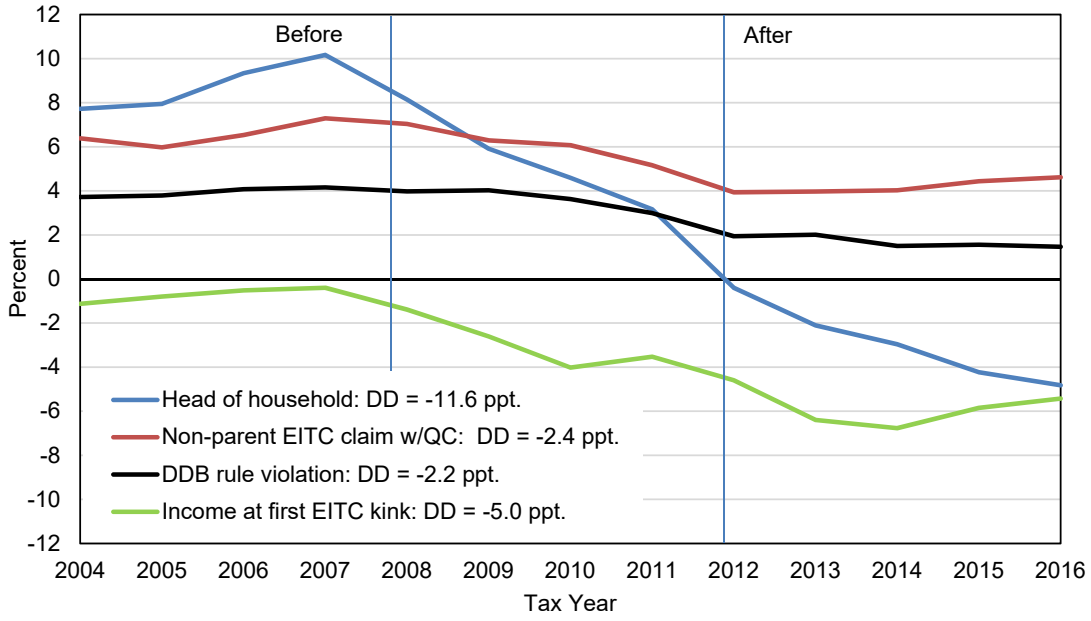
Figure 5 summarizes these relative trends. Breaking the 13-year period into pre-reform (2004–2008), reform (2009–2011), and post-reform (2012–2016) eras, Figure 5 depicts the relative potential EITC noncompliance rates, i.e., the frequency of potential error indicators for paid-preparer returns net of the frequency for self-prepared returns, by year. The figure also lists the before-and-after changes in the relative potential noncompliance rates, i.e., the difference-in-difference (DD) estimates. As shown, the relative rates were almost flat before reform and then declined substantially during the reform years, indicating a relatively more compliant outcome for paid-preparer returns as reform began. Except for the head-of-household indicator, for which the relative rate continued to decline, the relative rates for the other three potential EITC error indicators appeared to have reached a new steady state in the post-reform years. The DD estimates show the magnitudes of the average declines in these relative potential EITC noncompliance rates after the preparer industry reform.

FIGURE 4. Share Reporting Earned Income Around the First EITC Kink, if Self-Employed



SOURCE: Individual Returns Transaction File.

FIGURE 5. Relative Potential Error Rates: Paid-Preparer Returns v. Self-Prepared Returns



SOURCES: Individual Returns Transaction File and the IRS Dependent Database.

3.2 Estimation Strategy

The idea behind the difference-in-difference approach is that, in the absence of policy changes, the difference in potential error rates between the two preparation methods would remain constant, as illustrated in Figure 5 for the pre-reform years. The dips in the relative potential error rates seen after 2011, hence, are attributable to the effect of 2009–2012 industry changes on the treatment group, i.e., the paid-preparer returns, other things equal. The estimation describing this approach is

$$Y_i = \alpha + \beta\,(Paid_i * Post_i) + \gamma Paid_i + \delta Post_i + X_i\,\theta + \epsilon_i$$

In the equation, Y_i is the indicator of a potential error, taking the value of zero (0) or one (1), and X_i is a set of variables about taxpayer characteristics (age and gender of the primary taxpayer, and the State of residency) and the various tax schedules attached to the return. The latter is an indicator of return complexity and is potentially related to both return compliance and taxpayer choice of preparation method. The variable $Paid_i$ is a dummy for paid-preparer returns. The estimate γ thus measures paid-preparer returns' specific effect, i.e., the permanent difference between the two preparation methods. The variable $Post_i$ is a dummy for post-reform years, and the estimate δ thus measures the effect common to all returns in the post-reform years. I use data from Tax Years 2012–2016 for post-reform years and data from 2004–2008 for pre-reform years. Returns from Tax Years 2009–2011, i.e., the years with ongoing policy changes, are excluded from the regression analysis.

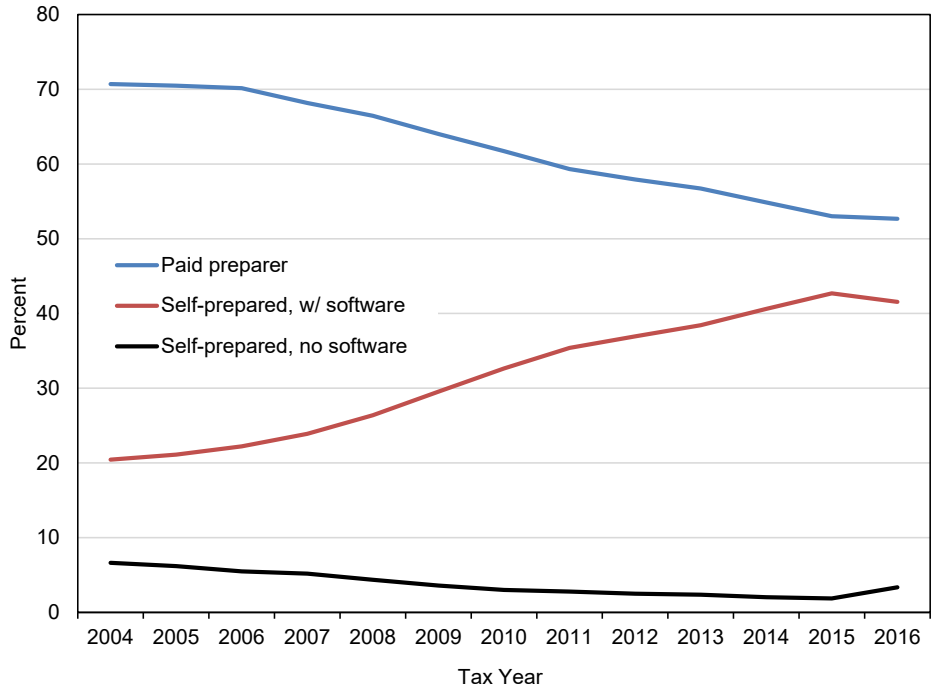
The estimate β would measure the true effect of the industry changes on reducing noncompliance if no omitted variables are correlated with both the noncompliance indicator (Y_i) and the post-reform return preparation method ($Paid_i * Post_i$). However, the relative composition between returns prepared by the two methods could well be shifted by reasons other than the policy's intended effect. For example, less compliant persons might have incentives to avoid paid preparers' due diligence scrutiny and thus switch to self-help. Likewise, compliant taxpayers might see submitting documents to preparers an unnecessary hassle and thus decide to prepare their own returns. The coefficient would be biased in either direction because post-reform preparation method is conditional on unobserved individual compliance.

In addition to taxpayer self-selection, omitted variable bias could be caused by preparer behavior. Some low-volume, unenrolled paid preparers might cease offering services as their business cost rises, consequently causing their clients to self-prepare or switch to another preparer. Furthermore, unethical preparers who

intend to escape from the new requirements might be paid to prepare a tax return but purposely do not sign it, creating the so-called "ghost preparer" phenomenon.[16] Because EITC claims filed by unenrolled or unethical preparers are more likely to contain errors, these preparer responses, if they happen, could alter the relative compliance between preparer-assisted and self-prepared returns, with the latter including returns filed by ghost preparers. Some of the errors, on the other hand, might be eliminated as taxpayers begin to use a more competent preparer. Either way, the relative compliance between returns prepared by the two methods is shifted due to reasons beyond the policy's intended effect.

To examine potential bias, Figure 6 depicts trends in paid-preparer and tax software use for EITC returns. There were substantial declines in the share of EITC returns using a paid preparer, decreasing from 70 percent in 2004 to just over 50 percent in 2016. Persistent decreases occurred throughout the years from 2007 to 2015. That is, declines in preparer use began prior to the 2009–2012 industry changes and did not take place exclusively during those years. Moreover, this pattern of preparer use was negatively associated with the rise of software-aided returns, with the share of software-aided, self-prepared returns doubling from 20 to 40 percent of all EITC claimants over the period. These trends likely reflect increased access to computer software for tax filing (e.g., the availability of the IRS Free File option for the low-income), and thus offer little insight into the relationship between the industry reform and possible composition shifts.

FIGURE 6. Return Preparation Methods and Software Use for EITC Returns



NOTE: Returns using Volunteer Income Tax Assistance (VITA), Tax Counseling for the Elderly (TCE), or IRS assistance are included in total but the share is not shown.
SOURCE: IRS Individual and Sole Proprietor (INSOLE) files.

To help control for potential composition shifts and isolate the policy effect, I include a number of variables associated with individual compliance in the estimation. Two of these variables are dummies indicating the taxpayer's compliance level in the pre-reform years, measured as having a Discriminant Function System (DIF) score ranked top 5 percent or top 5-to-20 percent of the DIF distribution within a specific examination activity code computed over taxpayers in the sample. DIF is one of the computer scoring systems the IRS uses for audit filtering. For post-reform taxpayers, I look for the DIF scores and examination activity codes assigned

[16] Further study is needed to estimate the prevalence of this phenomenon.

to the 2006–2008 tax returns on which they appeared, either as a filer or a dependent, and pick the latest one of those found. If no prior returns for 2006–2008 are found, taxpayers are treated as if they did not have a high DIF score before reform.[17]

Furthermore, I include a dummy variable indicating if the taxpayer had a child age 18 or younger, per Social Security data, in the year when the tax return was filed. This exogenous parent variable serves as a control for individual compliance because, as mentioned earlier, parents' EITC claims are less likely to contain qualifying child errors than nonparents' claims. Relatedly, I also include in the estimation the share of EITC returns with children, by State and by year, that use a paid preparer. Including this variable is intended to capture the compliance consequence, if any, of State-specific trends in preparer use among taxpayers claiming EITC with a qualifying child.[18]

For the bunching indicator, because the focus is on self-employed taxpayers, I use the interaction $(Paid_i * Post_i * SE_i)$ to estimate the reform's effect specific to self-employed returns, noted as SE_i.

The 1-percent sample of EITC returns from the IRS's individual income tax population files for Tax Years 2004–2008 (pre-preform) and 2012–2016 (post-reform) totals approximately 2.5 million tax returns. Columns (1) through (3) of Table 3 show the summary statistics of these taxpayers, in total and by paid preparer use. For taxpayers claiming the EITC, use of paid preparers is associated with being self-employed, having children younger than age 19, and having a complex return as indicated by the filing of various tax schedules.

Finally, as an alternative specification to circumvent selection bias into the EITC pool, I added to the estimation a one-percent sample of low-income filers who did not claim the EITC in Tax Years 2004–2008 and 2012–2016. Unlike the linear regression (OLS) model applied for the EITC sample, probit estimations are conducted on the combined EITC and non-EITC samples.[19] Probit is appropriate with the inclusion of non-EITC returns because the nonlinear specification allows the marginal effect of an independent variable to vary with the values of covariates, some of which, such as whether a child is present, markedly differ between EITC claimants and non-claimants. Column (4) of Table 3 lists the summary statistics for the 5.1 million non-EITC taxpayers.

TABLE 3. Summary Statistics

Variable	All EITC Returns (1)		EITC Returns, Paid-Preparer (2)		EITC Returns, Self-Prepared (3)		Low-Income, Non-EITC Returns (4)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
EITC>0	1.00	(0.00)	1.00	(0.00)	1.00	(0.00)	0.00	(0.00)
EITC with Qualiying Child	0.76	(0.43)	0.80	(0.40)	0.70	(0.46)	0.00	(0.00)
Self-employed	0.20	(0.40)	0.23	(0.42)	0.14	(0.35)	0.12	(0.33)
Head-of-household	0.50	(0.50)	0.51	(0.50)	0.49	(0.50)	0.09	(0.28)
Nonparent claim for EITC with Qualifying Child	0.22	(0.41)	0.24	(0.42)	0.19	(0.39)	0.00	(0.00)
DDB rule violation	0.16	(0.37)	0.17	(0.37)	0.16	(0.36)	0.00	(0.04)
Income around EITC kink, if self-employed	0.17	(0.37)	0.15	(0.36)	0.20	(0.40)	0.02	(0.12)

[17] I also tried including a separate dummy variable indicating that a taxpayer did not appear on any return in 2006–2008. This additional control did not change key coefficients.

[18] In additional analysis not shown in the paper, I directly tested the hypothesis of composition changes with respect to individual compliance and found no evidence of such changes that would cause a spurious policy effect. Specifically, I found a disproportionate decrease in the share of parent taxpayers and a disproportionate increase in the share with a high pre-reform DIF score for paid-preparer EITC returns relative to self-prepared EITC returns following the industry changes. These results indicate that the composition of the paid-preparer pool has gotten relatively less compliant in the post-reform years. Had the paid-preparer pool become more compliant, there would be concern about a spurious policy effect due to taxpayer self-selection into preparation methods or preparer behavior.

[19] The income limit for this low-income sample is set at $5,000 plus the statutory EITC income limit (rounded to the nearest $5,000) for the maximum number of qualifying children allowed for the year. As with EITC returns, I limit the non-EITC sample to those 15 to 64 years old based on the primary filer's age.

TABLE 3. Summary Statistics—Continued

Variable	All EITC Returns (1)		EITC Returns, Paid-Preparer (2)		EITC Returns, Self-Prepared (3)		Low-Income, Non-EITC Returns (4)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Paid-preparer return	0.63	(0.48)	1.00	(0.00)	0.00	(0.00)	0.50	(0.50)
Post-reform year	0.54	(0.50)	0.48	(0.50)	0.63	(0.48)	0.52	(0.50)
Age of primary taxpayer	37.57	(11.09)	38.02	(11.14)	36.81	(10.95)	34.76	(14.44)
Primary taxpayer is male	0.47	(0.50)	0.49	(0.50)	0.44	(0.50)	0.58	(0.49)
Schedule A	0.07	(0.26)	0.08	(0.26)	0.07	(0.25)	0.15	(0.36)
Schedule B	0.10	(0.30)	0.11	(0.31)	0.09	(0.29)	0.24	(0.43)
Schedule C	0.24	(0.43)	0.27	(0.44)	0.19	(0.39)	0.12	(0.32)
Schedule D	0.02	(0.15)	0.03	(0.16)	0.02	(0.13)	0.08	(0.27)
Schedule E	0.03	(0.18)	0.04	(0.20)	0.02	(0.12)	0.06	(0.23)
Schedule F	0.01	(0.08)	0.01	(0.10)	0.00	(0.05)	0.01	(0.09)
Have child(ren) younger than 19	0.63	(0.48)	0.65	(0.48)	0.60	(0.49)	0.14	(0.35)
Top 15-20% DIF	0.12	(0.32)	0.12	(0.33)	0.10	(0.31)	0.08	(0.27)
Top 5% DIF	0.05	(0.22)	0.05	(0.22)	0.05	(0.21)	0.05	(0.21)
Share of the State's EITC families using a paid preparer in the year (%)	65.59	(10.64)	67.27	(10.16)	62.77	(10.82)	65.21	(10.82)
Tax Year 2004	0.09	(0.28)	0.10	(0.30)	0.07	(0.25)	0.09	(0.28)
Tax Year 2005	0.09	(0.29)	0.10	(0.30)	0.07	(0.25)	0.09	(0.28)
Tax Year 2006	0.09	(0.29)	0.10	(0.30)	0.07	(0.25)	0.10	(0.30)
Tax Year 2007	0.10	(0.29)	0.11	(0.31)	0.08	(0.27)	0.10	(0.31)
Tax Year 2008	0.10	(0.30)	0.10	(0.31)	0.08	(0.28)	0.10	(0.29)
Tax Year 2012	0.11	(0.31)	0.10	(0.30)	0.12	(0.32)	0.10	(0.31)
Tax Year 2013	0.11	(0.31)	0.10	(0.30)	0.13	(0.33)	0.10	(0.30)
Tax Year 2014	0.11	(0.31)	0.10	(0.30)	0.13	(0.33)	0.10	(0.30)
Tax Year 2015	0.11	(0.31)	0.09	(0.29)	0.13	(0.34)	0.11	(0.31)
Tax Year 2016	0.10	(0.31)	0.09	(0.29)	0.13	(0.34)	0.11	(0.31)
Observations	2,474,616		1,550,304		924,312		5,101,963	

SOURCES: Individual Returns Transaction File and the IRS Dependent Database

4. Estimation Results

Table 4 presents the OLS results on the EITC sample, and Table 5 reports the marginal effects calculated from the probit estimations on the combined EITC and non-EITC samples. I use the probit estimates to evaluate the marginal effects of paid preparer use before and after reform. The estimated reform's impact, listed at the bottom of Table 5, is therefore measured as the before-and-after change in these marginal effects for a specific potential error indicator. The full probit estimates are provided in the Appendix.

Using the identification strategy described in Section 3.2, I find evidence of the reform's effects on potential EITC compliance. Column (1) of Table 4 suggests that the industry changes in the late 2000s and early 2010s have lowered the likelihood that paid-preparer EITC returns claim the head-of-household status by 7.36 percentage points. For comparison, the probit estimation suggests a slightly larger effect at 7.63 percentage points in Table 5. With the treatment group, i.e., returns prepared by a paid preparer, representing 55 percent of all EITC returns in the post-reform years, these estimates suggest a reduction in the head-of-household filing by 4.0 (7.36*0.55) to 4.2 (7.63*0.55) percentage points among all EITC returns.

Since the dependent variable—head-of-household filing—is a potential error indicator, not an actual error, the estimates do not tell us how many filing status errors contributing to EITC overclaims have been eliminated. It is possible that the reform's impact has extended to taxpayers who do not make a filing status mistake but are deterred to claim the head-of-household status. IRS (2014) estimated that about 1.0 million, or 4.2 percent, of the EITC returns in 2006–2008 contained a filing status error that resulted in an EITC overclaim.[20] This error rate implies that about 8.4 percent of head-of-household EITC returns likely made a filing status error, following Table 3 that about half of all EITC returns claimed as head-of-household.[21] If we simply assume that the resulting deterrence is independent of whether a filing-status error is committed, then the estimates suggest that the industry reform has lowered the filing status errors by about 0.34–0.35 percentage points (4.0 or 4.2*0.084). Of course, the error-reduction rate would be higher if the deterred filers consisted disproportionally of those making a filing status error.

On other covariates, paid-preparer returns are more likely to claim the head-of-household status than self-prepared returns, and returns from the post-reform years are more likely to use this filing status than returns from the pre-reform years. Moreover, using the head-of-household status is positively associated with having a child under age 19 as well as having a high DIF score prior to the industry reform. Conversely, head-of household status is less used among male primary taxpayers, and is negatively associated with the taxpayer's age and the presence of various tax schedules.

Next, both OLS and probit results reveal a lowered post-reform probability of nonparent claims for EITC with a qualifying child for paid-preparer returns. The effect is estimated to be 3.40 percentage points in Table 4 and 2.79 percentage points in Table 5 for paid-preparer returns. As this treatment group makes up 55 percent of the post-reform EITC population, these estimates suggest an effect of 1.9 or 1.5 percentage points for all EITC returns. Like head-of-household filing, nonparent claims for EITC with a qualifying child are not always erroneous. The extent to which the estimated effects are indicative of reduced EITC qualifying child errors depends on the share of deterred nonparental claims that are erroneous. Leibel, *et al.* (2017) find that slightly over one-half of the EITC children claimed by a nonparent failed the qualifying child tests. Hence, using the assumption that one-half of the deterred nonparent claims were erroneous, the implied reduction in the error rate would be about 0.75 to 0.95 percentage points (1.5 or 1.9*0.5). For reference, IRS (2014) estimated that about 3.0 million, or 12.7 percent, of the EITC returns in 2006–2008 contained qualifying child errors that resulted in an EITC overclaim.

TABLE 4. OLS Results

Variable	Head-of-Household Filing Status	Nonparent Claim	DDB Rule Violation	Income at the First EITC Kink
	(1)	(2)	(3)	(4)
Paid-preparer return	0.0826***	0.0872***	0.0377***	-0.0033***
	(0.0013)	(0.0011)	(0.0008)	(0.0005)
Post-reform year	0.0209***	0.0746***	0.0761***	-0.0115***
	(0.0022)	(0.0020)	(0.0017)	(0.0010)
Self-employed (SE)				0.1015***
				(0.0020)
Paid*Post*SE				-0.0536***
				(0.0029)
Paid*Post	-0.0736***	-0.0340***	-0.0155***	0.0068***
	(0.0016)	(0.0014)	(0.0012)	(0.0006)

Footnotes at end of table.

[20] The count and error frequency are based on EITC claims with known errors in the IRS random audit study (2014). Some claims were denied but the causes were unknown.

[21] This calculation assumes that all filing status errors come from claims of the head-of-household status.

TABLE 4. OLS Results—Continued

Variable	Head-of-Household Filing Status	Nonparent Claim	DDB Rule Violation	Income at the First EITC Kink
	(1)	(2)	(3)	(4)
Post*SE				0.0636***
				(0.0025)
Paid*SE				-0.0021
				(0.0022)
Age	-0.0008***	-0.0015***	-0.0040***	-0.0010***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Male	-0.3726***	-0.0042***	0.1037***	-0.0248***
	(0.0011)	(0.0010)	(0.0008)	(0.0004)
Schedule A	-0.0013	0.0121***	-0.0371***	-0.0310***
	(0.0015)	(0.0013)	(0.0010)	(0.0005)
Schedule B	-0.1123***	-0.0830***	-0.0704***	-0.0191***
	(0.0013)	(0.0011)	(0.0008)	(0.0005)
Schedule C	-0.0871***	-0.0317***	0.0198***	0.0196***
	(0.0011)	(0.0010)	(0.0008)	(0.0006)
Schedule D	-0.0701***	-0.0809***	-0.0444***	-0.0067***
	(0.0023)	(0.0019)	(0.0014)	(0.0011)
Schedule E	-0.1204***	-0.0601***	-0.0604***	-0.0178***
	(0.0023)	(0.0019)	(0.0015)	(0.0009)
Schedule F	-0.1381***	-0.0725***	-0.0697***	-0.0152***
	(0.0044)	(0.0039)	(0.0030)	(0.0019)
Have child(ren) younger than 19	0.1981***	-0.2777***	-0.0321***	-0.0216***
	(0.0011)	(0.0010)	(0.0008)	(0.0004)
Top 5-20% DIF	0.0854***	0.0334***	0.0019**	-0.0202***
	(0.0012)	(0.0012)	(0.0009)	(0.0005)
Top 5% DIF	0.0721***	0.0135***	-0.0059***	-0.0248***
	(0.0019)	(0.0018)	(0.0015)	(0.0008)
Share of State EITC families using paid preparers	0.0003** (0.0001)	0.0018*** (0.0001)	-0.0010*** (0.0001)	-0.0002*** (0.0001)
State dummies	Yes	Yes	Yes	Yes
Observations	2,474,616	2,474,616	2,474,616	2,474,616

NOTE: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Column (2) of Table 4 shows that nonparent claims for EITC with a qualifying child are more likely to come from returns that use a paid preparer than from self-prepared returns. Such claims were also more frequently filed in the post-reform years than in the pre-reform years. Moreover, the probability of making a nonparent claim is positively associated with having a high DIF score prior to the industry reform, itemizing deductions, and living in a State in which paid-preparer use is more prevalent among EITC families. The probability, on the other hand, is negatively associated with the taxpayer's age, being male, having a child under age 19, and filing the other tax schedules.

Both Table 4 and Table 5 indicate a post-reform decline in the probability of a DDB rule violation for paid-preparer returns. The effect is estimated to be 1.55 percentage points based on OLS and 1.81 percentage points based on the probit estimation. Because this rule violation is an indicator of potential qualifying child errors, these estimates reassure support for improved compliance with qualifying child tests after the industry

changes. Assuming a 20-percent false-positive rate for the IRS scoring system, the estimates imply a policy effect on the reduction in qualifying child errors by 0.68 (1.53*0.55*0.8) to 0.80 (1.81*0.55*0.8) percentage points.[22]

TABLE 5. Marginal Effects of Paid-Preparer Use from Probit Results

Variable	Head-of-Household Filing Status	Nonparent Claim	DDB Rule Violation	Income at the First EITC Kink
	(1)	(2)	(3)	(4)
Paid-preparer, pre-reform	0.1184***	0.0600***	0.0366***	
	(0.0008)	(0.0005)	(0.0004)	
Paid-preparer, post-reform	0.0421***	0.0321***	0.0185***	
	(0.0008)	(0.0005)	(0.0005)	
Paid-preparer, pre-reform, non-SE				-0.0006**
				(0.0002)
Paid-preparer, pre-reform, SE				0.0096***
				(0.0009)
Paid-preparer, post-reform, non-SE				0.0030***
				(0.0002)
Paid-preparer, post-reform, SE				-0.0119***
				(0.0009)
Implied reform effect	-0.0763	-0.0279	-0.0181	-0.0251

NOTE: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Rule-breaking is more common among paid-preparer returns than self-prepared returns, and also more prevalent in post-reform years. Male taxpayers and Schedule C filers are more likely to be identified in the IRS system as violating at least one DDB rule. In contrast, the probability of being identified as rule-breaking is lower among taxpayers with children and among those filing more complex returns with various tax schedules. Other things equal, having a high DIF score in the pre-reform years has a very small and indeterminate effect on the probability of violating DDB rules.

For reporting income around the first EITC kink, OLS regression generates a negative coefficient of 5.36 percentage points for the policy effect, larger in magnitude than the marginal effect of 2.51 percentage points under the probit estimation.[23] Because compliant, self-employed taxpayers are not likely to adjust their reported income because of the industry changes, the entire estimated effect can be expected to reflect a reduction in noncompliance of self-employment income, with a magnitude ranging from 0.33 to 0.70 percentage points for all EITC returns.[24] For reference, IRS (2014) estimated that about 3.1 million, or 13.1 percent, of the EITC returns in 2006–2008 contained self-employment income errors that resulted in an EITC overclaim. As for other covariates, reporting income around the first kink of the EITC schedule is positively associated with filing schedule C and negatively associated with nearly all other covariates, including a high DIF score in the pre-reform years.

[22] GAO (2014) documented that the no-change rate of IRS correspondence audits ranged from 11 to 21 percent for all closed cases in Fiscal Years 2009 through 2013, including defaults. A false positive rate of 20 percent is likely close to the higher bound.

[23] For non-claimants, bunching is evaluated by the taxpayer's reported AGI and the number of children at home claimed for exemption. Both the OLS and probit effects are difference-in-difference-in-difference estimates, with an additional comparison against wage earners' income bunching.

[24] The treatment rate is 13 percent as 55 percent of post-reform EITC returns used a paid prepares, of which 23 percent reported self-employment income.

5. Conclusion

From mandatory electronic filing, required PTINs, to heightened EITC due diligence requirements and enhanced preparer enforcement activities, a series of legislative, regulatory, and tax administrative developments that took place in the late 2000s and early 2010s has reshaped the tax return paid-preparer industry. A logical question to follow is whether these efforts have achieved their stated goal of improving tax compliance through enhanced preparer standards and competence.

In this paper, I find evidence of a modest, positive effect of these industry changes on EITC compliance. In particular, the share of paid-preparer EITC returns potentially making eligibility or income errors is found lowered after the industry reform, indicating the policy's effect. Under plausible assumptions, the results imply that, for all EITC returns, the frequency of filing status misreporting has declined by 0.34 to 0.35 percentage points, the frequency of qualifying child errors by 0.68 to 0.95 percentage points, and the frequency of self-employment income misreporting by 0.33 to 0.70 percentage points.

Further research will strengthen and complement the findings of this paper. Although the four potential noncompliance indicators analyzed in this paper cover main sources of EITC errors, they are neither comprehensive nor are they derived from audit data on EITC overclaims. It will be a reasonable exercise to examine the reform's compliance effect using random audit results when additional random audit data from the post-reform years become available. Next, it is of policy interest to understand the exact channels through which these compliance effects occur. Studying the compliance consequences by type of preparers—small-volume vs. large-volume paid preparers, unenrolled preparers vs. those from national tax preparation firms, etc.—or studying preparer longitudinal data will help shed light on the issues.

References

Acs, Gregory, and Elaine Maag, 2005. "Irreconcilable Differences?: The Conflict Between Marriage Promotion Initiatives for Cohabiting Couples with Children and Marriage Penalties in Tax and Transfer Programs." Paper No. B-66 in *New Federalism: National Survey of America's Families*. The Urban Institute, Washington, DC.

Allingham, Michael G., and Agnar Sandmo, 1972. "Income Tax Evasion: A Theoretical Analysis." *Journal of Public Economics* Vol. 1 No. 3–4, 323–338.

Book, Leslie, 2007. "Study of the Role of Preparers in Relation to Taxpayer Compliance with Internal Revenue Laws." In *National Taxpayer Advocate 2007 Annual Report to Congress* Vol. 2, 44–74. Internal Revenue Service, Washington, DC.

Chetty, Raj, John N. Friedman, Peter Ganong, Kara E. Leibel, Alan H. Plumley, and Emmanuel Saez, 2012. *Taxpayer Response to the EITC: Evidence from IRS National Research Program.* Retrieved October 9, 2018, from http://www.rajchetty.com/chettyfiles/eitc_nrp_tabs.pdf.

Dubin, Jeffrey A., Michael J. Graetz, Michael A. Udell, and Louis L. Wilde, 1992. "The Demand for Tax Return Preparation Services." *The Review of Economics and Statistics* Vol. 74 No. 1, 75–82.

GAO, 2006. *Paid Tax Return Preparers: In a Limited Study, Chain Preparers Made Serious Errors.* GAO-06–563T. GAO, Washington, DC.

GAO, 2014. *IRS Correspondence Audits: Better Management Could Improve Tax Compliance and Reduce Taxpayer Burden*, GAO-14–479. GAO, Washington, DC.

GAO, 2016. *Refundable Tax Credits: Comprehensive Compliance Strategy and Expanded Use of Data Could Strengthen IRS's Efforts To Address Noncompliance*, GAO-16–475. GAO, Washington, DC.

Internal Revenue Service, 2002. *Compliance Estimates for Earned Income Tax Credit Claimed on 1999 Returns*. Internal Revenue Service, Washington, DC.

Internal Revenue Service, 2009. *Return Preparer Review*. Internal Revenue Service Publication 4832 (Rev. 12–2009). Internal Revenue Service, Washington, DC.

Internal Revenue Service, 2014. *Taxpayer Compliance for the Earned Income Tax Credit Claimed on 2006–2008 Returns.* Internal Revenue Service Report 5162 (8–2014). Internal Revenue Service, Washington, DC.

Internal Revenue Service, 2016. *Tax Gap Estimates for Tax Years 2008–2010*. Retrieved October 9, 2018, from https://www.irs.gov/pub/newsroom/tax%20gap%20estimates%20for%202008%20through%202010.pdf.

Internal Revenue Service, 2018. *Internal Revenue Service Data Book, 2017*. Internal Revenue Service, Washington, DC.

Jones, Maggie R., 2017. *Tax Preparers, Refund Anticipation Products, and EITC Noncompliance.* Center for Administrative Records Research and Applications (CARRA) Working Paper Series 2017–10. U.S. Census Bureau, Washington, DC.

Klepper, Steven, Mark Mazur, and Daniel Nagin, 1991. "Expert Intermediaries and Legal Compliance: The Case of Tax Preparers." *Journal of Law and Economics* Vol. 34 No.1, 205–229.

Klepper, Steven, and Daniel Nagin, 1989. "The Role of Tax Preparers in Tax Compliance." *Policy Sciences* Vol. 22 No. 2, 167–194.

Langetieg, Patrick, Mark Payne, and Melissa Vigil, 2013. "Return Preparer Industry Analysis." In *2013 IRS Research Bulletin, Tax Administration at the Centennial: An IRS-TPC Research Conference*. Ed. Alan Plumley. Internal Revenue Service, Washington, DC. 17–44. Retrieved October 9, 2018 from https://www.irs.gov/pub/irs-soi/13rescon.pdf.

Leibel, Kara, 2014. *Taxpayer Compliance and Sources of Errors for the Earned Income Tax Credit Claimed on 2006–2008 Returns.* Internal Revenue Service Technical Paper Publication 5161 (8–2014). Internal Revenue Service, Washington, DC.

Leibel, Kara, Emily Y. Lin, and Janet McCubbin, 2017. "Social Welfare Considerations of EITC Qualifying Child Noncompliance." Mimeo, U.S. Department of the Treasury.

Lin, Emily Y., and Patricia K. Tong, 2012. "Marriage and Taxes: What Can We Learn from Tax Returns Filed by Cohabiting Couples?" *National Tax Journal* Vol. 65 No. 4, 807–826.

Lin, Emily Y., and Patricia K. Tong, 2014. "Effects of Marriage Penalty Relief Tax Policy on Marriage Taxes and Marginal Tax Rates of Cohabiting Couples." In *National Tax Association 2014 Conference on Taxation Proceedings*. Retrieved October 9, 2018, from https://www.ntanet.org/wp-content/uploads/proceedings/2014/018-lin-tong-effects-marriage-penalty-relief-tax.pdf.

Maag, Elaine, and Gregory Acs, 2015. *The Financial Consequences of Marriage for Cohabiting Couples with Children*. The Urban Institute, Washington, DC. Retrieved October 9, 2018, from http://webarchive.urban.org/UploadedPDF/2000366-the-financial-consequences-of-marriage.pdf.

McCubbin, Janet, 2000. "EITC Noncompliance: The Determinants of the Misreporting of Children." *National Tax Journal* Vol. 53 No.4 Part 2, 1135–1164.

Saez, Emmanuel, 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* Vol. 2 No. 3, 180–212.

Treasury Inspector General for Tax Administration, 2008. *Most Tax Returns Prepared by a Limited Sample of Unenrolled Preparers Contained Significant Errors*. 2008–40–171. Treasury Inspector General for Tax Administration, Washington, DC.

U.S. Department of the Treasury, 2016. *Report to Congress on Strengthening Earned Income Tax Credit Compliance Through Data Driven Analysis*. U.S. Department of the Treasury, Washington, DC. Retrieved October 9, 2018, from https://www.treasury.gov/resource-center/tax-policy/Documents/Report-EITC-Data-Driven-Compliance-2016.pdf.

U.S. Department of the Treasury, 2018. *Agency Financial Report: Fiscal Year 2018*. U.S. Department of the Treasury, Washington, DC.

Appendix A

TABLE A-1: Probit Result

Variable	Head-of-Household Filing Status	Nonparent Claim	DDB Rule Violation	Income at the First EITC Kink
	(1)	(2)	(3)	(4)
Paid-prepared return	0.4016***	0.4699***	0.3225***	-0.0081***
	(0.0027)	(0.0036)	(0.0038)	(0.0031)
Post reform year	0.1661***	0.3637***	0.4043***	-0.0771***
	(0.0046)	(0.0058)	(0.0063)	(0.0058)
Self-employed (SE)				0.2789***
				(0.0099)
Paid*Post*SE				-0.2021***
				(0.0101)
Paid*Post	-0.2602***	-0.2807***	-0.2117***	0.0505***
	(0.0034)	(0.0044)	(0.0046)	(0.0043)
Post*SE				0.2778***
				(0.0083)
Paid*SE				0.0823***
				(0.0078)
Age	0.0140***	0.0114***	-0.0004***	-0.0129***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Male	-0.7758***	-0.1285***	0.2102***	-0.1801***
	(0.0024)	(0.0028)	(0.0027)	(0.0020)
Schedule A	-0.2154***	-0.4067***	-0.4803***	-0.5829***
	(0.0030)	(0.0041)	(0.0046)	(0.0049)
Schedule B	-0.4219***	-0.5324***	-0.5223***	-0.1302***
	(0.0028)	(0.0039)	(0.0042)	(0.0031)
Schedule C	-0.1418***	0.1790***	0.2723***	0.2351***
	(0.0028)	(0.0031)	(0.0030)	(0.0074)
Schedule D	-0.3215***	-0.5293***	-0.4296***	0.0040
	(0.0053)	(0.0084)	(0.0092)	(0.0050)
Schedule E	-0.3359***	-0.3846***	-0.3751***	-0.1017***
	(0.0059)	(0.0080)	(0.0083)	(0.0062)
Schedule F	-0.5560***	-0.3779***	-0.3517***	-0.2271***
	(0.0184)	(0.0221)	(0.0204)	(0.0147)
Have child(ren) younger than 19	1.2343***	0.0166***	0.5887***	0.0537***
	(0.0022)	(0.0028)	(0.0027)	(0.0022)
Top 5-20% DIF	0.2498***	0.1826***	0.0792***	-0.0664***
	(0.0029)	(0.0036)	(0.0038)	(0.0034)
Top 5% DIF	0.6947***	-0.0567***	-0.0977***	-0.2305***
	(0.0042)	(0.0054)	(0.0056)	(0.0060)
Share of State EITC families using paid preparers	0.0029*** (0.0002)	0.0077*** (0.0003)	0.0005* (0.0003)	0.0022*** (0.0003)
State dummies	Yes	Yes	Yes	Yes
Observations	7,576,579	7,576,579	7,576,579	7,576,579

NOTE: Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Effects of Recent Reductions in the Internal Revenue Service's Appropriations on Returns on Investment

Janet Holtzblatt (Urban-Brookings Tax Policy Center) and
Jamie McGuire (Joint Committee on Taxation)[1]

Since 2010, funding for the Internal Revenue Service (IRS) has dropped by 24 percent, after adjustment for inflation.[2] The cuts have been deepest in enforcement activities, including audits and collections. As a consequence, the percentage of taxpayers who are audited has fallen by nearly half.

Nonetheless, a decline in resources could cause enforcement revenues to increase *relative* to the costs of audits and collections. For any level of appropriations, the IRS should—ignoring all other considerations—select the cases known to have the highest returns on investment (ROIs) to be audited. Therefore, as appropriations fall, the average return on investment should increase.

Other factors, however, may hinder the IRS from allocating resources based solely on historic ROIs. The period has been marked by other changes that affect efficiency. Additional responsibilities (resulting from newly enacted legislation, for example) and changing expectations for the agency (such as increased demand for high-quality customer service) place greater pressures on the agency's flexibility. Meanwhile, the IRS infrastructure is weakening as its skilled workforce retires and its computer systems become increasingly outdated. As a result, ROIs may, on average, decline. Because the impact of the new responsibilities and faltering infrastructure may hinder some types of enforcement actions more than others, the impact of funding reductions may also differ by the type of enforcement activity.

Still another constraint may be the IRS's concern over public perception. For example, if the cost of auditing lower-income taxpayers is substantially lower than the expense of examining high-income individuals, the efficient choice may be to allocate more resources to auditing people with limited ability to dispute the IRS's assessments. But that may not be the image that the IRS's officials and staff want to project—in part, because that perception may increase evasion by higher-income taxpayers.

In this paper, we use confidential IRS data to compare the costs and returns on examinations that were initiated or in progress in 2010 and 2017. Our estimates exclude the indirect effects of IRS enforcement activities—that is, the reduction in voluntary compliance that may occur when the IRS conducts fewer audits. We find, on average, that the ROI fell slightly between those 2 years. The average ROIs increased for low-cost audits (those conducted through the mail) but generally declined for the more expensive audits that require face-to-face interactions with taxpayers and which cover more complicated issues. We use the findings of this investigation to estimate the effect on the Federal budget if the IRS enforcement budget were restored to 2010 levels.

---

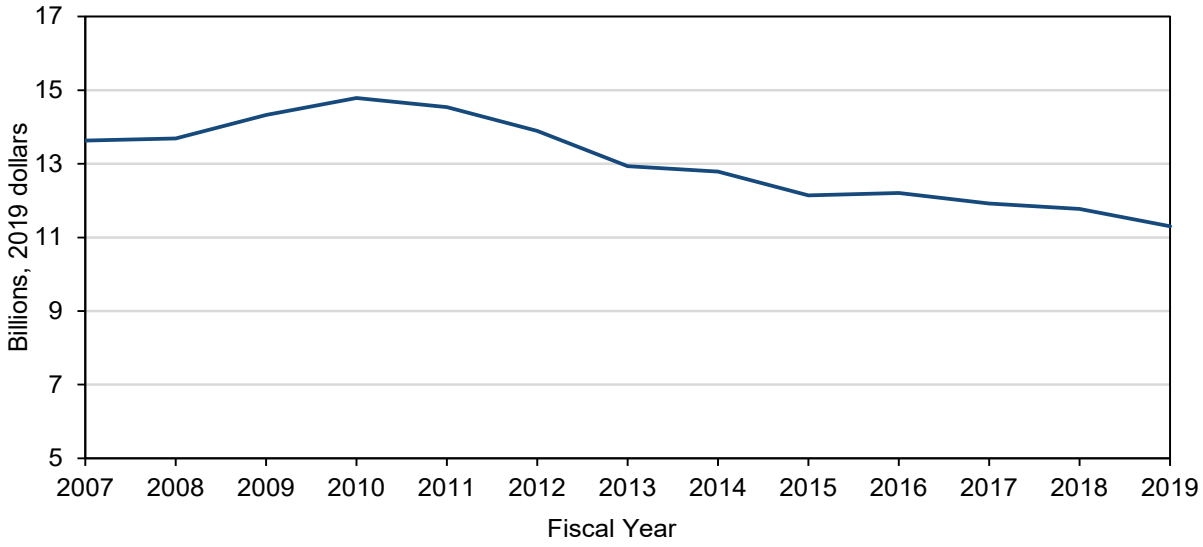
[1] This paper was prepared for the 9th Annual Internal Revenue Service-Tax Policy Center Joint Research Conference on Tax Administration, held in Washington D.C. on June 20, 2019. This paper embodies work undertaken for the staff of the Joint Committee on Taxation, but as members of both parties and both houses of Congress comprise the Joint Committee on Taxation, this paper should not be construed to represent the position of any member of the Committee. Janet Holtzblatt's work on this paper was made possible by a grant from Arnold Ventures and an anonymous funder. The statements made and the views expressed are those of the authors 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/aboutus/our-funding/funding-principles; further information on Brookings' donor guidelines is available at https://www.brookings.edu/donor-guidelines/. The authors wish to thank Thomas Barthold, Robert Harvey, Ronald Hodge, Mark Mazur, Alan Plumley, Kyle Richison, and Eric Toder for helpful comments and suggestions.

[2] Unless otherwise noted, all years referred to in the paper are fiscal years, and dollar amounts have not been adjusted for inflation.

IRS Resources

In 2019, the Internal Revenue Service received appropriations totaling $11.3 billion—about 24 percent less than it received in 2010. Appropriations for the IRS have been declining (in 2019 dollars) continuously over the past decade (Figure 1). These reductions have been accompanied by other developments that affect the IRS's efficiency—including the incremental expansions of the agency's role, the steady departure of its most skilled staff, and the continued aging of its computer systems.

FIGURE 1. IRS Appropriations, Fiscal Years 2007–2019



SOURCE: Appropriation Acts and Internal Revenue Service, *Budget in Brief*, various years. Amounts were adjusted to 2019 levels: For personnel costs, inflation was measured using the employment price index for wages and salaries of private industry workers; for all other spending, the measure of inflation was the chain-type price index for U.S. gross domestic product.

IRS Budget Accounts

Congress's Appropriations Committees distribute the IRS's funding among four different accounts:

- Taxpayer services, which funds pre-filing taxpayer assistance and submission processing;
- Enforcement, including examinations and collections;
- Operations support, which maintains the IRS's infrastructure (from facilities maintenance to computer systems used by more than one division); and
- Business systems modernization, which underwrites investments in technology.

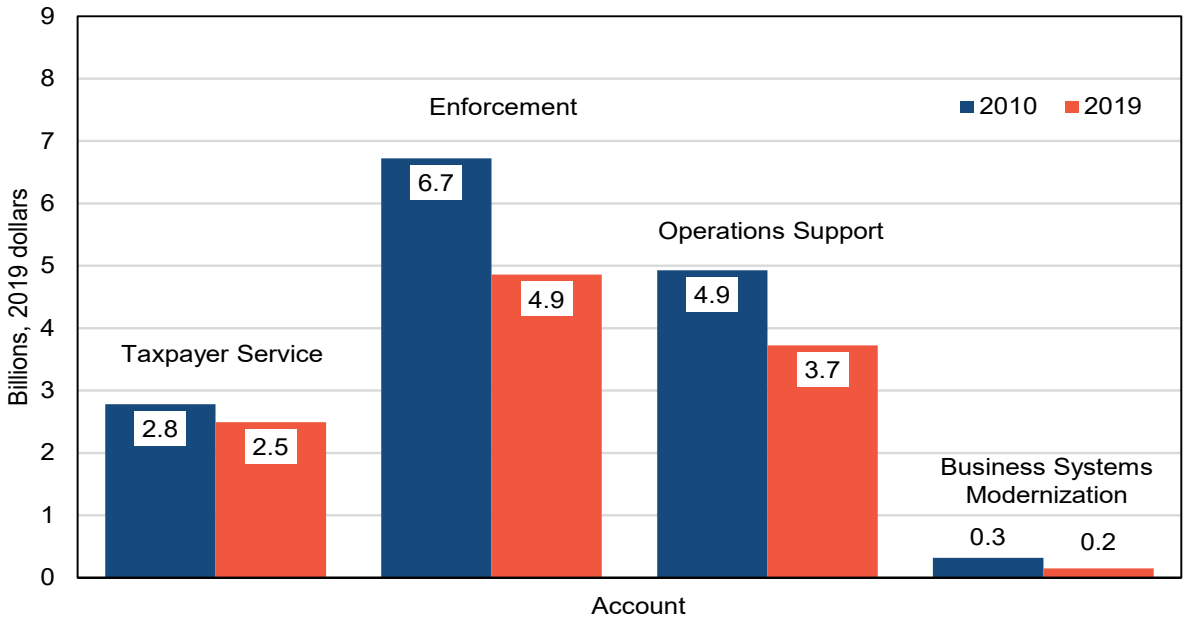
The largest of those accounts is enforcement, with funding in FY 2019 set at $4.9 billion, and the smallest is business systems modernization with a budget of $150 million.

Each budget account has experienced substantial cutbacks since 2010. The biggest cutbacks occurred in the enforcement account with funding falling by nearly $2 billion (28 percent) from 2010 levels (Figure 2).[3] In some years, the appropriated amounts for IRS enforcement have been further diminished because funds were shifted to other IRS accounts after the appropriations legislation had been enacted.[4]

[3] In dollars, the cuts in funding of business systems modernization were the smallest among the IRS accounts, but the relative reduction was greatest; appropriations were cut by half relative to 2010 levels.

[4] The actual amounts allocated to each of the IRS accounts sometimes differ from the amounts contained in appropriations acts. Transfers between the accounts have been allowed, though the amounts transferable are capped by the appropriations acts and are subject to the approval of the Appropriations Committees. The Consolidated Appropriations Act of 2017, for example, stated "Not to exceed 5 percent of any appropriation made available in this Act to the Internal Revenue Service may be transferred to any other Internal Revenue Service appropriation upon the advance approval of the Committees on Appropriations." Accordingly, $220 million was transferred from Enforcement to Taxpayer Services ($90 million) and Operations Support ($130 million).

FIGURE 2. IRS Appropriations by Account, Fiscal Years 2010 and 2019



SOURCE: Appropriation Acts and Internal Revenue Service, *Budget in Brief*, various years. Amounts were adjusted to 2019 levels: For personnel costs, inflation was measured using the employment price index for wages and salaries of private industry workers; for all other spending, the measure of inflation was the chain-type price index for U.S. gross domestic product.

Funding for New Responsibilities

As IRS appropriations have declined, Congress enacted legislation that assigned new responsibilities to the agency, including:

- Administration of new tax credits for health insurance coverage and the enforcement of health coverage mandates (Affordable Care Act—ACA—in 2010);
- Processing of reports of financial assets held abroad by U.S. citizens and related enforcement actions (Foreign Account Tax Compliance Act—FATCA—in 2010);
- Acceleration of processing and matching of W-2s to tax returns combined with a delay of payments of certain refundable tax credits so that claimants' earnings could be verified (Protecting Americans from Tax Hikes Act—PATH—in 2015); and
- Major changes to the tax code and forms in the 2017 tax act (P.L. 115-97, commonly referred to as the Tax Cuts and Jobs Act or TCJA).

Without additional funding, those new responsibilities would compete with the ongoing activities of the IRS for resources.

In some cases, funds were provided for a portion of the costs of implementing the new responsibilities. The largest amount—$488 million over a 3-year span—was for the development of the administrative infrastructure for the ACA. Those resources were on top of the annual IRS appropriations (and thus not included in Figure 1) and came from the Health Insurance Reform Implementation Fund. The fund was managed by the Department of Health and Human Services and provided support for Government agencies with ACA-related responsibilities. After 2012, the IRS no longer received funding specifically intended for administration of their new health responsibilities.

Funding specifically provided for the implementation of other initiatives was either smaller than the amounts transferred from the health fund (and like ACA, temporary) or nonexistent. The IRS received about

$397 million in the 2018 and 2019 appropriations acts for implementation of the 2017 tax act. The 2016 appropriations act—the first appropriations act after PATH's enactment—did not explicitly contain funding for carrying out PATH, which required the costly acceleration of the matching of tax returns and information returns. However, the act included $290 million to be spread—at the IRS's discretion—across taxpayer services, enforcement, or operations support for several different initiatives, one of which was the improvement of prevention of refund fraud and identity. No arrangements were made for the implementation of the new FATCA requirements, despite an estimate by the Treasury Inspector General for Tax Administration that its implementation cost $380 million (largely for information technology) from 2010 through 2017 (Treasury Inspector General for Tax Administration (2018)).

Personnel

Even before the recent cutbacks in IRS funding, the agency's workforce was shrinking. The average number of full-time-equivalent (FTE) staff fell from about 112,000 in 1990 to nearly 95,000 in 2010. From 2010 to 2018, however, the number of FTEs decreased by more than it had in the prior 20 years—a drop of an additional 21,000 FTEs.

Over two-thirds of the decrease in personnel occurred in examinations and collections, with FTEs declining from about 45,000 in 2010 to nearly 31,000 in 2018. Revenue agents and officers are trained to conduct the most difficult types of audits, but their numbers fell from nearly 20,000 in 2010 to about 12,000 in 2018.

Technology

The IRS's ability to enforce the tax code is also closely linked to the state of its computers and the long-term challenges of the IRS's computer modernization program. A 2016 report from the Government Accountability Office (GAO) found that the IRS's individual master file and business master file still relied on computer programming language developed more than 50 years ago (Government Accountability Office (2016)). Former IRS Commissioner John Koskinen likened the IRS's technology "to driving a Model T automobile that has satellite radio and the latest GPS system" (Koskinen (2015)).

The legacy design of the IRS's computer system often delays implementation of its new responsibilities. For example, PATH moved up the deadline for filing W-2s to January 31st from the end of February for paper forms and from March 31st for electronic transmission. The intent of that acceleration was to enable the IRS to match the W-2s to tax returns during processing and to verify earnings before refunds were paid. PATH also delayed payment of the earned income tax credit (EITC) and the refundable portion of the child tax credit (formally called the additional child tax credit) until mid-February. In 2017, however, the W-2s could be downloaded to the IRS's computer system only once a week, and the resulting delays in matching prevented the agency from verifying the earnings of more than half of returns claiming those two credits before refunds were paid out (Government Accountability Office (2018)). A year later, the IRS was able to verify wage income on 87 percent of returns claiming the EITC and additional child tax credit.

Have There Been Offsetting Efficiency Savings?

Reductions in administrative costs could free up money, which could be used instead for enforcement or customer services. The IRS's officials often cite examples of efforts to reduce the costs of its organization. Those examples include reductions in training and travel, with savings of $248 million between 2010 and 2014, and initiatives to reduce office space and rent payments, with annual savings of about $50 million (Koskinen (2015)).

Although the direct monetary savings from reductions in training, travel, and rent are easily observable, quantifying the impact of those cutbacks on the IRS's efficiency is not. Reducing office space shrinks the amount of space the IRS leases and hence the rent it pays, but it may also reduce workers' productivity. To some extent, the IRS may respond by offering more workers the option to work remotely, but it is unclear whether workers' productivity increases because of fewer disruptions from their colleagues or declines due to fewer face-to-face interactions with those colleagues or less monitoring by supervisors.

A stronger indicator of efficiency improvements is the growth in electronic filing, which grew from 50 percent of returns in 2010 to over 70 percent in 2018. In 2013, the IRS spent 18 cents to process each electronic return—compared to $3.54 per paper return (Government Accountability Office (2014b)). Although those savings do not directly affect the enforcement budget, the greater utilization of electronic filing has the potential of reducing enforcement costs because more information can be captured at lower cost than from paper returns.

IRS Enforcement

Nearly everything that the IRS does can be characterized as a way to improve compliance, but certain of its activities are more directly related to enforcement than others. In this paper, we focus on the cost of the automated underreporting program, audits, collections, and appeals.

Automated Underreporting Program

One type of enforcement action is the automated underreporting (AUR) program. After the processing season ends, information returns (such as W-2s and 1099s) are edited and matched to tax returns.[5] Discrepancies between the items reported on tax returns and the amounts shown on the information returns may result in a notice to the taxpayer indicating the additional amount of taxes that may be owed. Although its name suggests that the process is mechanical, it becomes more labor intensive beyond the initial identification of discrepancies.

Not all discrepancies lead to notices. Before a notice is sent, the AUR examiners review the return and other information available to the IRS to determine whether the discrepancy can be resolved without contacting the taxpayer. If the discrepancy cannot be resolved internally, the IRS will send a CP 2000 notice informing the taxpayer of the additional taxes that are owed. Taxpayers who agree with the assessment pay the additional taxes due. Those who disagree with that assessment can provide the IRS with documentation in support of their reported income. If the examiner determines that the documentation is insufficient, a notice of deficiency is issued, and the taxpayer can appeal to the U.S. Tax Court.

Because of resource constraints, the IRS sets criteria—including a dollar threshold—to prioritize the cases worked. Over the period from 2009 through 2013, only about one in five discrepancies was selected for investigation each year (Treasury Inspector General for Tax Administration (2015)). The threshold (which is not publicly disclosed) is probably raised when funding and personnel decline as occurred over the past decade. Between 2010 and 2018, the number of closed cases dropped by 30 percent as the number of AUR staff fell by 40 percent.

Audits

IRS conducts three different types of audits, which vary in scope and the types of taxpayers affected. The three types also differ in timing and cost.

The simplest type of audit is conducted through correspondence with the taxpayer and focuses on a small number of selected items (such as whether a child is related and resides with the taxpayer, which are criteria for determining eligibility for certain child-related tax benefits). In response to a letter from the IRS, taxpayers must provide documentation in support of their claims. For about half of correspondence audits in 2013, the taxpayer's refund was frozen until the disputed issues were resolved (GAO (2014a)). Taxpayers who claim the EITC are the subject of about half of correspondence audits, and most EITC audits are conducted through correspondence.

The scope of the other two types of audits extends to items reported on the entire tax return and requires face-to-face interaction between the IRS and the taxpayer. Those interactions can occur in the IRS office (office audits) or in the taxpayer's home, place of employment, or elsewhere (field audits).

[5] PATH accelerated the matching of W-2s and tax returns to occur during the filing season. However, those forms are still not completely edited at that point, which may result in more notices sent to taxpayers after the filing season has ended.

In total, the number of audits fell from 1.7 million in 2010 to 1 million in 2018 (*Internal Revenue Service Data Books*, various years). The audit rate—defined as the ratio of the number of audits closed in a given fiscal year to the number of tax returns filed in the prior tax year—fell from 0.9 percent in 2010 to 0.5 percent in 2018. For individual filers, the audit rate declined at about the same rate—by 1.1 percent to 0.6 percent. About three-quarters of audits in both 2010 and 2018 were conducted through correspondence.

Certain segments of the population are more likely to be targeted for audits than others. In 2010, the audit rate was 0.5 percent for taxpayers with relatively simple returns: they had positive income under $200,000, no self-employment income, and did not claim the EITC. Among EITC claimants, the audit rate was 2.4 percent. Audit rates were generally higher for individuals reporting business income and for C-corporations. For individual taxpayers with gross business (excluding farm) receipts in excess of $200,000, the audit rate was 3.3 percent. And nearly all the largest corporations—with $20 billion or more of assets—were subject to audits in 2010.

Since 2010, audit rates have fallen across all groups of taxpayers. In 2018, only 0.2 percent of the simplest returns were audited. For EITC claimants, the audit rate fell by one percentage point—to 1.4 percent. For individual taxpayers with $200,000 or more of gross business income, the audit rate declined by 1.4 percentage points to 1.9 percent. And for the largest corporations, the audit rate dropped by about half.

Post-Audit

After audits, taxpayers and the IRS may continue to interact—either with collections or with taxpayers' challenges to the examiners' assessments and the collection process.

The collection process begins once a taxpayer underpays taxes and is not limited to examinations. Taxpayers first receive a notice informing them of the taxes owed plus penalties and interest (which will continue to accrue if taxes are not paid). If not paid, the next steps become increasingly labor intensive. They may involve establishment of an installment plan, requests for hardship delays, and imposition of a Federal tax lien on the taxpayer's property.

Taxpayers can turn to the IRS's Office of Appeals if they dispute a proposed tax assessment or have difficulty with collections. The appeals officials do not raise new issues or reopen issues on which the taxpayers and auditors have already reached agreement. If the taxpayer submits new information or evidence, Appeals will return the case to Examination for further review. A similar process occurs with respect to resolution of collection problems.

Disputes between taxpayers and the IRS may also end up in court. If the taxpayer disagrees with the IRS's assessment, a notice of deficiency is issued. Taxpayers have 90 days to file a petition with the Tax Court and are not required to pay the amounts owed before the case is settled. Attorneys in the IRS's Chief Counsel's offices represent the Government in those cases and others that end up in bankruptcy court. Other tax cases may be litigated in other Federal courts by attorneys from the Justice Department, in coordination with the IRS lawyers.

Enforcement Revenues

The IRS measures enforcement revenue as the sum of the amount collected from taxpayers as the consequence of the three major enforcement programs: the automated underreporting program, examinations, and collections. The agency's measure includes the amount collected in a fiscal year, which will include tax, interest, and penalties from multiple years.

As enforcement funding fell after 2010, enforcement revenue also declined—from $66 billion in 2010 to $60 billion in 2018 (2019 dollars). However, the average return on investment—defined by the IRS as enforcement revenue collected in a fiscal year to enforcement funding in that year—increased from $8.80 (Treasury Department (2013)) for a dollar of funding to $10.70 (Treasury Department (2019)). The IRS measure of the return on investment does not necessarily link the costs of audits in a particular year to the amounts ultimately collected for those audits, unless the audit and the amount collected occurred within the same year.

Methodology

We define the average return on investment as the amount collected by the IRS relative to the costs of the enforcement activities. Following Hodge, et al. (2016), we limit the amount collected to taxes, excluding the interest and penalties attributable to the late payment.[6] We include any amounts paid by taxpayers audited in 2010 or 2017 from the beginning of an enforcement activity (starting with the automated underreporting program) through the collections process. Similarly, the costs of the enforcement action include all those associated for the individuals who were audited in either 2010 or 2017—and thus extend over several years. As a result, our estimates of enforcement revenue and costs differ from those reported in the IRS's budget documents, which include interest and penalties and do not necessarily link the costs of an audit of a taxpayer to the amounts ultimately collected from that particular taxpayer.

Only the direct enforcement revenues are included in our estimates. Enforcement actions may spur taxpayers to become more compliant, but the estimation of the indirect savings from those improvements in compliance is outside the scope of this analysis. Nor can we determine whether the IRS's assessments of the amounts owed by taxpayers were correct.

Enforcement Data

Our analysis relies on the Enforcement Revenue Information System (ERIS), an IRS data set that follows each tax return from the inception of an enforcement activity (including the use of more-automated systems that may precede or supplant audits) through collections. ERIS contains information on the issues that triggered the enforcement action, how the issues were identified, the duration of the process, and the number of hours worked by the IRS's personnel, and their grade (level) on the Government's pay scale. Information on the amounts of assessments and enforcement revenues is also included in ERIS.

We focus on all tax returns (including both individuals and corporations) undergoing examinations in 2010 and 2017. Because we are interested in the impact of the IRS's declining enforcement resources, we include both audits initiated in 2010 (or 2017) as well as those that began in a prior fiscal year but had not been completed before October 1, 2009 (or 2016). Moreover, some audits in our sample were not completed in 2010 (or 2017). Thus, the total number of audits in our analysis is greater than the counts shown in the *IRS Data Books*: in our sample, 3.2 million audits in 2010 and 1.8 million in 2017, compared to 1.7 million and 1 million, respectively, in the *Data Books*. Moreover, the decrease in audits between 2010 and 2017 was driven by reductions in newly-initiated audits. The number of newly-initiated audits in 2017 was nearly half of the number of audits begun in 2010, whereas previously-initiated examinations fell by about a third relative to 2010.

ERIS also contains more detailed data by type of audit than contained in the *IRS Data Books*. Whereas the *Data Books* combined office and field audits into one category, ERIS separates the two categories. In both years, the number of field audits was about double that of office audits, but about two-thirds of all examinations were conducted by mail (Table 1).

TABLE 1. Types of Examinations, by Percentage Distribution, Fiscal Years 2010 and 2017

Type of Examination	2010	2017
Percentage distribution of examinations		
Field	23	21
Office	12	11
Correspondence	65	68
Total	100	100
Number of examinations (thousands)	3,206	1,827

SOURCE: IRS Enforcement Revenue Information System

[6] In their analysis of ROIs, Hodge, *et al.* exclude interest and penalties because it is not the IRS's objective to maximize such payments.

Despite the fall in the total caseload, the composition of the caseload did not change substantially. The share of total returns undergoing correspondence audits rose by 3 percentage points, offset by a drop in the share of field examinations (by 2 percentage points) and office examinations (by 1 percentage point). But compared to 2010, a substantially smaller share of field examinations in 2017 were new: 56 percent of all field examinations in 2010 had been initiated that year, whereas 45 percent of field examinations in 2017 were new. In contrast, the share of correspondence examinations that were new declined by only 4 percentage points between 2010 and 2017. This difference suggests that a greater share of audits will be conducted through correspondence over time, if those trends continue.

Measuring Costs

Our estimates of the IRS costs are limited to labor compensation, including both salaries and benefits, of the IRS employees directly involved in the enforcement activity (i.e., we do not have information on the time spent by support staff or managers). We thus do not account for the costs of buildings, computers, and other physical infrastructure that support the work of the IRS. That omission may not significantly affect our estimates. Ninety-four percent of the IRS enforcement budget is attributable to personnel compensation with the remainder—about $275 million in 2017—paying for travel, rent, utilities, operations and maintenance of facilities, research and development, equipment, and so forth (*IRS Data Book* (2018)). Most of the infrastructure supporting enforcement activities is shared with other programs and would probably be needed by the IRS even if the enforcement budget was not increased.

We also limit our analysis to the IRS's costs, although other agencies may incur expenses related to tax enforcement. For example, tax disputes that end up in Federal and State courts (other than in the United States Tax Court or bankruptcy courts) are often tried by lawyers in the Justice Department's Tax Division. In 2017, the division's appropriation was set at $107 million, and the Justice Department estimates that its civil litigation (about three-quarters of its budget) brought in $451 million each year, on average, between 2013 and 2017 (Justice Department (2019)).

To estimate the labor costs of IRS, we multiplied the number of hours worked by personnel by their hourly wage. From ERIS, we know the worker's pay grade at each point in the enforcement process, but we do not know their step within the grade. We assumed that each person was in the middle of the pay schedule for their grade (which is about step 5). Each year, the Office of Personnel Management adjusts Federal salaries for differences, by major metropolitan areas, for the disparity between public and private sector pay as well as for differences in the cost of living. IRS enforcement personnel are spread throughout the country, but workers' locations vary by their roles: Different types of enforcement activities are concentrated in different regions of the country. To account for those differences, we computed the median locality adjustment by type of activity. We then applied the activity-specific median to the basic wages of workers in that particular category. To the hourly wage estimates, we added the costs of employee benefits, based on findings of Congressional Budget Office's reports on Federal pay (Congressional Budget Office (2012); Falk (2012); Congressional Budget Office (2017)).

Although our sample is limited to taxpayers who were subjects of audits in 2010 and 2017, we include all costs incurred—from any associated with the automated underreporting program that precedes the audits to those associated with appeals, counsel, and collections. Thus, some costs may have been incurred prior to 2010 or 2017 or after those years.

Adjustments to Data

Before computing the average ROIs, we made several adjustments to the data. The first two adjustments removed audits for which the data were incomplete. A third adjustment excluded a very small number of audits, which concluded with extremely large payments that were not representative of 99.5 percent of tax returns examined in 2010 or 2017. The fourth adjustment was to limit taxpayers' payments to those made over the same number of months after either 2010 or 2017 (Table 2). In later sections of the paper, we consider the extent to which the third and fourth adjustments affected our estimates of the overall ROIs.

TABLE 2. Number of Tax Returns in Examination and Resulting Enforcement Revenue, Fiscal Years 2010 and 2017

Item	Tax Returns (thousands)		Enforcement Revenue ($ millions)[1]	
	2010	2017	2010	2017
Total[2]	3,206	1,827	80,566	37,191
Exclude returns with Earned Income Tax Credit	2,302	1,185	80,157	36,920
…and exclude if reported hours = 0	1,921	1,080	77,208	35,536
…and exclude "outliers"[3]	1,912	1,074	9,674	4,760
…and limit to cases where enforcement completed by March 31, 2012 (2019)[4]	1,333	918	4,434	3,930

[1] In nominal dollars
[2] For both years, total enforcement revenue through April 30, 2019. Hence, the amount of enforcement revenue received as a result of audits in 2010 covers over 8 years but only 30 months for the audits occurring in 2017.
[3] Outliers are audits resulting in enforcement revenue in the top 0.5 percent. The cutoffs were: 2010: Field: $1,500,000; Office: $42,000; Correspondence: $38,000; 2017: Field: $1,100,000; Office: $52,000; Correspondence: $34,000
[4] Limiting to cases where enforcement was completed by March 31, 2012 (2019) reduced the outlier cutoffs. The adjusted outliers are audits resulting in enforcement revenue in the top 0.5 percent. The cutoffs were: 2010: Field: $630,000; Office: $33,000; Correspondence: $23,000; 2017: Field: $980,000; Office: $49,000; Correspondence: $32,000
SOURCE: Enforcement Revenue Information System

Pre-refund audits. ERIS does not categorize the amount of savings resulting from a pre-refund audit as enforcement revenue. There are two barriers to inferring the amount of "protected revenue" resulting from a pre-refund audit. First, pre-refund audits are not flagged in the ERIS data and are grouped with other revenue protection projects. Second, the protected revenue cannot be distinguished from prepayments.

As a proxy, we excluded audits that had been prompted by an EITC-related issue. Three-quarters of EITC-related audits with W-2 earnings and 90 percent of those with self-employment income occur before refunds are paid out (Guyton, *et al.* (2019)). Excluding tax returns with EITC-related audits reduced the number in our sample by about 900,000 in 2010 and by over 600,000 in 2017. Almost all the reduction occurred in the correspondence audit category, causing the number of correspondence audits in our sample to fall by over 40 percent.

The impact of the exclusion of the EITC audits on the overall ROI is uncertain. Correspondence audits—as will be shown—cost less, on average, than office or field examinations. All other things equal, inclusion of the EITC audits in the analysis would cause the overall ROI to increase. However, EITC audits also probably yield less revenue, on average, than other audits, even among correspondence audits. According to the *IRS Data Book*, the average assessment for an EITC correspondence audit in 2017 was about $4,700 compared to an average of $6,000 for all correspondence audits of individual income tax returns. Whether the EITC exclusion caused the overall ROI to rise or fall depends on which of those two effects dominates.

Hours worked. After removing the EITC tax returns from the sample, there were about 400,000 returns in 2010 and 100,000 in 2017 for which no hours had been recorded. We excluded those returns from our analysis.

Outliers. In each year, a small number of examinations resulted in very large final payments. Those payments were large enough to skew the average return on investment to levels that vastly overstated the impact of nearly all other audits on enforcement revenue. For our main analysis, we removed the outliers, defined as returns above the 99.5th percentile in tax collected as a result of the IRS's enforcement for each type of audit. Although the number of taxpayers removed from the sample was very small, their removal caused enforcement revenues to decline by about 80 percent in each year, with nearly all the excluded revenue coming from field examinations of taxpayers filing corporate income tax returns.

Years. One key difference between our samples for 2010 and 2017 is that our data covered at least a decade for the former but just 30 months for the latter. For the portion of our analysis that compared 2010 and 2017 returns on investments, we limited the analysis to examinations and subsequent activities that had been closed by the end of March 2012 or March 2019. That exclusion eliminated an additional 600,000 returns from the 2010 sample and 160,000 from the 2017 sample.

Data for analysis. In combination, the first three restrictions reduce the sample from 3.2 million to 1.9 million audits in 2010 and from 1.8 million to 1.1 million audits in 2017. Constraining the years of enforcement activity further reduces our sample to 1.3 million audits in 2010 and 900,000 in 2017. The restrictions shift the composition of audits somewhat more in the direction of field and office audits.

Comparison of Average Returns on Investments in 2010 and 2017

We compare the average return on investment (ROI) for examinations occurring in 2010 and 2017, using the sample that retains only enforcement activities ending, for 2010 audits, by March 31, 2012 or, for 2017 audits, those ending by March 31, 2019. The estimates of the ROIs depend on the number of hours worked, labor costs, and collections.

Hours

Not surprisingly, the amount of time spent on correspondence examinations and follow-up activities is much lower than on office and, in particular, field examinations. For the 2010 examinations, the average hours worked per examined return ranged from 2 for correspondence audits to 45 for field audits (Table 3). Overall, over 95 percent of the hours working a case occurred during the examination period, with the remaining time largely split between appeals and collections.

TABLE 3. Average Hours in Enforcement Activities for Examinations Conducted in Fiscal Years 2010 and 2017

Type of Examination	Hours	
	2010	2017
Field	45	57
Office	10	11
Correspondence	2	2
Average for all types	15	19

NOTE: Reflects enforcement activities ending by March 31, 2012 (for 2010 examinations), or by March 31, 2019 (for 2017 examinations).
SOURCE: Enforcement Revenue Information System

Although the number of completed examinations dropped by 31 percent for the 2017 sample relative to 2010, the total number of hours devoted to examinations fell by only 15 percent, largely because the average hours spent completing field audits rose from 45 for audits in 2010 to 57 in 2017. In contrast, the average time spent on correspondence and office audits changed very little during this period—remaining, on average, at about 2 and 10 to 11 hours, respectively.

Another difference across type of examinations was in the pay grade levels of examiners assigned to a case. The highest-paid examiners—typically at the general schedule (GS) grades ranging from 12 to 14—were responsible for most of the hours worked on field examinations. For office and correspondence audits, nearly all the hours were attributed to IRS personnel at lower grades: GS grades 7 through 11 for office audits and 5 through 8 for correspondence audits. Between 2010 and 2017, there was a reduction in the share of hours spent

by grade 7 employees on office audits that was offset by an increase in the share done by grade 8 employees. For correspondence audits, though, the share of hours worked by grade 4 employees increased from 1 percent to 13 percent over the period, whereas the share of hours spent by workers in grades 5 through 7 declined.

Costs

For examinations in place in 2010, total labor costs equaled $1.2 billion. Of that amount, 89 percent was attributable to field examinations, with the remainder nearly evenly split between office and correspondence audits. Labor costs remained at about $1.2 billion for the examinations occurring in 2017, with the split by type of audit remaining about the same as in 2010.

Though there was wide variation in the cost per tax return by type of audit, the gaps in the hourly costs were much narrower. In 2010, the cost per return ranged from $78 for a correspondence audit to $2,861 for a field audit (Table 4). But the hourly cost was $39 for correspondence audits and $63 for field audits. Overall, average costs rose from 2010 to 2017 by 40 percent per return and by 16 percent by hour, with the greatest increase in the cost per return of a field audit (from $2,861 to $4,148).

TABLE 4. Average Costs Per Tax Return and Per Hour for Enforcement Activities Related to Examinations Conducted in 2010 and 2017

Type of Examination	Per Return		Per Hour	
	2010	2017	2010	2017
Field	2,861	4,148	63	72
Office	442	552	46	52
Correspondence	78	97	39	43
Average for all types	913	1,278	60	69

NOTES: In nominal dollars. Reflects enforcement activities ending by March 31, 2012 (for 2010 examinations), or by March 31, 2019 (for 2017 examinations).
SOURCE: Enforcement Revenue Information System, Office of Personnel Management, and Congressional Budget Office

Collections

Additional tax payments totaled $4.4 billion as a result of the 2010 audits and dropped to $4 billion from the 2017 audits. In 2010, the average collection per return for field audits was 10 times the comparable amount for correspondence audits; that gap narrowed for 2017 audits, with the average for field audits equal to 8 times that of correspondence audits (Table 5). In contrast, the average collection per hour for correspondence audits was about double the amount for both office and field audits in 2010 and was about triple the amount for the 2017 audits.

TABLE 5. Average Enforcement Revenue Per Tax Return and Per Hour for Enforcement Activities Related to Examinations Conducted in 2010 and 2017

Type of Examination	Per Return		Per Hour	
	2010	2017	2010	2017
Field	9,019	11,186	199	195
Office	2,036	2,551	211	239
Correspondence	867	1,440	435	636
Average for all types	3,327	4,284	218	230

NOTES: In nominal dollars. Reflects enforcement activities ending by March 31, 2012 (for 2010 examinations), or by March 31, 2019 (for 2017 examinations).
SOURCE: Enforcement Revenue Information System, Office of Personnel Management, and Congressional Budget Office

Average Return on Investment

Overall, the ROI was $3.60 for a dollar of appropriations for the 2010 audits and $3.40 for the 2017 audits (Table 6). Given that the number of hours and the hourly labor cost were much lower for correspondence audits than for the more intensive types of examinations, it is not surprising that the average ROI for the former was much higher than the overall levels: $11.10 for a dollar of appropriations for the 2010 examinations, rising to $14.90 for the 2017 audits. That growth was generated by the rise in average collections and increased reliance on lower-grade staff, with average hours remaining about the same. Largely because of the increase in the average hours worked, the ROI for field examinations fell from $3.20 for a dollar of appropriations to $2.70. On net, the increase in the average costs of field examinations outweighed the higher net returns from correspondence audits, causing the overall ROI to be lower for the 2017 audits compared to those conducted in 2010.

TABLE 6. Average Dollars of Return on $1 of Investment for Enforcement Activities Related to Examinations Conducted in Fiscal Years 2010 and 2017

Type of Examination	2010	2017
Field	3.2	2.7
Office	4.6	4.6
Correspondence	11.1	14.9
Average for all types	3.6	3.4

NOTE: Reflects enforcement activities ending by March 31, 2012 (for 2010 examinations), or by March 31, 2019 (for 2017 examinations).
SOURCE: Enforcement Revenue Information System

Average Return on Investment in 2010 through April 2019

For audits that occurred in 2010, we can estimate the average return on investment for a period spanning nearly a decade. We expand the sample to include examinations and collections that were still ongoing after March 31, 2012. For this analysis, we again exclude returns claiming the EITC, with zero hours reported, or in the top 0.5 percent of enforcement revenues. Extending to the longer time span increases the number of returns in the sample by 600,000, also causing the thresholds for outliers to rise (especially for field audits, where the threshold more than doubles). With the longer-time span, the average return on investment increases slightly from $3.60 (Table 6) for a dollar of appropriations to $3.70 (Table 7).

TABLE 7. Average Dollars of Return on $1 of Investment for Enforcement Activities Related to Examinations Conducted in Fiscal Year 2010, Including Collections Through April 2019

Type of Examination	2010
Field	3.2
Office	5.3
Correspondence	12.0
Average for all types	3.7

SOURCE: Enforcement Revenue Information System

Outliers

As noted earlier, we excluded 0.5 percent of returns that resulted in the largest amounts of collections. For returns examined in 2010, the collection thresholds for the period through April 2019 were $38,000 for correspondence audits, $42,000 for office audits, and $1,500,000 for field audits, resulting in nearly 10,000 returns being dropped from our analysis.

We excluded the outliers for two reasons. First, they skewed the ROI substantially upwards and were not reflective of the bulk of examinations conducted by the IRS. We also were concerned about the reliability of some of the data in the top 0.5 percent. For example, over half of the outlier cases in 2010 were examinations conducted solely through correspondence. Even though they resulted in only 4 percent of total tax collections from the outlier cases through April 2019, the average tax collection per correspondence examination was over $450,000, which seems remarkably high for audits conducted solely through the mail.

The findings on the other types of audits are more consistent with our expectations. Nearly all audits of C corporations are conducted in the field. Although C corporations represented less than a quarter of all outlier audits, about 84 percent of the tax collections attributed to the outliers' examinations were from audits of C corporations. And although the outlier C corporations represented only about 3 percent of C corporations audited in 2010, nearly all the tax collections attributable to examinations of C corporations came from the outlier cases.

For all outlier cases (including the correspondence audits), the average costs and average collections over that period were much larger than for the typical returns. On average, per return, outlier examinations cost about $54,000 and yielded over $7 million per return—resulting in an average ROI of $129 for a dollar of appropriations. For corporate outliers, the ROI was $128; in contrast, the ROI for the non-outlier corporations was less than $2.

Inclusion of the outliers would have boosted the overall ROI from $3.70 to $25. In future research, we will look at the characteristics of the outlier audits in greater depth and, in particular, whether those findings are repeated in other years.

Revenue Effect of Restoring 2010 IRS Enforcement Budget

In 2019, the IRS received an appropriation for enforcement of $4.9 billion—$1.8 billion less than its appropriation for this account in 2010 (2019 dollars). Restoring that funding would increase revenues, but the effect would not be immediate because the IRS would have to hire and train new employees. Moreover, the magnitude of the revenue effect would depend on how those additional funds were used.

Thus far in this paper, we have estimated average ROIs for examinations that occurred in 2010 and 2017. When estimating the revenue savings from a new initiative, however, the appropriate measure is not the average ROI but the marginal ROI—that is, the additional amount of revenues received from an additional dollar of funding (Holtzblatt and McGuire (2016)). We made several adjustments to our estimates of average ROIs to move them closer to being marginal measures:

- We assume that it would take 3 years for the IRS to hire and train new examiners. As a result, the IRS enforcement activities do not reach full potential until the third year after the funding is restored.
- Over time, however, taxpayers begin to identify the types of issues that are more likely to be targeted for the additional new audits. Our analysis accounts for taxpayers adjusting their evasion methods to avoid detection and the resulting decline in ROIs.
- That effect is somewhat offset because we anticipate that the IRS would revise its detection algorithms in response to taxpayers' adoption of new forms of noncompliance.
- With additional increments of funding, the IRS would select more difficult cases to audit.

For this analysis, we started with the average ROIs for audits conducted in 2010 and included all costs and collections through April 2019. Unlike the ROIs shown earlier, interest payments and penalties were added to the taxes collected, when estimating the budgetary impact of an increase in appropriations. We did not include the outlier cases in the ROIs.

Under our options, funding would be used entirely to initiate new examinations; that is, none of the funding would be used to intensify or prolong examinations begun before funding was increased. Moreover, we assume that the existing infrastructure—buildings, computers, managers—would be sufficient to support the new hires. Thus, the funding would go entirely to hiring and training staff to conduct the new audits.

We estimate the impact of the increased funding for two options. In both cases, funding would be increased in three increments, roughly rising by an additional $600 million a year over a 3-year period. After 3 years, the total additional appropriations would reach $2 billion and would remain at that level, with adjustments for inflation.

Under the first option, funding would be allocated across the new audits in the same proportions as was the case for new audits in 2010 (27 percent for field, 15 percent for office, and 58 percent for correspondence). For the initial increase of funding of $620 million, the estimated return on investment in the first year would be $1.40 for an additional $1 of appropriations and would rise to a peak of $5.70 in the third year when new employees were hired and fully trained. The ROI, however, on the level and types of audits financed by the first installment of funding would fall to $4.60 by the end of the decade as taxpayers shifted to less detectible forms of tax evasion.

Over the next 2 years, appropriations would rise in increments: by an additional $640 million in 2021 (on top of the added funding in 2020) and by another $665 million in 2022 (on top of the added funding in 2020 and 2021). With each increment, the IRS would select returns with increasingly difficult issues. As a result, the ROI on cases selected with each additional increment of funding would fall—at full implementation, from the peak of $5.70 for the audits funded by the first installment to a peak of $5.40 for the additional audits funded by the 2021 increment and to a maximum of $5.00 for the 2022 increment.

Over a 10-year period, the option would increase appropriations by $20 billion, causing revenues to rise by $65 billion. On net, the option would reduce the deficit by $45 billion.

Under the second option, the additional appropriations would fund new field audits only. After 3 years, the ROI would peak at $5.00 for an additional dollar of appropriations. The option would increase revenues by $57 billion over 10 years. With increased appropriations still totaling $20 billion, the net reduction in the deficit would be $37 billion. Still, the ROI estimates for the outlier cases suggests much larger net savings if some of the new field audits were targeted at very large corporations.

Conclusions

Both the decline in the IRS's funding and audit rates have been widely publicized (Kiel and Eisinger (2018)). Less known has been the impact of those reductions on the IRS's efficiency. We find the average ROI for enforcement activities fell by 6 percent between 2010 and 2017. It is likely that decline was related to the decrease in funding, but other factors—such as an unrelated change in compliance behavior or IRS detection algorithms—might have contributed to the reduction in the average ROI. That overall effect, however, masks differences across types of enforcement actions. The ROI associated with field examinations declined by 16 percent, largely because of an increase in the number of hours worked on each case. That increase could reflect the departure of experienced revenue agents and officers. In contrast, the ROI for correspondence audits increased by 34 percent, in part fueled by a greater reliance on lower-grade employees.

Although the impact of outliers is difficult to interpret, the analysis suggests that the ROI for auditing some corporations is very high. Given that those audits and follow-up actions probably extend over many years, we do not yet have enough data to evaluate the impact of the recent cutbacks in appropriations on the ROI associated with audits of large corporations. However, the halving of the audit rate of those very large corporations probably has resulted in sizable reductions in enforcement revenue.

Future research should extend the analysis to other years as well as to other types of IRS enforcement activities. One key extension would include pre-refund exams in the analysis. Another would expand the focus to other types of IRS enforcement-related activities. In particular, the IRS has more flexibility in dealing with returns where "mathematical and clerical" errors are detected during return processing. Whereas math error procedures (the common shorthand used to refer to that process) were originally limited to inconsistencies in the returns, the scope has broadened since the mid-1990s to apply to more compliance-related items (such as provision of invalid social security numbers when claiming certain child-related tax benefits). Understanding the costs and revenue savings from those extensions would provide more insight into the effectiveness of different types of IRS enforcement actions.

Finally, this study does not reflect two very recent changes to the IRS's enforcement procedures. First, PATH allows the IRS to conduct audits at the partnership level, beginning after December 31, 2017. And in May 2019, the IRS announced a new large corporate compliance program (LCC) that will employ automated techniques to identify large corporations and then data analytics to detect the returns with the highest compliance risk. Both changes have the potential to increase the IRS's efficiency in the long term.

References

Congressional Budget Office. 2012. *Comparing the Compensation of Federal and Private-Sector Employees*. Washington, DC.

Congressional Budget Office. 2017. *Comparing the Compensation of Federal and Private-Sector Employees, 2011 to 2015*. Washington, DC.

Consolidated Appropriations Act, 2017, P.L. 115–31, 131 Stat. 135.

Falk, Justin. 2012. *Comparing Benefits and Total Compensation in the Federal Government and the Private Sector*. Congressional Budget Office Working Paper 2012–04.

Government Accountability Office. 2014a. *IRS Correspondence Audits: Better Management Could Improve Tax Compliance and Reduce Taxpayer Burden*. GAO-14–479.

Government Accountability Office. 2014b. *Tax Filing Season: 2014 Performance Highlights the Need To Better Manage Taxpayer Services and Future Risks*. GAO-15–163.

Government Accountability Office. 2016. *Information Technology: Federal Agencies Need To Address Aging Legacy Systems*. GAO-16–468.

Government Accountability Office. 2018. *Tax Fraud and Noncompliance: IRS Can Strengthen Pre-refund Verification and Explore More Uses*. GAO-18–224.

Guyton, John, Kara Leibel, Dayanand Manoli, Ankur Patel, Mark Payne, and Brenda Schafer. 2019. *The Effects of EITC Correspondence Audits on Low-Income Earners*. National Bureau of Economic Research Working Paper 24465.

Hodge, R.H., A.H. Plumley, K. Richison, G. Yismaw, N. Misek, M. Olson, and H.S. Wijesinghe. 2016. "Estimating Marginal Revenue/Cost Curves for Correspondence Audits." Internal Revenue Service, *IRS Research Bulletin*, Publication 1500.

Holtzblatt, Janet, and Jamie McGuire. 2016. *Factors Affecting Revenue Estimates of Tax Compliance Proposals*. Congressional Budget Office Working Paper 2016–05.

Internal Revenue Service. Various years. *Data Book*. Publication 55B.

Kiel, Paul, and Jesse Eisinger. 2018. "How the IRS Was Gutted." *ProPublica*. December 11, 2018.

Koskinen, John A., Commissioner, Internal Revenue Service. 2015. Written testimony before the Senate Appropriations Committee Subcommittee on Financial Services and General Government. https://www.irs.gov/newsroom/written-testimony-of-john-a-koskinen-before-the-senate-appropriations-committee-subcommittee-on-financial-services-and-general-government.

Treasury Inspector General for Tax Administration. 2015. Automated Underreporter Program Tax Assessments Have Increased Significantly: However, Accuracy-Related Program Tax Assessments Were Not Always Assessed When Warranted. 2015-30–037.

Treasury Inspector General for Tax Administration. 2018. *Despite Spending Nearly $380 Million, the Internal Revenue Service Is Still Not Prepared To Enforce Compliance With the Foreign Account Tax Compliance Act*, 2018-30–040.

U.S. Department of Justice, 2019. FY 2020 Congressional Budget Submission for Tax Division. https://www.justice.gov/doj/fy-2020-congressional-budget-submission.

U.S. Department of the Treasury. 2013. FY 2014 Congressional Justification of Appropriations. https://home.treasury.gov/system/files/266/10.-IRS-CJ-FINAL-v2-FY2014.pdf.

U.S. Department of the Treasury. 2019. FY 2020 Congressional Justification of Appropriations. https://home.treasury.gov/system/files/266/02.-IRS-FY-2020-CJ.pdf.

3

Improving the Digital Taxpayer Experience

Gay

ten Brink ♦ Scollan

Kerber ♦ Papa ♦ Sauser

IRS Online Account User Testing: Improving the User Experience Through Iterative Design and Research

Heather Gay (Mediabarn, Inc.)

Mediabarn, Inc., on behalf of the Internal Revenue Service (IRS), has conducted over 15 rounds of iterative design and user testing with taxpayers for "online account." In talking with nearly 200 research respondents, we uncovered an assortment of findings related to both taxpayers' expectations for accessing their tax information online and their perspectives on interacting with the IRS in general. This paper shares some of our key discoveries and how these findings can inform IRS decisions to optimize the online account application to do a better job of addressing user needs.

What Is Online Account?

Online account was conceived by the IRS in 2014. For taxpayers who must interact with the IRS (e.g., balance due, notice, or refund amount), online account provides a self-service, one-stop shop for personalized tax assistance. Unlike calling, mailing, faxing, or visiting the IRS, online account provides a quick, easy, and secure on-demand service.

There were several business priorities established in 2014 and reaffirmed in 2018 that drove the IRS's decision to create and build the online account application. Their goals were to:

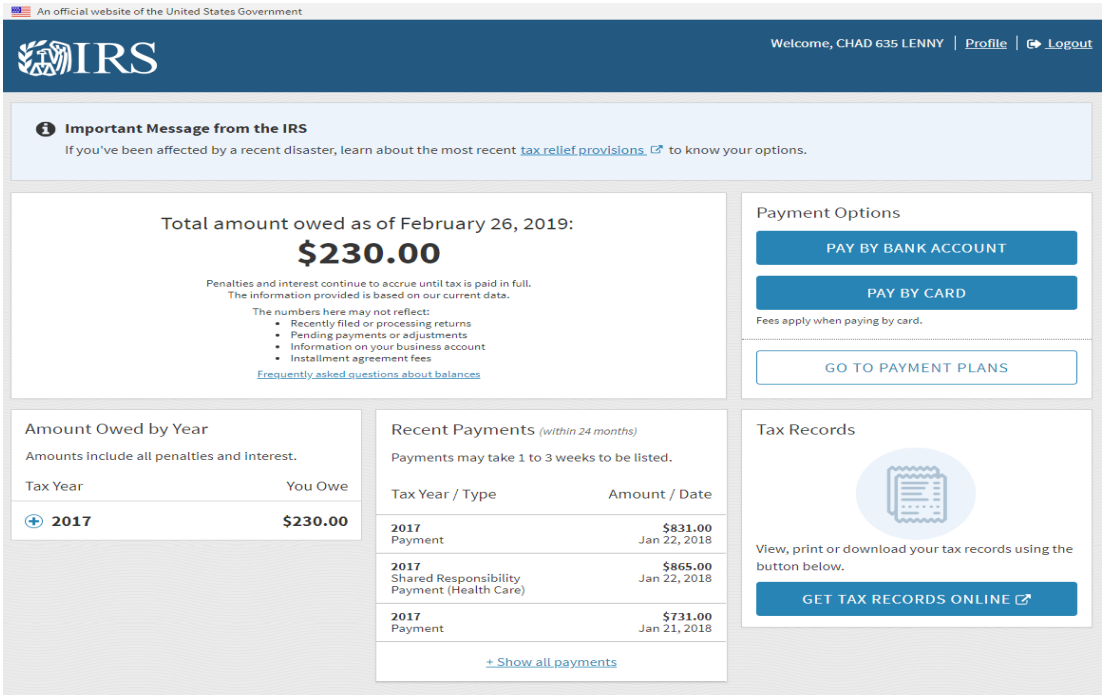
- Make it easy to use so that it is spontaneously used by taxpayers and they can help themselves;
- Reduce the number of phone calls to the IRS and save taxpayers money;
- Increase public trust by providing taxpayers with secure access to their own tax data; and
- Improve voluntary compliance.

Online account was launched as an application on IRS.gov in the fall of 2016 with four key features:

- *Balance Due*: allows taxpayers to see if they owe a balance to the IRS, with the ability to see details by tax year;
- *See Payments*: shows a list of recent payments made by the taxpayer to the IRS;
- *Make A Payment*: allows taxpayers the ability to pay their tax debt or other payments through either their bank account or credit/debit card and gives taxpayers the option to setup an online payment agreement; all of these currently hand off to other systems, and
- *Tax Records*: allows taxpayers to access transcripts of their previously filed returns.

Figure 1 is a screen shot of the landing page for online account as it was in production (live on the IRS.gov Website) in the spring of 2019.

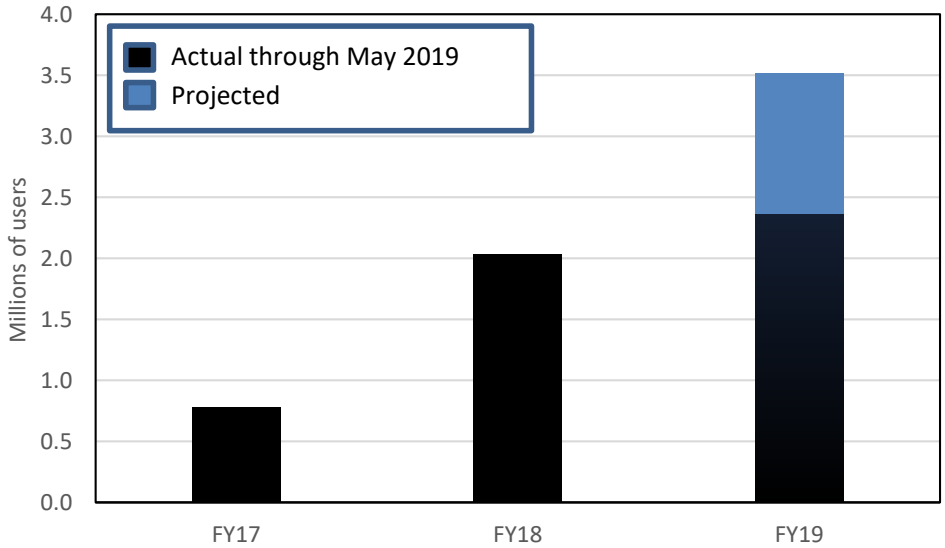
FIGURE 1. IRS Online Account Landing Page, Spring 2019



NOTE: Chad 365 Lenny is a fictitious name.

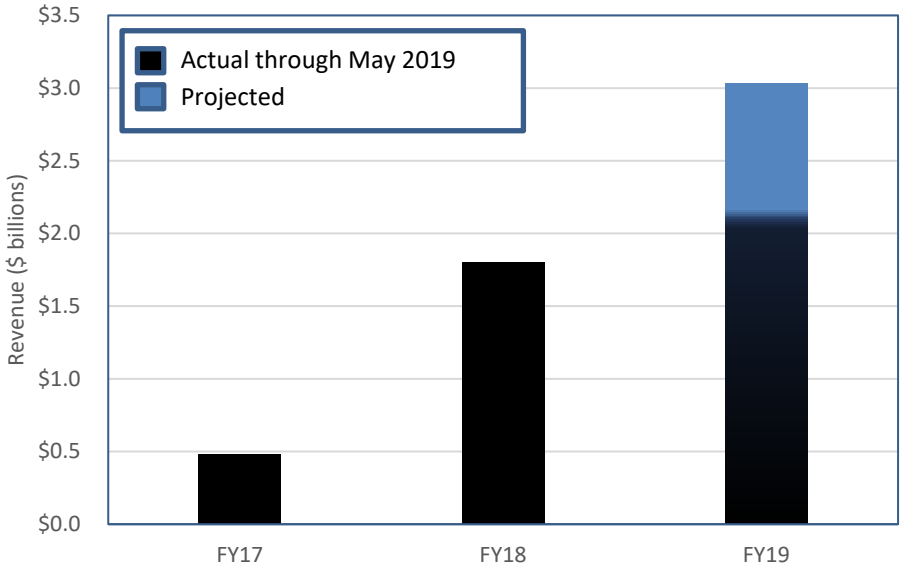
Figure 2 shows that, since online account launched in November 2016, there have been over 5 million unique users tracked.

FIGURE 2. Number of Online Account Unique Users, Fiscal Years 2017, 2018, and 2019 (through May 2019)



Clicks from online account to IRS Direct Pay (i.e., "Pay by Bank Account") resulted in a transactional value of $4.3 billion from the day it was launched in November 2016 through May 2019 (see Figure 3). In addition, almost 198,000 installment agreements have been established by online account users.

FIGURE 3. Online Account Revenue Generated Through IRS Direct Pay, Fiscal Years 2017, 2018, and 2019 (through May 2019)



Overview of User Testing

Methodology

User testing is a form of qualitative research. Qualitative research is defined by small sample sizes, as opposed to quantitative research that often includes statistically significant sample sizes. Qualitative research focuses on the quality of the interactions, depth of the responses, and user behaviors rather than on statistical measurements.

Typically, we, the design team, identify user problems and develop hypotheses to provide solutions. These help to define the information included within the design prototype. In many cases, we test alternate designs to evaluate which is best at meeting user needs. We also develop validation metrics within the context of the Lab test; this adds an aspect of quantitative measurement to the test, which helps us to determine how well the design solutions address taxpayer issues.

One round normally includes a sample of 8 to 12 representative respondents who participate in guided and moderated one-on-one sessions. These sessions are usually scheduled over a 1- to 2-day period. Respondents are asked a series of questions that identify whether they qualify to participate in the study; only those who qualify are invited to participate. Depending on the focus of a specific round of testing, we ask for information such as how they file their tax returns, whether they have received a notice from the IRS regarding a balance due, or whether they have set up an installment agreement with the IRS.

In a user experience test session, respondents typically interact with a design prototype of online account and are asked to accomplish specific tasks. The moderator observes their ability to complete each task successfully. In addition, the moderator asks the respondents indepth questions to gain actionable insights into their thoughts, perceptions, and reactions while interacting with both the online account application and the IRS.

The sessions we conducted were always completed with simulated information within a prototype. In our testing sessions, taxpayers never interacted with their own information on the live Website. (The Wage and Investment Strategy and Solutions Division at IRS conducted research sessions with taxpayers and their individual data; we collaborated with this group to get a full-lifecycle picture of how taxpayers interacted with online account, pre- and post-launch.)

Once the sessions have been completed, the research and design teams work with stakeholders to synthesize the results and iterate for future tests. In this way, findings from frequent rounds of testing inform the questions and design hypotheses for future rounds.

What We've Tested So Far

Through 17 rounds of research sessions, we tested a variety of designs and evaluated many common user tasks, including:

- Online account page layout—both a modular/widget view and a transactional view;
- Payment option flows—nonmodal pages/wizards to include IRS Direct Pay, Pay1040.com, and a path to payment plans;
- Future balance-due calculator;
- Overview by Tax Year/Amount Owed by Year table;
- Recent Payments table;
- Tax Records/Get Transcript module;
- Frequently Asked Questions link—supported by the Taxpayer Advocate Service;
- Expanding and collapsing tables/widgets;
- Desktop and mobile layouts; and
- Text/language of buttons, links, and educational copy.

Simple, Impactful Improvements

Some examples of simple yet impactful improvements that we've made to online account as a direct result of user-centered research and design can be found in the following changes to the language used:

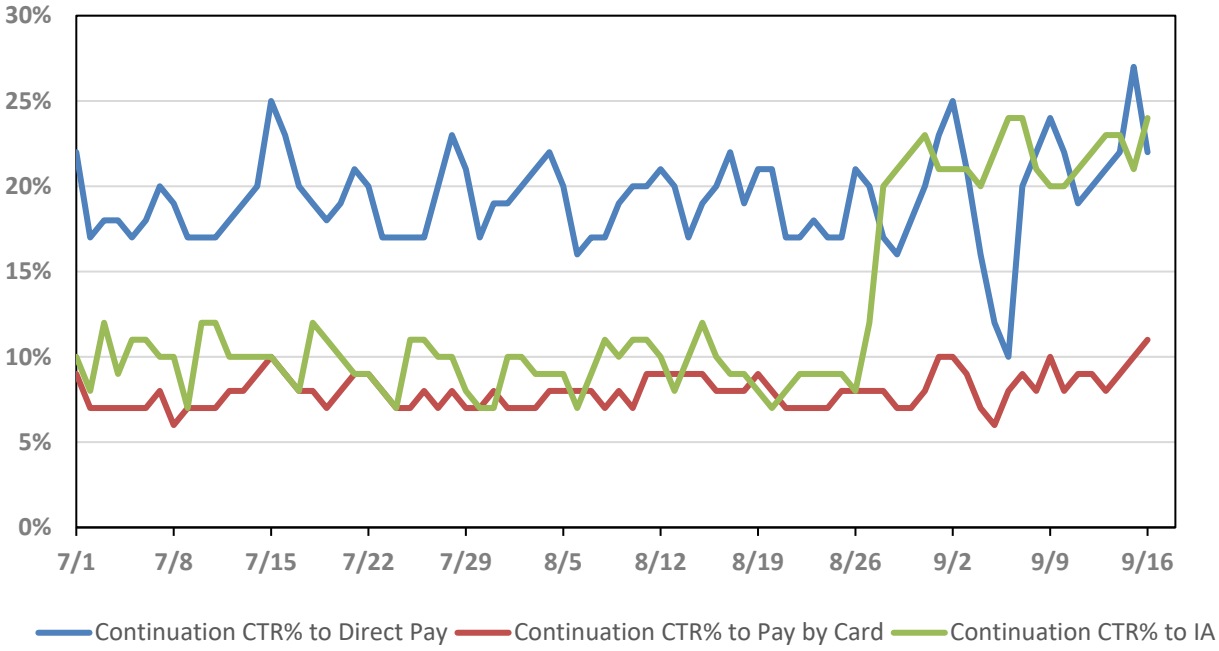
a) The wording for the *Get Transcript Online* button required taxpayers first to consider what is a "transcript." For many, their first thought was a school grade report, which did not correspond to taxes or the IRS. In reality, some taxpayers needed access to a transcript of their previously filed tax return, usually to apply for a loan. This button was relabeled as *Get Tax Records Online,* which aligns more with the way in which they thought about the information.

b) The *Recent Payments* table showed the few most recent payments a taxpayer had made to the IRS and could be expanded to show additional payments made within the past 24 months. This table could be expanded and collapsed. In earlier rounds of testing, the link to collapse the table was labeled *Minimize Payments*. Who would not want to minimize payments owed to the IRS? To avoid any misunderstanding, it was relabeled as *Show Fewer Payments*.

c) Call-to-action buttons generally strive to include the action the user can take when selecting a button. The button labeled *Need More Time to Pay?* did not clearly indicate where it would take the user. This was relabeled as *Go To Payment Plans*, which more clearly indicated that the user would next see information about setting up an installment agreement.

Figure 4 shows the effect of changing the wording of the *Payment Plans* button. Based on user testing, we uncovered that changing the wording from *Need More Time to Pay?* to *Go To Payment Plans* increased online account sessions continuing on to an installment agreement by over 100 percent (from 10 percent to 22 percent) overnight—without cannibalizing click-throughs to *Direct Pay* (bank account) or *Pay by Card*.

FIGURE 4. IRS Online Account Payment Options Daily Continuation CTR Rate, July 1, 2017 to September 16, 2017



NOTES: CTR stands for click-through rate. Label change relaunched August 26, 2017. Continuation CTR to Direct Pay: 15 percent. Continuation CTR to Pay by Card: 7 percent. Continuation CTR to Installment Agreement (IA): 10 percent.

What We've Learned—Key Themes

In addition to these very specific enhancements that user testing uncovered and allowed us to improve, there are some overarching themes that emerged from our time spent talking with nearly 200 taxpayers. The primary themes include:

- Taxpayers regard the IRS as a financial institution;
- There is an unmet and increasing taxpayer need for both digital communications and solutions from the IRS;
- Taxpayers want to see more of a connection between their balance due and the payments they have made;
- Taxpayers want to see all payment options before deciding how to pay;
- When faced with owing a balance, taxpayers first consider how much to pay, then when to pay, followed by the form of payment, in that order; and
- Taxpayers expect online account to be one, integrated system.

Uncovering these themes will likely take online account in new directions. Having this knowledge should help to bridge the gap between the mental model that taxpayers have and the way in which information is presented within online account.

We discovered after conducting the first few rounds of research interviews that taxpayers unmistakably view their relationship with the IRS as a financial one. Therefore, although the IRS is not a lending institution, taxpayers often expect that it will conform to their prior experiences with viewing their other financial information online. This is important because it has far-reaching implications for the design of digital products, such as online account. It can also impact voluntary tax compliance.

Taxpayers' expectations for interacting with the IRS online are similar to other financial transactions, such as a bank accounts, credit card statements, auto loans, student loans, mortgages, etc. These expectations include access to information about the balances they owe, which is a combination of their initial balance, the credited payments they have made, accrued interest and penalties, and the current balance due. This is important because it could impact the way in which information is displayed within online account. Providing a way to connect monies paid more closely with balances owed may help taxpayers feel a greater sense of satisfaction while they're "chipping away" at their balance.

Since many taxpayers are familiar with making payments against a loan or other debt, many understand the time value of money. Also, their payments to the IRS are not made in a vacuum separate from their other budget items. Taxpayers have expressed a need to understand clearly the advantages and disadvantages of paying a portion or all of their balance due and paying now versus paying over time, in addition to choosing a form of payment. They would like to see how their total payment will differ based on the choices they make.

Currently, online account acts as a portal for making a payment, allowing taxpayers to begin the process of making a payment, but then sending users out to other applications (such as online payment agreement, IRS Direct Pay, or third-party credit card processors) to complete the transaction. The transition between online account and IRS payment systems is not as seamless as it could be, and the experience from a user perspective may cause distrust. This can be frustrating, confusing, or time-consuming for taxpayers, all of which are counter to the purpose and vision of online account.

The Future of Online Account

The ultimate goal for online account is to offer individualized, personalized information. Through user testing as well as other research, the IRS has identified a host of taxpayer needs and expectations for which online account could provide a solution.

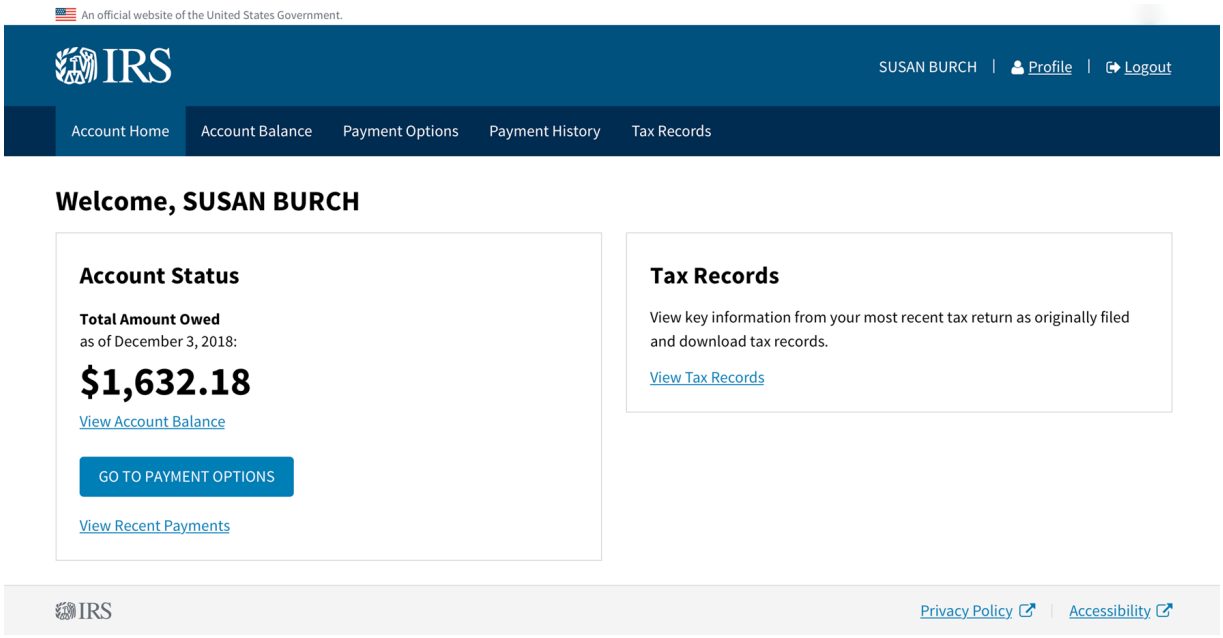
As the IRS contemplated the addition of new features and functionality, it became clear that it was necessary first to evolve the design of the one-page application to something more robust that could accommodate these extras. The IRS made the decision to alter the architecture of online account and grant the application the capacity to add on new features yet still offer an uncluttered visual design.

Future features will likely include greater integration of installment agreement and payment application programming interfaces (APIs), the ability to sort and filter payment history, and the ability for the IRS to send digital notices to taxpayers.

After completing five rounds of design iterations and user testing sessions for this new architecture, the IRS is working to update online account to add future features. Taking the steps to validate the new designs allowed the IRS to build it with confidence. This updated application launched in July 2019.

Figure 5 is a screen shot of the new landing page for online account as it was launched in July 2019. The tabbed navigational structure provides the flexibility to add in new features and functionality without providing too much information at one time.

FIGURE 5. IRS Online Account Landing Page, July 2019



NOTE: Susan Burch is a fictitious persona used in the testing materials.

Usability of Biometric Authentication Methods for Citizens with Disabilities

Ronna ten Brink and Rebecca Scollan (The MITRE Corporation)[1]

1. Introduction

Today, 27.2 percent of people living in the United States (U.S.) experience a disability, which is defined as a functional limitation that affects one or more major life activities (Taylor (2018); (National Center on Disability and Journalism (2018)). Approximately 17.6 percent of those who report having a disability describe it as a severe disability. As we age, our likelihood of having a disability increases. The current percentage of the population with a disability is assumed to be a low assessment because census data are collected from those who live in households. It is not collected from those who live in nursing or assisted living facilities, the large majority of whom have a disability (Taylor (2018)). Generally, people are living longer both in the U.S. and across the globe. From 2015 to 2030, it is estimated that the elderly population will grow from 9 percent to 12 percent of the global population (Roberts, *et al.* (2018)). As our aging population grows larger, the number of adults with a disability will grow as well.

Single-factor authentication with a username and password has long been known to be vulnerable to cyberattacks, both social engineering and brute-force, as well as a usability challenge due to contradictory advice and the cognitive burden of managing many complex passwords (Bonneau, *et al.* (2014)). A smartphone allows for greater use of more convenient authentication methods. Smartphone ownership increased 42 percent from 2011 to 2018 (Pew Research Center (2018)), and 77 percent of adults living in the U.S. now own smartphones. Widespread smartphone use has made two-factor and multifactor authentication more prevalent (Bonneau, *et al.* (2014)). Two-factor authentication combines information that someone knows, such as a password, with something that they own, like a smartphone. Multifactor authentication provides an additional security factor unique to the subject, typically a physical or behavioral biometric (Aleksandr, *et al.* (2018)), that can be input via a smartphone. The use of multifactor authentication will likely continue to grow within the U.S. as e-commerce adopts recommendations from the National Cybersecurity Center of Excellence (NCCoE) (Newhouse, *et al.* (2018)) at the National Institute of Standards and Technology (NIST) to use multifactor authentication online to reduce the growing problem of online-purchase fraud.

We investigated the usability of biometric authentication schemes for users with and without disabilities. We comparatively evaluated three biometric authentication schemes (fingerprint, eye, and palm recognition) and one nonbiometric authentication scheme (a personal identification number or PIN) on effectiveness, efficiency, and perceived usability. This research contributes to the development of a standardized Methodology to evaluate the usability and accessibility of authentication technologies intended for use with public Government services. Our initial focus is a comparative usability study on biometric authentications and PIN; these methods were chosen for their current popularity and potential use in the future. We worked with the HYPR Corporation, who provided a Fast IDentity Online (FIDO) Universal Authentication Framework (UAF) client for Android and iOS devices. HYPR offers an inherently multifactor, decentralized authentication solution designed to eliminate passwords and shared secrets as a means for authenticating users and to provide a more secure means that is easier to use. Using a working demonstration application provided by HYPR, we conducted our usability study on a range of popular biometric schemes.

We chose to work with two large populations of adults with disabilities: those who are low vision or blind, and those who are hard of hearing or deaf. In Taylor's Census Report on estimates of disability prevalence

The authors wish to thank Katja A. Sednew, Michelle R. Schumaker, Melanie Shere, Jared M. Batterman, Kristen M. Klein, and Erika L. Darling for their support in this work.

based on the Social Security Administration (SSA) Supplement to the 2014 Survey of Income and Program Participation, 12.3 million U.S. adults over the age of 18 had serious difficulty seeing, of which 1.6 million are legally blind (Taylor (2018). Approximately 17.1 million adults reported a serious hearing difficulty; of these individuals, 3.4 million were deaf. We selected these two populations due to their large size, as well as, practical and logistical considerations due to time constraints, the research team's familiarity with both populations, and the assistive technologies used. Ultimately, 30 individuals were recruited: 10 participants who reported having hearing loss, 10 who reported having low vision or who were legally blind, and 10 who reported no disability.

This research contributes to a better understanding of the user experience of smartphone-based biometric authenticators and the eventual increased usability and accessibility of online Government services, leading to higher adoption and wider access to these services. Our results can also be generalized to any secured web services, e.g., banking and healthcare services.

1.1. Background

A growing community of people living with one or more disabilities creates a challenge to Federal agencies looking to digitize more personalized services. Government services received low customer satisfaction scores (Comer (2018)) for their websites and customer services. Despite this challenge, there is recognition that services must be modernized, personalized, and moved to online channels to reduce costs and improve citizen services (U.S. Congress (2018, March 22); Konkel, (2018, April 4)). For example, the President's Management Agenda (PMA) CAP goal 4 aims to "provide a modern, streamlined, and responsive customer experience across Government, comparable to leading private-sector organizations" and "improv[e] the experience citizens and businesses have with Federal services whether online, in-person, or via phone" (OMB (n.d.)) The 21st Century Integrated Digital Experience Act (21st Century IDEA), passed in December 2018, sets a "minimum accessibility, searchability and security standards for all new and existing Government websites, and requires agencies to adopt web analytics tools to constantly improve sites' functionality. Organizations would also need to make all sites mobile-friendly and comply with website standards set by the General Services Administration" (Corrigan (2018)). As Federal agencies work towards meeting these challenges, providing services that are both usable and secure is tantamount, and the designs of identity proofing and authentication address critical early touchpoints for users.

Federal agencies' digital services face unique usability challenges. Registering for an online service with a Federal agency might be a citizen's first interaction with that agency. Such services might be used only once in a lifetime or be accessed infrequently. The audience for these services is often diverse, spanning all ages, incomes, geographies, and abilities. Additionally, key services may include access to one's own personally identifiable information (PII), implying significant risk to both the institution and to the user. However, moving such services online offers Federal agencies the great benefits of increased citizen satisfaction and reduced costs. Federal agencies typically have no competition and are the only place to contact when citizens have questions or problems. An average business cost for a call-center call is $5.50 versus an online services' cost of $0.10 by serving those who find answers or resolutions online (Cardello and Farrell (2017)). Some agencies face even higher call-center costs; the average call to the IRS costs $41 (Konkel (2018, February 9)). Agencies must comply with Section 508 and the new IDEA Act mandates on accessibility when designing and implementing digital services. Section 508, an amendment made to the 1973 Rehabilitation Act in 1998, mandates that Federal agencies provide accessible electronic content and technologies (GSA (2018, May)). The 21st Century IDEA Act mandates that Fenderal Agencies comply with Section 508 for all hardware, software, and documentation (S.3050—21st Century IDEA (n.d.)).

The 2017 update to the NIST Special Publication (SP) 800-63-31, *Digital Identity Guidelines,* which includes SP 800-63B, *Authentication and Lifecycle Management,* now requires two-factor authentication: either a multifactor authenticator or a combination of two, single-factor authenticators to achieve Authentication Assurance Level 2 (Grassi, *et al.* (2017)). Many Federal agencies' online services meet the criteria for Authentication Assurance Level 2. Biometrics are growing in popularity (Kessem (2018)) and may be used in a multifactor authentication design. NIST defines biometrics as both physical and behavioral characteristics, including them as factors provided they are part of multifactor authentication that includes a physical authenticator (with a

device like a smartphone meeting security requirements of proving "something you have," and the biometric, "something you are").

But are biometrics captured by smartphones usable and accessible to all citizens? While widespread smartphone ownership has made biometrics more available (German and Barber (2018)), there is little evidence (Blanco-Gonzalo, *et al.* (2018)) to support that mobile-based biometrics will be accessible to, or usable by, everyone. Federal agencies looking to leverage multifactor authentication need more data-driven insight into the usability and accessibility of these technologies. NIST recommends observational usability testing for assessing multifactor authentication and biometrics (Theofanos, Stanton, and Wolfson (2008)). However empirical comparison of authentication schemes, including biometrics, is not common. The historical lack of a standard usability metric in authentication research contributes to difficulty comparing usability across schemes (Ruoti, Roberts, and Seamons (2015); Trewin, *et al.* (2012)).

1.2. Related Work

The body of literature on both the accessibility and usability of authentication schemes is growing, but currently remains small relative to the body of work on authentication usability. It has been noted that accessibility has not received adequate attention in biometric system design (Sasse and Krol (2013)). This section discusses prior research relevant to our focus. We build on existing literature by evaluating the relative usability of authentication schemes based on their effectiveness, efficiency, and perceived usability metrics for users with and without disabilities.

Ruoti, *et al.* (2015) emphasized the importance of empirical research when evaluating authentication schemes. The authors explored seven web-based authentication systems to determine what was most usable and what features participants valued most. They compared the usability of authentication techniques like e-mail-based and QR-based systems in a tournament-style "championship," measuring usability using the System Usability Scale (SUS) questionnaire. The authors recommend using SUS as a standard metric for future evaluation of new authentication systems. Our usability comparison employed the UMUX-LITE perceived usability scale, which has been shown to be an acceptable alternative to SUS (Lewis (2018); Lewis, Utesch, and Maher (2015); Borsci, *et al.* (2015); Berkman and Karahoca (2016)).

In 2012, Trewin, *et al.* (2012) conducted a lab study of three biometric schemes and one password scheme. They observed six experimental conditions: PIN, voice, face, gesture, face and voice together, and gesture and voice together. The authors collected biometric performance, interaction time, error rates, memory recall success rate, and self-reported reactions using modified SUS. They observed that despite the fact that the voice biometric condition resulted in the least errors and performance time, participants found it lacking in usability, gracing it with a SUS score of "D." The authors proposed this might have been due to the volume required for participants to provide an acceptable voice sample. We too used time and a SUS variant as usability metrics. Trewin, *et al.* emphasized the importance of providing appropriate feedback on achieving proper facial biometric alignment to reduce the number of errors and the time for biometric recognition to occur; a similar conclusion is discussed later in this paper.

Blanco-Gonzalo, *et al.* (2018) performed a comparative study on the usability and accessibility of mobile biometrics. They investigated the accessibility of voice, face, fingerprint, PIN, and pattern schemes and compared the usability and accessibility of the more traditional authentication method of PIN to biometric authentication techniques. They also included multiple groups of participants with disabilities (upper body, lower body, visual, and cognitive) and a control group of participants without disabilities. The authors measured task time, satisfaction, and errors. Similarly, we compared traditional and biometric authentication schemes (PIN, fingerprint, eye, palm), and worked with participants with low vision, participants who were blind, and participants with no disabilities. We measured similar metrics, although our error data were ultimately not usable for analysis. Unlike Blanco-Gonzalo, *et al.*, we included participants with hearing loss and did not examine pattern authentication. Our study also required participants to perform the tasks on their own devices so as to gain a better understanding of usability in the context of personalized assistive technologies. This study's results echo Blanco-Gonzalo, *et al.*'s findings for their control group. Our participants who had vision loss

preferred biometrics that did not require positioning (Section 2.2 explains positioning biometrics), similar to Blanco-Gonzalo, *et al.*'s finding that participants with visual disabilities disliked the face biometric.

2. Study Design

We recruited 30 diverse participants, including participants having limited or no vision or having hearing loss. We evaluated and compared six authentication modalities: PIN, palm, eye, face, face and voice together, and fingerprint. Two modalities, face and face and voice together, were removed from analysis because technical set-up difficulties caused too small of a sample size for these schemes. The International Organization for Standardization (ISO)'s definition of usability was employed: the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO 9241-11:2018 (2018)). We did not request nor were we provided performance data on the biometrics from the prototype application partner. This research focuses only on usability measures.

2.1. Mobile Application Prototype

HYPR provided a real, working system and hosting resources to support a prototype of several modes of biometric authentication on iOS and Android devices. HYPR uses Fast IDentity Online (FIDO) and a "decentralized" authentication concept. The user's device application allowed six authentication schemes for "unlocking" a private key. Biometric privacy precautions are discussed in Section 3.5.

On installing the application on an iOS or Android device, participants were prompted to enroll their biometrics within the application. PIN, palm, face and voice, fingerprint, and eye were available within the iOS application. PIN, palm, face, fingerprint, and eye were available within the Android application. Enrollment included text and illustrations on how to position the phone to capture the biometric best. Some also contained text or visual cues during the enrollment process, such as text content suggesting where to move a phone, green bars lighting up when the user's eyes were properly positioned within a bounding box, or displaying a red circle to show where to position their palm on the screen. After enrolling one or more authenticators, a dashboard was enabled for participants, showing icons representing each authenticator enrolled. On selecting an icon on the dashboard, the participant was able to try to log in using the corresponding scheme.

2.2. Hypotheses

PIN is considered a baseline similar to the most common authenticator, passwords, where users enter characters or numerals using a keyboard. From observations in pilot sessions and informal interviews with people with disabilities, we created a dynamic-positioning versus a nondynamic-positioning categorization for biometrics. We define dynamic positioning as interactions where users are required to position and hold their device in relation to a specific point on their frame (dynamic-positioning actions). We define nondynamic positioning as interactions where users are not required to position or hold their device in relation to a specific point on their frame (static-positioning actions). We predicted three patterns would emerge in our study:

H1: User performance (efficiency and effectiveness) will be different between PIN and biometric schemes;

H2: User performance will be different between positioning biometrics (eye, palm) and nonpositioning biometrics (fingerprint); and

H3: For the user group with vision loss, user performance will be better with nonpositioning biometrics than it will with positioning biometrics.

2.3. Task Performance Metrics

2.3.2. Efficiency and Effectiveness

Efficiency was operationalized as response time on an authentication task. Response time was captured by measuring elapsed time on task from the start and end of screen prompt page loads. The mobile authentication application was reviewed to identify common start and end screens for the login task. The task start was

considered the first page loaded after selecting a biometric or PIN login icon. The start time was the moment when the mobile application page fully rendered in the session screen recording. On biometric recognition, the mobile application displayed a "success" page, and in fingerprint, "success" was represented by a pop-up message. The app displayed a failure message if authentication was not successful. Task end times were collected on success or failure page or pop-up load. Time in milliseconds was manually captured from video of the mobile screens.

All participants were provided time on each authentication task with no support from the facilitators. Some participants requested assistance mid-task. In these cases, they were given lightweight verbal guidance such as "try that again," or more detailed verbal and/or physical guidance if requested, like frequent verbal directional instructions (ex. "try moving the phone closer to your face"). We therefore categorized completion types (*effectiveness*) as:

- Independent success;
- Success with light guidance (few light verbal prompts);
- Success with heavy guidance (frequent, detailed verbal guidance and/or physical guidance); and
- Failure.

Independent success and success with light guidance were grouped as *trial success* in our analysis. Success with heavy guidance and failure, including instances when participants chose to end the trial, are both considered *trial failure.* Generally, choosing to end the trial only happened after a number of errors had occurred.

2.3.2. Perceived Usability

The 10-item long System Usability Scale (SUS) is an industry standard method of assessing a user's *perceived usability* of a system and has been recommended as a standard metric for comparing usability of authentication systems (Ruoti, *et al.* (2015)). We deemed requiring participants to complete the 10-item questionnaire several times as too cumbersome for participants in our specific study design. Due to the long task set-up times, short task times, and rapid switching between tasks, we selected a shorter perceived usability questionnaire, the UMUX-LITE. Figure 1 shows the two questionnaire items.

FIGURE 1. UMUX-LITE questionnaire items

Item 1: This system's capabilities meet my requirements.
Item 2: This system is easy to use.

UMUX-LITE (Lewis, Utesch, and Maher (2013)) is a two-item questionnaire based on the Usability Metric for User Experience (UMUX) questionnaire. It has been shown to have high reliability and validity. A regression-adjusted version called the UMUX-LITEr has been found to correspond closely with the SUS in assessing user satisfaction in a given system (Lewis (2018); Lewis, Utesch, and Maher (2015); Borsci, *et al.* (2015); Berkman and Karahoca (2016)). We report results in UMUX-LITE format, but interested readers may use this adjustment to transform perceived usability data into SUS equivalency scores, which combine results of both UMUX-LITE items. The conversion is:

$$\text{SUS equivalency score} =$$

$$.65*(((\text{UMUX LITE Item1} - 1)+(\text{UMUX LITE Item2} - 1))*(100/12))+22.9$$

2.4. Ensuring Accessibility

Because we examined usability for populations with specific disabilities, it was especially important to ensure test materials and environments were accessible for people with limited or no vision and people with hearing loss. Lab environments and building entrances were checked for accessibility prior to sessions. All equipment that was not the subject of testing was accessible to and comfortable for participants. We confirmed that all

elements in the prototype application could be read by a screen reader. Signature guides were provided for users with low or no vision to use on consent forms. Consent forms were provided digitally ahead of the session to participants with vision loss to give them time to review the information. Upon scheduling, participants with hearing loss were asked if they desired American Sign Language (ASL) interpretation. If requested, an ASL interpreter was present to facilitate communication during the study as well as during introductions, consent discussion, debriefing, and other immediate pre- and post-session interactions. Participants who used hearing aids in everyday life used them during the study.

Based on informally received advice within the usability community on working with people with disabilities, we chose to select participants who were willing to use their personal smartphones and install the application needed for the study. Personal devices help ensure that the hardware used in research is easily accessible to participants as it enables participants to use their personal assistive technology configurations. This method also allows facilitators to observe the individual approaches to using a smartphone by participants with audio and visual disabilities' and how the authentication methods in question interact with participants' everyday assistive technologies. Using personal devices provides privacy advantages as well (see Section 3.5).

3. Methodology

We conducted a lab study comparatively evaluating the usability of three biometric authenticators (fingerprint recognition, eye recognition, and palm recognition) and one nonbiometric authentication scheme (PIN). Participants completed a presession survey, described in Section 3.1, before the session. After giving informed consent at the start of the session, a facilitator assisted participants through prototype setup. During the session, participants used each authentication scheme to perform login tasks using the smartphone application. After the task portion, facilitators engaged the participant in structured interviews to gather their opinions on the accessibility and usability of each authentication mode, their general preferences between the schemes, and their thoughts regarding personally using the technology to authenticate into online services. The structured interviews are not described further in the Methodology as they are beyond the focus of this paper. The study ran for 2½ weeks during the summer of 2018, with 29 participants taking part in the study.

3.1. Presession Survey

A survey on authentication use and behaviors was developed to ascertain participants' technical acumen and security awareness, and to surface meaningful relationships between experimental results, demographics, and technology perspectives. Survey analysis is outside of this paper's current focus. It is only discussed here for transparency and insight into performance results. The survey first gathered the types of technologies participants regularly use, including assistive technology. It then evaluated their awareness and use of authentication technologies such as passwords, patterns, and biometrics. Finally, it attempted to identify how security-minded participants were by including questions about password-sharing practices, software update habits, and types of sensitive accounts they access from their devices. The behaviors surveyed were constrained to best practices for securing a sensitive application on a personal smartphone. Participants were offered the options of completing the study online in advance of the session, completing the study on paper, or completing the study verbally at the beginning of their scheduled appointment.

3.2. Study Setup

Sessions took place in conference rooms. Environments were accessible, and light levels were controlled to ensure minimal interference with camera-based authentication actions.

Audio recordings and top- and side-view video recordings centered on the participant's interactions with the prototype were captured. If participants had iOS devices and agreed to it, their screens were captured using iOS screen sharing to a researcher's laptop. Recordings started after the participant provided an informed signed consent and explicitly consented to being recorded. Video and audio recordings were later used to manually calculate response times and to double-check live session notes. When an ASL interpreter was present, they sat in full sight of the participant and aided communication between participant and facilitator.

Participants used their personal mobile devices for the study. The facilitator guided the participants through downloading and installing the mobile application and enrolling their authentication information to the prototype, providing aid if needed. Enrolling included performing each authentication action and thus served as an introduction to unfamiliar authentication schemes and a practice for all schemes. Before enrolling in any schemes, the participant was instructed not to use any passwords or PINs they had used before or planned on using outside of the study.

Participants were informed that facilitators could answer questions related to the study at any time during the session and answer questions related to using the authentication schemes during setup and after the tasks, but not during experiment trials. Since the prototype used unlabeled icons as elements to navigate to authentication tasks, the icons were explained to participants and a visual cheat sheet of the icons was provided during registration and tasks.

Participants took 60 to 90 minutes to complete the study and were compensated $100 U.S. dollars (USD) in cash. Regardless of completion of the session, participants with disabilities received an additional $25 USD incentive to compensate for added travel time and expense.

3.3. Tasks

Tasks began at the home screen of the prototype application. The facilitator described a fictional scenario in which the user's goal was to use their mobile device to log in to a Government service called MyUSA Account in order to download a digital copy of their latest tax returns. Participants were aware that the service was not real but were asked to place themselves in the scenario. It was used to ground experiences in real-life application and introduce using biometric authentication for digital Government services.

The facilitator directed the participant to authenticate using a particular scheme. To start a task, the participant tapped the corresponding authenticator icon. A "trial" began when the application instructed the participant to attempt the authentication interaction. The trial continued until a Success or Fail was achieved. Before each PIN trial, the participant was reminded not to use passwords that they had used before or planned on using outside of the study. *Tasks* consisted of two sequential trials using the same scheme. A *trial* was an individual attempt to authenticate using the task scheme.

Trials could contain multiple authentication interactions if errors occurred. If an error occurred (known by the appearance of an error message), the participant was told to try authenticating again. The participant completed the trial by achieving a success or failure (criteria in Section 2.3.1). After successful trials, the participant was returned to the app's home screen. Task order was counterbalanced to control for the possibility of task ordering patterns influencing results. After the first or second trial ended, the facilitator asked the participant to rate their agreement with each UMUX-LITE item on a scale of 1 (strongly disagree) to 7 (strongly agree). What trial the ratings were collected after was randomized to reduce the risk of repetitive questioning influencing participant responses.

Face recognition and face/voice combination were tested during the sessions by all participants who had registered those schemes. However, unexpected updates to the prototype application during the weeks the experiment took place caused technical difficulties with registration. Not enough participants were able to successfully register the two schemes to achieve a useful sample size, so face recognition and face/voice recognition are excluded from this paper's analysis.

Participants with disabilities were encouraged to use their normal assistive technologies during the study. Participants with limited or no vision used VoiceOver, screen magnification, and color filtering assistive technologies to complete the tasks, depending on their needs. Participants with hearing loss did not use assistive technology on their mobile devices, but some made use of ASL interpretation.

3.4. Participants

We worked with a professional usability recruitment firm to recruit 30 participants who were U.S. citizens living in the Northern Virginia and Baltimore region. We aimed to balance the sample overall for gender and include participants across the following age groups: 18–24 years old; 25–34 years old; 35–44 years old; 45–54

years old; 55–64 years old; and 65 years old or over. All participants were required to be fluent in English or ASL.

One of the 30 participants did not show for their session and could not be rescheduled, giving an overall participant count of 29 (13 women and 16 men). Two participants were unable to set up the prototype due to technical difficulties, giving a final count of 27 participants supplying task performance data. These two participants still took part in the survey and structured interview. Participant ages skewed older. Table 1 gives the number of participants in each age range.

TABLE 1. Number of participants in each age range, 29 participants total

	Age range (years)				
	25–34	35–44	45–54	55–64	65+
Number of participants	1	6	8	5	9

All participants were required to own a smartphone and agree to install a mobile phone application for the duration of the study. Smartphones were Android OS 4.4+ or iOS models 5s and above or iOS 9.1+ and had operational fingerprint sensors and operational front-facing cameras. Participants were requested to bring all assistive technology they use regularly with their mobile devices to their study session. Six participants owned an Android device and 23 owned an iOS device.

Participants were grouped into those with no disabilities (control), participants with hearing loss, and participants with limited or no vision. We aimed for recruitment of 10 participants with moderate to profound hearing loss (phrased as "hearing impairment") with no more than 5 participants who required an ASL interpreter, and 10 participants with moderate to profound vision loss (phrased as "vision impairment"). An additional requirement was that these participants not have any other disabilities. All disabilities and levels were self-reported by participants to the usability recruitment firm. The following definitions were provided to the firm for recruitment guidance:

Visual impairment (at the participant's presenting corrected vision level) (World Health Organization (n.d.)):

- Low vision, consisting of partially sighted, moderate visual impairment or severe visual impairment; and
- Profound visual impairment, legally blind, or totally blind.

Hearing impairment (World Health Organization (2017); Canadian Association of the Deaf (2015); Clason (2015)):

- Moderate impairment or hearing loss, or hard of hearing; and

Detail: Someone with a moderate level of hearing loss has difficulties hearing regular conversational speech, even at close distances. This includes people who use technology that allows them to operate at a less severe hearing loss level, ex. cochlear implants, hearing aids.

- Severe to profound impairment or hearing loss, deaf, or total hearing loss.

Detail: Someone with a severe or profound level of hearing loss does not hear conversational speech. Someone with a severe level may hear very loud speech or loud sounds in the environment, such as a fire truck siren or a door slamming. Someone with a profound level or someone who is deaf does not hear conversational speech and may perceive loud sounds as vibrations. They cannot understand speech (with or without hearing aids or other devices) using sound alone (i.e., no visual cues such as lipreading).

Table 2 details participants per group and level, as reported by the recruitment firm. It also presents how many participants completed enrollment in each authentication scheme.

TABLE 2. Number of participants enrolled in each authentication scheme.

Authentication schemes	Disability type and level						
	Visual		Hearing			None	Total
PIN	7		11			9	27
Finger print	7		11			9	27
Eye print	7		9			9	25
Palm print	5		9			6	20
Face	1		2			3	6
Voice/Face	1		6			6	13
Total participants	**9**		**11**			**9**	**29**
Total participants by level	**Total 6**	**Moderate 3**	**Total 4**	**Severe 2**	**Moderate 5**		

3.5. *Ethics and Privacy*

The experimental design was approved by The MITRE Institutional Review Board (IRB). At the start of each session, the participants were given a consent form to sign, detailing the study and their rights as participants. Consent forms were provided in accessible formats and with longer review times, when appropriate. We took care to treat all participants with respect and performed accessibility checks of materials and lab settings before sessions (see Section 2.4). Participants were informed that, among other participant rights, they would receive a prorated incentive if they chose to end the session early.

Facilitators reminded participants frequently not to use any past or future personal passwords or PINs. The simulation prototype did not include any identity verification steps. All passwords, PINs, and biometric data created during the study were stored locally on the participant's personal smartphone and were not transmitted from the device or out of the application. Facilitators supervised participants securely installing and uninstalling the prototype at the start and finish of each research session. Participants were made aware of these precautions.

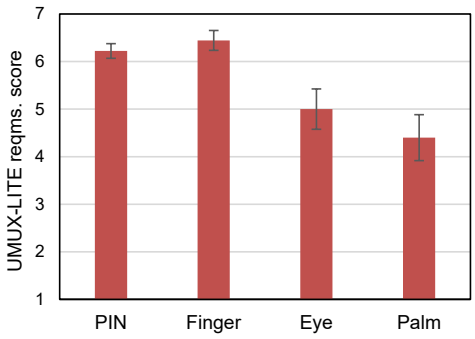
4. Results

This section reports the quantitative data gathered, organized by metric. We also show participants' prior exposure to biometric authentication schemes, as reported on presession questionnaires. As this paper focuses on task performance data, analysis of qualitative interviews is deferred to future publications. Tables 4 to 9 in the appendix present further details on the results.

4.1. *Perceived Usability*

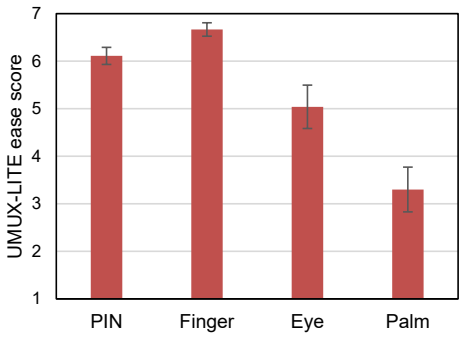
Perceived usability was measured through responses to UMUX-LITE items, shown in Figures 2 and 3. Note that 1 corresponds to the "strongly disagree" response and 7 to the "strongly agree" response.

FIGURE 2. Mean UMUX-LITE requirements item scores across authentication schemes and all populations



NOTE: Error bars indicate standard errors.

FIGURE 3. Mean UMUX-LITE ease item scores across authentication schemes and all populations



NOTE: Error bars indicate standard errors.

UMUX-LITE data were not normally distributed, therefore nonparametric tests were needed. Due to the interval nature of the data, k independent samples analysis was performed.

A Kruskal-Wallis H test, a one-way ANOVA on ranks for nonparametric data, showed that there were no statistically significant differences in requirements item scores between the different populations ($\chi^2(2) = 2.000$, $p = 0.368$, with a mean rank score of 45.17 for no disability, 54.25 for hearing loss, and 49.62 for vision loss); nor in ease-of-use item scores between the different populations (χ^2 (2) = 0.415, p = 0.813, with a mean rank score of 49.33 for no disability, 52.01 for hearing loss, and 47.75 for vision loss).

There were significant differences in the UMUX-LITE requirement ratings between schemes; χ^2 (3) = 19.000, p = 0.000, with a mean rank score of 55.56, 42.36, 64.54, and 32.43 for PIN, eye, fingerprint, and palm, respectively. Mann-Whitney post-hoc tests, the nonparametric alternative to the independent sample t-test, found significant differences in requirements item scores between several schemes. Median requirements ratings were significantly higher for PIN (6) than palm (5); (U = 135.000, p = 0.003). Median requirements ratings were significantly higher for fingerprint (7) than eye (6); U = 197.500, p = 0.005. Finally, median requirements ratings were significantly higher for fingerprint (7) than palm (5); U = 45.000, p = 0.000.

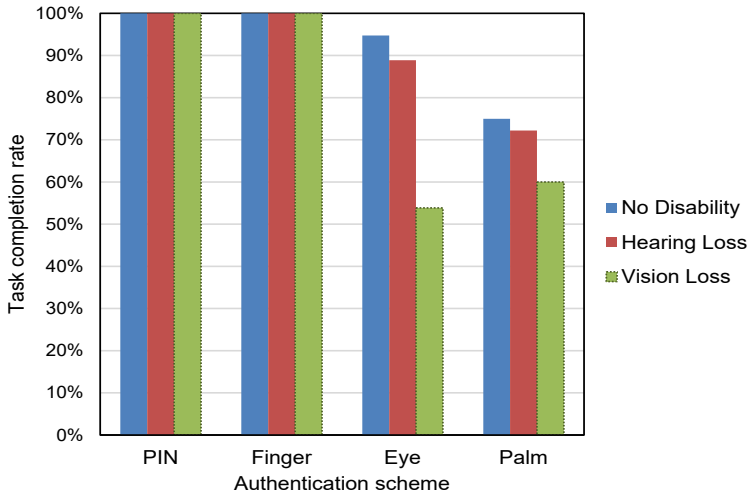
A Kruskal-Wallis H test showed significant differences in ease-of-use item scores between schemes; χ^2 (3) = 33.048, p = 0.000, with a mean rank score of 54.50, 45.76, 68.94, and 23.65 for PIN, eye, fingerprint, and palm. According to a Mann-Whitney post-hoc test, median UMUX-LITE ease ratings were significantly higher for fingerprint (7) than PIN (6) (U = 228.500, p = 0.007), eye (6) (U = 187.000, p = 0.002), and palm (3) (U = 100.000, p = 0.000). Median ease scores were significantly higher for PIN than palm (3); U = 75.000, p = 0.000. They were also significantly higher for eye than palm; U = 143.000, p = 0.013.

A Mann-Whitney post-hoc test was run to test the third hypothesis about the experiences of participants with vision loss. It determined that median requirements item scores were significantly higher for PIN (7) than eye (4), (U = 5.500, p = 0.011); and palm (3), (U = 5.500, p = 0.036). Median requirements scores for fingerprint (7) were significantly higher than eye, (U = 5.500, p = 0.011); and palm, (U = 5.500, p = 0.036). Finally, median ease item scores were significantly higher for fingerprint (7) than for eye (3) (U = 9.000, p = 0.033); and for palm (2) (U = 5.500, p = 0.036).

4.2. Effectiveness

Effectiveness was assessed through measuring completion rate. Task completion rate is the number of successful task completions out of the number of attempted task completions (which are also the number of successful scheme registrations). Note that each participant had two task attempts (one per trial). Figure 4 shows completion rate results.

FIGURE 4. Mean completion rates across authentication schemes and participant groups



A logistic regression was performed to ascertain the effects of population on the likelihood that participants successfully completed the tasks. The model explained 5.4 percent (Nagelkerke R^2) of the variance in completion rate and correctly classified 89.4 percent of cases. Population was found to have an effect, with participants with no disability being 3.690 times more likely to be successful than those with vision loss; χ^2 (1) = 4.372, p = 0.037.

Every participant who registered a PIN and fingerprint was able to successfully complete the PIN and fingerprint tasks, regardless of participant group. No participant group had 100 percent task completion rates for eye and palm tasks. However, a logistic regression performed to examine the effects of the authentication scheme on the likelihood that participants successfully completed the tasks found no significant differences between completion rates due to mechanism. The model explained 35.5 percent (Nagelkerke R^2) of the variance in completion rate and correctly classified 89.4 percent of cases.

To ensure no learning effects were at play, a logistic regression was used to ascertain the effects of number of trials (1 or 2) on the likelihood that participants successfully completed the tasks. The model explained 0.2 percent (Nagelkerke R^2) of the variance in completion rate and correctly classified 89.8 percent of cases. There were no significant differences between completion rates due to trial number and thus no learning effect due to number of trials experienced; χ^2 (1) = 0.185, p = 0.667.

A logistic regression was performed to ascertain, specifically for the vision-loss group, the effects of scheme on likelihood that participants successfully completed the tasks. The model explained 47.6 percent (Nagelkerke R^2) of the variance in completion rate and correctly classified 80.8 percent of cases. There were no significant differences for this group between completion rates due to scheme.

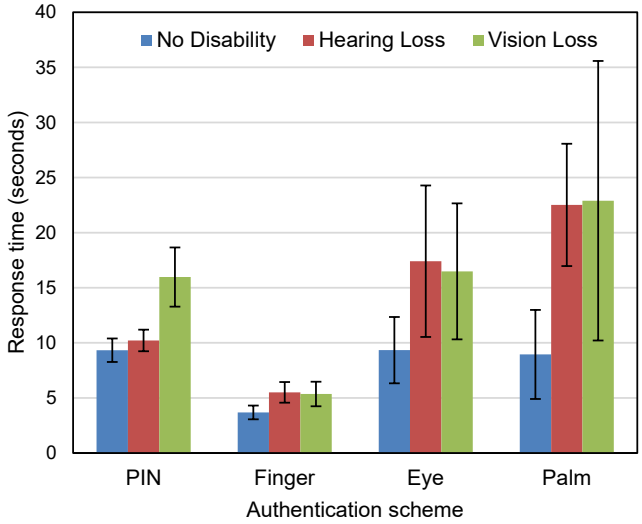
We planned to examine error rate as a component of effectiveness, with an error defined as an instance when the participant does not fail the task but must redo the authentication action. The prototype gave error prompts such as "Incorrect Match, This palm does not match the saved value," "Authentication Aborted, The eye authenticator timed out," and "Unable to authenticate, Eye verification not matched." However, prompts also included descriptions like "An Error Occurred, Unexpected HTTP status code received" and simply "Authentication Failed" with no explanation. Since some error messages were opaque and the prototype was created and managed by a third party, we were unable to accurately diagnose the genesis of each participant error or to guarantee that all errors were user-caused and never the result of a technical glitch (as the HTTP status code message implied). Therefore, we consider error rate data unfit for the same degree of scrutiny as completion rate, and do not report it here.

4.3. Efficiency

Efficiency is operationalized as response time, specifically, the time elapsed from when the prototype app instructed the participant to attempt the authentication interaction until the interface's indication of task success or failure (overall, length of a trial). Data from all success task trials are reported here. Figure 5 presents response-time results.

Mauchly's Test of Sphericity, which tests the assumption that the variance between the levels of independent variables are equal, indicated that the assumption of sphericity had been violated, ($\chi^2(5)$ = 44.308, p = 0.000). A Greenhouse-Geisser correction, typically used when the assumption of sphericity is violated, was used. A repeated measures ANOVA found that population had no significant effect on response time; F(2, 20) = 2.246, p = 0.132. A repeated measures ANOVA also found that scheme had no significant effect on response time; F(1.741, 34.823) = 3.260, p = 0.057.

FIGURE 5. Mean response time from all success trials across authentication schemes and participant groups



NOTE: Error bars indicate standard errors

However, the lack of power ($\eta = 0.546$) may have limited the ability to find a significant effect. Because differences between schemes were hypothesized, post-hoc tests were still performed on scheme comparisons. Additionally, many post-hoc procedures are designed to control familywise error rates in the absence of a significant prior omnibus analysis. Simple contrast post-hoc tests with Bonferroni correction, a correction made to p-values when several statistical tests are performed on a single dataset, found significant differences in response times. Specifically, the mean response time for fingerprint (4.86s) was significantly faster than mean response times for all other schemes (PIN (11.41s), $F(1, 20) = 37.520$, $p = 0.000$; eye (13.71s), $F(1, 20) = 5.339$, $p = 0.032$; and palm (18.24s), $F(1, 20) = 10.421$, $p = 0.004$).

To test specifically within the vision-loss participant group, a one-way repeated measures ANOVA was performed to ascertain the effect of scheme on reaction time, with planned pairwise comparisons. No significant differences between schemes were found $F(3,12) = 1.154$, $p = 0.367$. The lack of power ($\eta = 0.236$) may have limited the ability to find a significant effect. Because differences within the vision-loss group were hypothesized, post-hoc tests were performed, but planned pairwise comparisons found no significant differences between schemes on reaction time for the group with vision loss.

4.4. Biometric Authentication Scheme Experience

In the presession survey, 30 participants reported on the authentication schemes they had previously used to secure both their device and any personal accounts (such as a banking account). Items were phrased: "Do you have experience with the following ways to__?" Illustrative examples were included, like banking account for personal account and RSA token for digital key. Table 3 shows their responses. Password, PINs, two-factor with email and SMS, and fingerprint biometrics were all widely used. All participants reported experience using passwords to secure both devices and individual accounts, and over 80 percent reported experience using a PIN or pattern. Most participants had experience with some form of two-factor authentication, with the majority of the experience with a code received by email or SMS or with using a security question. A majority (83 percent) had used a biometric fingerprint to unlock their smartphone, and 60 percent had used fingerprint to unlock a personal account. A small number reported experience with face or voice biometrics to secure their phone (3 with voice, 2 with face). None had used palm or eye biometrics before for securing devices or accounts for web services.

TABLE 3. Participant responses to questionnaire items about prior experience with authentication methods

Authentication method	Number of "yes" responses to the following questionnaire items:	
	… secure your personal devices to access a web service?	… secure your personal accounts to access a web service?
Passwords	30	30
PIN or pattern	25	24
Two-factor using code received by email	23	22
Two-factor using security question	21	22
Two-factor using code received by personal cellphone or smartphone	20	19
Two-factor using standalone device with digital key	7	5
Two-factor using a code received by landline phone	6	8
Two-factor using an online or software digital key (e.g., Google Authenticator, Duo)	4	4
Biometric—fingerprint	25	19
Biometric—voice	3	2
Biometric—face	2	1
Biometric—iris	0	0
Other	0	0

5. Discussion

5.1. Traditional Authentication and Biometric Authentication

Performance data partially supported the hypothesis that PIN and biometric authentication schemes would differ in the metrics we collected. PIN had significantly lower perceived usability (specifically, ease of use) and lower efficiency (slower response time) than fingerprint. PIN had significantly higher perceived usability than palm (both items). Counter to expectations, no significant differences were seen between PIN and eye in any metric, and no significant differences in completion rates were seen for any scheme.

PIN and Fingerprint

The PIN/fingerprint difference could have been caused by the two schemes' different memory requirements and their required target acquisition actions. To use PIN successfully, participants had to recall a six-digit pattern, while they did not have to remember anything for fingerprint. For PIN, users performed six input actions in selecting six digits in the keypad entry interface; for fingerprint, they simply had to touch one input location (the touch sensor). In both the recall and the physical input differences, PIN's actions have a longer inherent time burden than do fingerprint's, which could explain the response-time difference. When using PIN with a screen reader during sessions, participants often had to cycle through digits listening for the correct one before selection—again, a possible time sink. Recall also brings in a cognitive element that the fingerprint scheme does not need. Preferences against the need to create and remember PINs could have affected perceived usability ratings.

The added cognitive component and the speed differences might have contributed to participants rating PIN and fingerprint differently for ease-of-use. The lack of difference in the "meets my requirements" aspect of perceived usability could indicate that participants held expectations of a minimum threshold of usability required to fulfill their needs, and that both schemes met such a threshold, causing a ceiling effect. Participants may also have viewed PIN and fingerprint similarly in terms of the security that the schemes provide.

PIN and Palm

PIN and palm's perceived usability difference as authentication schemes could again stem from different cognitive requirements and different time burdens. The palm authentication interaction of positioning the palm parallel to the phone's screen-side camera and adjusting accordingly does not easily compare time-wise to PIN's classic target acquisition and selecting numbers on an onscreen keypad. That said, palm had a longer mean response time (18.24s) than PIN. Basic times for both actions could be assessed, for example with Goals, Operators, Methods, Selection Rules (GOMS) model analysis (Kieras (1999)), to delve deeper into comparisons of the schemes' inherent time burdens. PIN and palm's cognitive actions differ as well; remembering a number sequence is a one-time recall, while reaching and maintaining a correct relative hand position involves continuous spatial monitoring and adjustment.

Differences in time to authenticate and in cognitive actions required, as well as in perceived security provided by each scheme, could have contributed to the differences in perceived usability between the schemes. Prior exposure could have had an effect as well, since a majority of participants reported having used PIN or pattern before the session and no participants reported using palm authentication before the session.

The palm print condition had the smallest sample size since fewer participants were able to successfully enroll palm print than other schemes. The sample shrunk further for response-time data as only results from successful trials were included in efficiency analysis. The lack of a significant efficiency difference does not align with expectations, but it may have been caused or affected by the lack of power and the high variability in palm response-time results.

PIN and Eye

Counter to expectations, there were no significant differences between PIN and eye schemes. Eye's low sample size could have impacted the ability to find a significant difference if there was one, although eye's sample size was larger than palm's sample size. From observation, eye seems to be more similar to palm than to PIN. Like palm, there is no recall needed, and the user continuously monitors and adjusts their relative hand positions. Unlike palm, in eye, a hand holding the mobile device is positioned relative to the user's head, and authentication requires assuming a specific head posture and face configuration (eyes open, gaze on the phone). In fact, eye and palm differed significantly in their ease of use item scores.

Within sighted participants, the prototype app feedback for eye seemed easier for users to monitor than did feedback for palm. During palm authentication, some sighted users shifted their hand away from and back over the screen as well as tilted their hand to peek under it to view the screen more fully. Some users remarked on these actions. No such actions or comments on ability to perceive feedback were seen during eye authentication sessions with sighted participants (perception of feedback being different from understanding of feedback).

We are ultimately unsure as to why participant performance did not differ significantly between the PIN and eye authentication schemes. There were no statistically significant effects of participant group on perceived usability or efficiency, but participants with no disability were 3.69 times more likely to complete tasks successfully than were participants with vision loss. This suggests that vision-loss participants' different experiences of PIN and eye bear further study.

Overall

PIN-fingerprint and PIN-palm comparison differences were supported by a subset of performance data, though not by completion rate (addressed in Section 5.3). A PIN-eye difference was unsupported. This mixed bag suggests that there might not be a clear usability divide between traditional authentication methods and biometric schemes. Another possibility is that traditional methods may indeed have distinct usability differences from some biometrics, but that grouping the biometrics examined here into a single usability category is an overreach.

Biometrics offer many advantages over traditional authentication schemes like PIN and password. They do not require recall, which cuts down on cognitive burden as well as time. However, some biometrics, such as

palm and eye, require additional monitoring of spatial information. This comparison merits further research to empirically evaluate the usability of PIN and other biometrics that can be captured by smartphone cameras or sensors. Future studies could explore: comparisons with use over time, for example authenticating several times over the course of months; comparing with stringent PIN or password creation requirements; use in field settings; larger sample sizes; and users with other single or concurrent disabilities.

5.2. Dynamic-Positioning Interactions in Authentication

Fingerprint, the nondynamic-positioning biometric authentication scheme, had significantly higher perceived usability (both items) and better efficiency than eye and palm, the dynamic-positioning biometrics. This supports the hypothesis that biometric authentication schemes' performance results would divide along the dynamic-positioning aspect. Counter to expectations, no scheme showed significantly different completion rates.

As discussed earlier, eye and palm authentication schemes share similarities—no need for recall, and a continuous spatial information monitoring by the user. Fingerprint also does not require recall, but neither does it need hand and/or head position perception and adjustment. It simply requires the user to locate and select a single, nonmoving target with tactile edges. In cases where the user is holding the phone in one hand, they can even brace their fingerprint-input hand against their phone-holding hand. Dynamic-positioning actions require more granular and frequent monitoring and adaptation of the body part's location as well as movement and pausing in space, generally with no physical bracing or tactile breakpoints. This difference in the use of dynamic positioning—positioning one body part relative to another, whether hand-to-hand or hand-to-head—is a likely cause for the performance differences seen between schemes.

There were no significant differences in completion rates between the comparison of finger and eye and finger and palm authentication schemes. This lack of significance could stem from a small sample size, or from differing levels of familiarity with the schemes. Most participants had previous experience using a fingerprint scheme and none had used eye or palm authentication schemes before their sessions.

Results partially support the prediction that biometric schemes would exhibit a usability split along dynamic positioning lines. Further research is needed to confirm this split and to explore its nature. Are there important distinctions within the types of biometrics captured by smartphone cameras and sensors? Are there meaningful groupings within the dynamic positioning conglomeration? Do individuals with certain disabilities experience disproportionally better or worse usability from positioning biometrics? Might different feedback channels (ex. audio tone, audio text, haptic vibration) of positioning guidance mitigate the effects of the split, so much so as to erase the dynamic positioning performance difference?

Results gave some support to the third hypothesis that the user group with vision loss would experience better performance with nonpositioning biometrics than with positioning biometrics. Low vision and blind participants reported significantly better perceived usability (both items) with fingerprint than with eye or palm. Also, participants with vision loss were far less likely to complete tasks successfully with given schemes than were control group participants (3.69 times). Since all enrolled participants had 100 percent completion rates only with PIN and fingerprint, this lower-success effect is likely occurring with eye and palm. No significant differences were found between completion rates due to scheme within the vision loss group, but this pattern is noteworthy and should be explored further in future. These results suggest that dynamic positioning is an important aspect of biometric usability and accessibility for users with low or no vision.

However, there were no significant efficiency differences between schemes for the group with vision loss. This could be affected by the lack of power.

It should be noted that the palm sample size of users with visual disabilities was small at five participants (other participants in the group were unable to enroll the scheme successfully). While five is not considered out of the ordinary for usability testing, it is a very small sample to support statistical analyses. Palm's sample size may have impacted results.

5.3. Effectiveness Metric

The completion rate did not vary significantly due to the authentication scheme. This was surprising, as the PIN and fingerprint schemes had 100 percent task completion rates and eye and palm had lower rates (mean 82.35 percent and 70 percent, respectively, over all participants). It could be that there was not enough power to see a significant effect. Levels of prior exposure to the schemes might have impacted the completion rate results; 83 percent of participants had used fingerprint and PIN or pattern schemes before, while no participants reported experience with eye or palm schemes before the study. There was no learning effect due to trial number, but familiarity could have had an impact larger than what the experience from registration and two trials could correct for.

Population significantly affected effectiveness, with participants with no disability being 3.69 times more likely to be successful than those with vision loss. All unsuccessful vision loss participants had been able to register the authentication schemes and could technically access the application content, but baseline access did not mean they could successfully use the schemes. Therefore, we recommend using the completion rate as a consideration in assessing technology usability and accessibility for low vision and blind users.

6. Limitations and Future Research Directions

The response-time measurement method was prone to human error. As described in the Methodology, researchers manually calculated response times from videos of the prototype screen. Though care was taken to move through videos at low-frame rates, measurements may have gained errors during this process. We recommend automating task time capture instead.

As detailed in the Results, useful error data could not be captured consistently due to prototype limitations. We believe error rate and diagnosis would be useful for future work.

Enrollment or registration performance was outside the scope of this study. Enrollment performance data, such as how difficult the participant found enrollment in a scheme and how many registration fails they caused, could give interesting insights.

What trial the perceived usability ratings were collected after was randomized to reduce the risk of repetitive questioning influencing participant responses. In retrospect, the risk of question repetition influence may have been lower than risk of effects due to uneven experience with the system. To address this, we recommend gathering self-reported ratings after every trial or after the same number of trials, and/or building in more participant interactions with the system to pursue a high enough level of familiarity, so that the lack of experience does not have an effect. The latter is the better option, as it would also combat difference in general levels of familiarity with particular schemes, as participants' prior exposure to authentication schemes could have had an effect, especially on results that showed high variability. Prior experience with the tested technology has been shown to affect SUS scores (McLellan, Muddimer, and Peres (2012)). Previous exposure should be examined in future studies for possible impact on perceived usability or other performance results or should be further controlled for.

Some metrics may be better suited to testing *across* disabilities and some to testing *between* disabilities. Response time might not be a useful metric for comparisons between groups where groups have different disabilities. It could be a more useful metric in within-group situations, since the functional effects of the disability on response time (ex. effects of poor fine motor control) would be standardized. Assistive tech may additionally influence task time and would also be better standardized within groups. Completion rate and self-reported reactions (ex. SUS scores), on the other hand, can more easily be compared across groups.

It is possible that slower response time does not always indicate inferior usability. Users might consider a scheme usable as long as it meets a minimum response time threshold and might at that point not be concerned with what scheme is faster.

Our findings should be validated through replication of the experiment with larger participant pools. Our study size was small due to the difficulty of recruiting participants with disabilities, resource limitations, and technical difficulties. We hope this work is expanded further by studying more types of disabilities and by

investigating the effects of severity levels within disabilities. More biometrics should be compared to expand authentication design guidance to other schemes that will become more common in the future, as well as to the face and voice biometrics for which not enough data could be collected. More research into the directionality of usability differences for people with disabilities would also be valuable as it could contribute to clear, evidence-based guidance toward selecting certain schemes over others.

We recruited participants into groups based on their self-reported disabilities. During the study, there was confusion over the definition of disability severity levels ("moderate," "severe," "total"). Many users did not describe their disabilities with this terminology. We recommend instead including assistive technology use when forming participant groups, as that may be more indicative of the type and degree of a hearing or vision loss. We also recommend a focus on testing authentication schemes with populations whose disabilities map to the scheme's interaction requirements, as these may have more immediate value. We observed usability decrements for participants with vision loss using schemes with a greater reliance on visual feedback, while users with hearing loss and control participants did not seem to have markedly different experiences with our analyzed schemes, none of which involved audio or speech-based interactions.

This work prompts ideas for future pursuit. Considering how the specific interactions that a biometric requires relates to the abilities of the user could surface more accessibility considerations like dynamic positioning that can be used to guide accessible authentication design. Further, it is not uncommon for people to have more than one disability. Usability for participants with multiple disabilities should be investigated.

We are also interested in how learnability may play a role in biometric accessibility. Participants with vision loss often expressed excitement and interest in eye and palm authentication during the study but sometimes could not employ them without verbal and occasionally physical assistance from facilitators. However, these participants said they were optimistic about their ability to learn to use the schemes over time. During informal background interviews, several technology users who had vision loss indicated that they frequently used iOS FaceID to secure their smartphones. They reported that the interactions were difficult at first, but that after some practice, they were highly satisfied with face recognition authentication and used it regularly. With repeated, possibly guided practice, certain authentication schemes that are initially difficult for participants with a disability to use may become easy and even preferred.

7. Conclusion

Our study found that there is not a clear usability divide between the traditional authentication method and all biometric schemes as a group. There may be no marked usability distinction, or it may be that fingerprint, eye, and palm are too distinct to consider together. The question of differences between traditional authentication schemes, like PIN or password, and biometric authenticators that can be captured by smartphones merits further exploration.

The results of our study partially supported a "dynamic positioning" split among the biometrics tested, with participants showing markedly different usability experiences between fingerprint and eye and between fingerprint and palm. The nonpositioning fingerprint scheme seemed somewhat more usable for participants with visual disabilities than the positioning eye and palm. Findings add weight to the positioning split. We propose research questions to further probe this categorization and other questions raised during the study, share thoughts on the metrics deployed in this usability evaluation, and discuss limitations in the experiment.

Based on the evidence collected, we propose dynamic device positioning as a new consideration for biometric usability evaluations. This new principle is operationalized as two actionable recommendations, to be used in authentication process design. Our recommendations were created with the accessibility and usability needs of citizen-facing Federal agencies in mind. Our work also contributes empirical findings on the usability of biometric authentication schemes for users with disabilities, expanding the body of work and demonstrating methods for comparative biometric usability evaluation with an accessibility focus.

7.1. Dynamic Positioning as an Accessibility Consideration

Smartphones offer a wide range of biometric capture, from fingerprint, eye, iris, face, and voice to emerging biometrics like ear shape. They offer conveniences to all users, including those with disabilities, but based on our research we feel that a better understanding of the accessibility of different biometrics is needed. There is little indepth usability guidance for designers to consult when integrating multifactor authentication into their services. Decision-makers at Federal agencies with accessibility mandates need to choose authentication techniques relatively early in the design process. They typically do not have the resources nor the time to perform rigorous experimentation on their web service's usability for people with disabilities. We seek to provide evidence-based knowledge to guide them in evaluating authentication options for people with disabilities and propose dynamic device positioning as a new consideration for usability evaluations of biometrics.

Participants with vision loss were far less likely to successfully complete tasks with given schemes than were control group participants. With this in mind, we suggest that completion rate is a key metric to consider when populations with disabilities are involved.

The fingerprint-eye and fingerprint-palm perceived usability and efficiency differences suggest that dynamic positioning could have an impact on biometric accessibility for users with low or no vision, though the relationship should be studied further and with larger participant pools.

We see positioning used alongside accessibility principles such as text alternatives for non-text content (WAI (n.d.)). Based on our findings, we offer the following recommendations to guide decision-makers in selecting biometric authentication techniques:

- A dynamic-positioning biometric should never be the sole authentication scheme.
- Multifactor authentication using biometrics should offer at least one nondynamic-positioning biometric. Fingerprint is a good option until other schemes are empirically shown to be more accessible.

References

Aleksandr, O., S. Bezzateev, N. Mäkitalo, S. Andreev, T. Mikkonen, and Y. Koucheryavy. (2018). Multifactor Authentication: A Survey. *Cryptography* 2(1): 1.

Berkman, M. I., and D. Karahoca. (2016). Re-assessing the Usability Metric for User Experience (UMUX) Scale. *Journal of Usability Studies* 11(3), 89–109. Retrieved from http://uxpajournal.org/assessing-usability-metric-umux-scale/.

Blanco-Gonzalo, R., C. Lunerti, R. Sanchez-Reillo, and R.M. Guest. (2018). Biometrics: Accessibility challenge or opportunity? *PLOS ONE: A Peer-Reviewed Open Access Journal* 13(3): e0194111. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5864003/.

Bonneau, J., C. Herley, F.M. Stajano, and P.C. Oorschot. (2014). Passwords and the evolution of imperfect authentication. *Communications of the ACM* 58(7): 78–87 (June 2015). Retrieved March 1, 2019, from https://dl.acm.org/doi/10.1145/2699390.

Borsci, S., S. Federici, S. Bacci, M. Gnaldi, and F. Bartolucci. (2015). Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *International Journal of Human-Computer Interaction* 31(8):484–495.

Canadian Association of the Deaf—Association des Sourds du Canada. (2015, July 3). *Definition of "Deaf."* (Canadian Association of the Deaf—Association des Sourds du Canada) Retrieved March 29, 2019, from http://cad.ca/issues-positions/definition-of-deaf/.

Cardello, J., and S. Farrell, S. (2017, July 27). *HealthCare.gov's Account Setup: 10 Broken Usability Guidelines.* Freemont, CA: Nielsen Norman Group. Retrieved March 1, 2019, from https://www.nngroup.com/articles/affordable_care_act_usability_issues/.

Clason, D. (2015, April 10). *From Mild to Profound: Understanding the Degrees of Hearing Loss.* Los Angeles, CA: Healthy Hearing. Retrieved March 29, 2019, from https://www.healthyhearing.com/report/41775-Degrees-of-hearing-loss.

Comer, Jr., J. (2018, November 29). 21st Century Integrated Digital Experience Act Floor Speech. Washington, DC: U.S. Congress. Retrieved March 1, 2019, from https://votesmart.org/public-statement/1306374/21st-century-integrated-digital-experience-act#.XHjELlNKjfb.

Corrigan, J. (2018, November 29). *The IDEA Act would require agencies to upgrade their websites to meet basic security and usability standards, but lawmakers did make some changes.* Washington, DC: Government Executive Media Group (www.nextgov.com). Retrieved March 1, 2019, from https://www.nextgov.com/it-modernization/2018/11/house-passes-bill-improve-governments-digital-services/153162/.

German, R. L., and K.S. Barber, (2018). *Consumer Attitudes About Biometric Authentication* (UT CID Report 18-03). Austin, TX: The University of Texas at Austin Center for Identity.

Grassi, P. A., J.L. Fenton, E.M. Newton, R.A. Perlner, A.R. Regenscheid, W.E. Burr, J.P. Richer, N.B. Lefkovitz, J.M. Danker, Y.Y. Choong, K.K. Greene, M.F., Theofanos. (2017, June). *Digital Identity Guidelines: Authentication and Lifecycle Management.* Washington, DC: U.S. Department of Commerce, National Institute of Standards and Technology (NIST Special Publication 800–63B). Retrieved March 1, 2019, from https://pages.nist.gov/800-63-3/sp800-63b.html.

International Organization for Standards (ISO) (2018) ISO 9241–11:2018: *Ergonomics of Human-System Interaction*—Part 11: *Usability: Definitions and Concepts.* Retrieved March 1, 2019, from https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en.

Kessem, L. (2018, January 29). IBM Study: *Consumers Weigh in on Biometrics, Authentication, and the Future of Identity.* Retrieved March 1, 2019, from https://securityintelligence.com/new-ibm-study-consumers-weigh-in-on-biometrics-authentication-and-the-future-of-identity/.

Kieras, D. E. (1999). *A Guide to GOMS Model Usability Evaluation Using GOMSL and GLEAN3.* Ann Harbor, MI: University of Michigan.

Konkel, F. (2018, April 4). *Bill Would Create Federal Customer Service Standards.* Washington, DC: Government Executive Media Group (www.nextgov.com). Retrieved March 1, 2019, from https://www.nextgov.com/cio-briefing/2018/04/bill-would-create-federal-customer-service-standards/147197/.

—. (2018, February 9). *It Costs Taxpayers $41 Per Phone Call To IRS.* Washington, DC: Government Executive Media Group (www.nextgov.com). Retrieved March 1, 2019, from https://www.nextgov.com/emerging-tech/2018/02/it-costs-taxpayers-41-phone-call-irs/145870/.

Lewis, J. (2018). Measuring Perceived Usability: The CSUQ, SUS, and UMUX. *International Journal of Human–Computer Interaction* 34(12): 1148–1156.

Lewis, J. R., B.S. Utesch, and D.E. Maher. (2015). Investigating the Correspondence Between UMUX-LITE and SUS Scores. In I*nternational Conference of Design, User Experience, and Usability* (pp. 204–211). 4th International Conference, DUXU 2015, Held as Part of HCI International 2015. Los Angeles, CA, USA, August 2–7, 2015.

—. (2013). UMUX-LITE: When There's No Time for the SUS. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099–2102). Paris, France, April 27–May 2. Retrieved March 1, 2019 from https://dl.acm.org/doi/10.1145/2470654.2481287.

McLellan, S., A. Muddimer, and S.C. Peres. (2012). The Effect of Experience on System Usability Scale Ratings. *Journal of Usability Studies* 7(2): 56–67.

National Center on Disability and Journalism. (2018). *Disability Language Style Guide.* Phoenix, AZ: Arizona State University, Walther Cronkite School of Journalism and Mass Communication. Retrieved from https://ncdj.org/style-guide/.

Newhouse, W., B. Johnson, S. Kinling, B. Mulugeta, and K. Sandlin. (2018). *Multifactor Authentication for E-Commerce Risk-Based, FIDO Universal Second Factor Implementations for Purchasers*. Washington, DC: U.S. Department of Commerce, National Institute of Standards and Technology, National Cybersecurity Center of Excellence.

Office of Management and Budget (OMB). (n.d.) *President's Management Agenda*. Retrieved March 1, 2019, from https://www.performance.gov/PMA/Presidents_Management_Agenda.pdf.

Pew Research Center. (2018, February 5). *Mobile Fact Sheet.* Retrieved March 1, 2019, from Office of Management and Budget (OMB) Pew Research Center for Internet & Technology: http://www.pewinternet.org/fact-sheet/mobile/.

Roberts, A. W., S.U. Ogunwole, L. Blakeslee, and M.A. Rabe. (2018, October). *The Population 65 Years and Older in the United States: 2016*. Washington, DC: U.S. Census Bureau, American Community Survey Report ACS-38. Retrieved March 1, 2019, from https://www.census.gov/content/dam/Census/library/publications/2018/acs/ACS-38.pdf.

Ruoti, S., B. Roberts, and K. Seamons. (2015). Authentication Melee: A Usability Analysis of Seven Web Authentication Systems. In *Proceedings of the 24th International Conference on World Wide Web*, 916–926. Florence, Italy, May 18–22, 2015.

S.3050—21st Century IDEA. (n.d.). Retrieved March 1, 2019, from https://www.congress.gov/bill/115th-congress/senate-bill/3050/text#idcd4f3d9b09ee463ebfa6567e047573fe.

Sasse, M. A., and K. Krol. (2013). Usable Biometrics for an Ageing Population. In *Age Factors in Biometric Processing* (pp. 303-320). Stevenage, United Kingdom: Institute of Engineering and Technology (IET) Digital Library. Retrieved from https://digital-library.theiet.org/content/books/10.1049/pbsp010e_ch16.

Taylor, D. M. (2018, November). *Americans With Disabilities: 2014*. Washington, DC: U.S. Census Bureau, Current Population Report P-70–152. Retrieved March 1, 2019, from https://census.gov/library/publications/2018/demo/p70-152.html.

Theofanos, M., B. Stanton, and C.A. Wolfson. (2008, June 11). *Usability and Biometrics Ensuring Successful Biometric Systems*. Washington, DC: U.S. Department of Commerce, National Institute of Standards and Technology. Retrieved March 1, 2019, from https://www.nist.gov/publications/usability-and-biometrics-ensuring-successful-biometric-systems.

Trewin, S., C. Swart, L. Koved, J. Martino, K. Singh, and S. Ben-David. (2012). Biometric Authentication on a Mobile Device: A Study of User Effort, Error and Task Disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference*, 159–168. Orlando, Florida, USA, December 3–7, 2012.

U.S. Congress. House. Committee on Oversight and Government Reform. (2018, March 22). H.R. 5402-*Government Customer Service Accountability and Improvement Act of 2018*. Washington, DC: 115th Congress. Retrieved March 1, 2019, from https://www.congress.gov/bill/115th-congress/house-bill/5402.

U.S. General Services Administration (GSA). (2018, May). *Section 508.gov GSA Government-wide IT Accessibility Program*. Washington, DC: U.S. General Services Administration. Retrieved March 1, 2019, from https://www.section508.gov/about-us.

Web Accessibility Initiative (WAI). (n.d.) *Accessibility Principles*. Retrieved March 1, 2019, from https://www.w3.org/WAI/fundamentals/accessibility-principles/.

World Health Organization. (2017, November 21). *Grades of hearing impairment*. (World Health Organization) Retrieved March 29, 2019, from https://www.who.int/pbd/deafness/hearing_impairment_grades/en/.

—. (n.d.). *Change the Definition of Blindness*. Retrieved March 29, 2019, from https://www.who.int/blindness/Change%20the%20Definition%20of%20Blindness.pdf?ua=1.

Appendix A

This appendix presents additional details on usability performance results.

TABLE 4. Perceived usability results for all participant groups combined

Item	N	Min	Max	Median	Mean	Std Error	St Dev
PIN reqms	27	5	7	6	6.22	.154	.801
PIN ease	27	4	7	6	6.11	.180	.934
Finger reqms	27	2	7	7	6.44	.209	1.086
Finger ease	27	4	7	7	6.67	.141	.734
Eye reqms	25	1	7	6	5.00	.424	2.121
Eye ease	25	1	7	6	5.04	.456	2.282
Palm reqms	20	1	7	5	4.40	.483	2.162
Palm ease	20	1	7	3	3.30	.471	2.105

TABLE 5. Perceived usability results for the control participant group

Item	N	Min	Max	Median	Mean	Std Error	St Dev
PIN reqms	9	5	7	6	5.89	0.26	0.78
PIN ease	9	5	7	6	6.11	0.26	0.78
Finger reqms	9	2	7	7	6.11	0.54	1.62
Finger ease	9	4	7	7	6.56	0.34	1.01
Eye reqms	9	2	7	5	4.78	0.57	1.72
Eye ease	9	2	7	6	5.22	0.57	1.72
Palm reqms	6	3	7	5.5	5.33	0.56	1.37
Palm ease	6	1	7	3.5	3.67	0.88	2.16

TABLE 6. Perceived usability results for the hearing loss participant group

Item	N	Min	Max	Median	Mean	Std Error	St Dev
PIN reqms	11	5	7	6	6.2	0.26	0.87
PIN ease	11	4	7	6	6.0	0.36	1.18
Finger reqms	11	5	7	7	6.5	0.25	0.82
Finger ease	11	5	7	7	6.7	0.19	0.65
Eye reqms	9	3	7	7	6.3	0.44	1.32
Eye ease	9	1	7	7	6.0	0.67	2.00
Palm reqms	9	1	7	5	4.1	0.79	2.37
Palm ease	9	1	6	3	3.0	0.60	1.80

TABLE 7. Perceived usability results for the vision loss participant group

Item	N	Min	Max	Median	Mean	Std Error	St Dev
PIN reqms	7	6	7	7	6.7	0.18	0.5
PIN ease	7	5	7	6	6.3	0.29	0.8
Finger reqms	7	6	7	7	6.7	0.18	0.5
Finger ease	7	6	7	7	6.7	0.18	0.5
Eye reqms	7	1	7	4	3.6	0.97	2.6
Eye ease	7	1	7	3	3.6	1.04	2.8
Palm reqms	5	1	7	3	3.8	1.16	2.6
Palm ease	5	1	7	2	3.4	1.29	2.9

TABLE 8. Completion rate results for all participant groups and schemes

Scheme	Trial Group	N	Mean	Std Error	St Dev
PIN	All participants	54	1.00	0.0	0.0
	Control	18	1.00	0.0	0.0
	Hearing Loss	22	1.00	0.0	0.0
	Vision Loss	14	1.00	0.0	0.0
Fingerprint	All participants	53	1.00	0.0	0.0
	Control	18	1.00	0.0	0.0
	Hearing Loss	22	1.00	0.0	0.0
	Vision Loss	13	1.00	0.0	0.0
Eye	All participants	51	0.82	0.05	0.39
	Control	18	0.94	0.06	0.24
	Hearing Loss	18	0.89	0.08	0.32
	Vision Loss	14	0.57	0.14	0.51
Palm	All participants	40	0.70	0.07	0.46
	Control	12	0.75	0.13	0.45
	Hearing Loss	17	0.71	0.11	0.47
	Vision Loss	11	0.64	0.15	0.50

TABLE 9. Response time results (in seconds) from success trials for all participant groups and schemes

Scheme	Trial Group	N	Min	Max	Median	Mean	Std Error	St Dev
PIN	All participants	54	5.02	35.8	8.9	11.4	0.94	6.87
	Control	18	5.02	23.5	8.4	9.3	1.07	4.52
	Hearing Loss	22	5.37	24.5	9.1	10.2	0.98	4.59
	Vision Loss	14	5.73	35.8	15.2	16.0	2.69	10.04
Fingerprint	All participants	54	0.82	17.1	3.4	4.9	0.52	3.86
	Control	18	1.47	11.7	2.7	3.7	0.62	2.65
	Hearing Loss	22	0.82	17.1	3.6	5.5	0.94	4.40
	Vision Loss	14	1.20	12.3	3.5	5.4	1.11	4.17
Eye	All participants	42	1.74	108.7	7.7	13.7	3.10	20.12
	Control	17	1.74	56.4	6.7	9.3	3.01	12.41
	Hearing Loss	16	3.74	108.7	8.1	17.4	6.87	27.50
	Vision Loss	8	3.44	56.5	9.9	16.5	6.17	17.45
Palm	All participants	28	1.00	81.0	8.3	18.2	3.98	21.04
	Control	9	0.93	33.1	4.0	9.0	4.04	12.12
	Hearing Loss	13	3.48	57.1	12.1	22.5	5.55	20.01
	Vision Loss	6	2.30	81.0	8.2	22.9	12.68	31.07

Customer Experience Research Leads to Better Design and Increased Adoption

Nichole Kerber, Kristen Papa, and Jacob Sauser (Booz Allen Hamilton)

What is Customer Experience?

"Most Americans may not think about the Federal Government every day — but when they need Government services, they expect them to work."[1]

Customer experience is defined as how customers perceive their interactions with an organization (Forrester Research).[2] Customers and stakeholders alike expect to communicate with the IRS through a variety of service channels that accommodate their diverse set of needs, challenges, and preferences. Customer experience principles can be applied across service channels, from traditional face-to-face interactions to virtual and digital environments. While the IRS is working to optimize the customer experience across all service channels, the research presented in this paper is primarily focused on digital interactions, specifically the payment experience on IRS.gov.

One of the most common misconceptions about the phrase, customer experience, is that if an organization has one product or service that is generally received well by their customer or user base, they are "doing" and integrating experience principles. The mark of a great end-to-end experience is not about one great product or experience but, rather, it's about continuously deploying seamless interactions across people, processes, and technology platforms to establish an orchestrated ecosystem that meets customers' needs at the time they need it. Furthermore, it not only includes external consumers of a product or service, but also front-line employees, such as call center customer service representatives and IT help desk agents, who interact and deliver the experience to those customers. Through continuous listening, monitoring, and management, organizations can uncover gaps, identify opportunities, and design improvements that address the real needs of the people they serve. By engaging organizational leaders and employees, they can introduce and sustain transformation programs, creating a customer-centered culture.

With the definition and understanding of customer experience in mind, the primary objective of the IRS' customer experience research is to understand taxpayers' perceptions of the agency and ultimately how that perception translates into engagement and advocacy. Through a framework that focuses on people, process, and technology, the IRS seeks to uncover taxpayer needs, wants, and opinions to better understand the challenges they may face now and in the future. This is done by eliciting direct taxpayer feedback via phone and in-person interviews, Website and application data analysis, and cross-collaboration with departments to share research recommendations and insights. The end-goal is to provide critical information to the agency so that it can begin to optimize existing services, develop new services, and encourage adoption or migration of users between channels and other touchpoints.

One of the focal points of the IRS' customer experience practice framework requires customers and stakeholders to be involved throughout the research, design, and deployment processes. In that way, the capability is meant to go beyond understanding the customer to also recognizing operational needs and improvements. It aims to understand how the organization functions and processes work to get an accurate picture of the backend operations driving the experience.

---

[1] President's Management Council (2018).

[2] Manning, H. (2010).

Measuring Customer Experience (CX)

According to Forrester Research, Government agencies' average Customer Experience Index Score for digital channels has remained flat while the average score for nondigital channels has risen by three points since 2015. Customers consider their experiences with Federal digital channels to be ineffective, difficult, and emotionally negative. As it specifically relates to the IRS, the organization ranks near the bottom of Government agencies in terms of public perception and customer satisfaction. In the *2017 American Customer Satisfaction Index (ASCI) Federal Government Report*, the Department of Treasury received a score of 61, well below the Government average of 69.7 (Morgeson (2018)). In addition, Forrester's U.S. Federal CX Index, 2018: Rankings of U.S. Federal Government Agencies, ranks the IRS as "very poor" with a rank of 12 out of 15 (Parrish, *et al.*). The CX Index score measures how successfully an organization delivers customer experiences that create and sustain loyalty on a variety of factors that influence the customer experience on a scale from zero to 100. The private sector average score for CX is 69 whereas the Federal average score is 59.

Noting these scores underlines the importance of understanding how to measure the impact of CX on an organization. Even though the potential benefits differ by industry and organization, a concerted effort and focus on the needs of a customer base generally increases in three main areas called "The Three E's" (Forrester Research): [3]

1. Emotion—Is the product or service enjoyable or emotionally engaging so that people want to use them?
2. Effectiveness—Is the product or service useful as to deliver value?
3. Ease—Is the product or service easy to find and engage with?

In addition to "The Three Es," Forrester Research proposes that organizations also measure seven drivers of quality regardless of channel or touchpoint. These drivers, ranked in order of importance, include components of information seeking; respect; completing transactions and scheduling appointments across desktop and mobile channels; the ability to obtain benefits or services; plain language and communication outreach; providing payment methods and clearly outlining payment methods that customers like to use; and, offering services at times that are most convenient for customers. When these quality drivers, along with "The Three Es," are taken into consideration, when aligned to organizational Key Performance Indicators (KPIs), it almost guarantees that organizations like the IRS will see an increase in overall compliance, engagement, and advocacy.

Customer Experience in Tax Administration

Due to recently passed legislation and bills introduced to Congress, the IRS, along with other Government agencies, are now required to consider customer experience and satisfaction when delivering services and, to foster an organizational culture focused on providing customer experiences that a citizen can take advantage of regardless of location, task complexity or touchpoint.

> *"Government exists to serve citizens, and this bill ensures Government leverages available technology to provide the cohesive, user-friendly online service that people around this country expect and deserve,"*
>
> *—Rep. Ro Khanna*

President's Management Agenda (PMA)

On June 29, 2018, the Office of Management and Budget (OMB) introduced a new aspect of its Circular A-11 that instructs Government agencies to craft customer experience frameworks. The changes guide agencies on how to manage their customer experience efforts and provides sample Key Performance Indexes (KPIs) that

[3] Parrish, et al. (2018).

Government will need to report on annually starting in the first quarter of Fiscal Year (FY) 2019. As part of this circular, 14 Cross-Agency Priority (CAP) goals were established "to target those areas where multiple agencies must collaborate to elect change and report progress in a manner the public can easily track" (President's Management Council (2018)).

Several CAP goals tie directly back to the mission and services of the IRS. CAP Goal 4, "Improving Customer Experience with Federal Services," mandates that the Treasury Department and related agencies comply with the following objectives:

- Transform the customer experience by improving the usability and reliability of our Federal Government's most critical digital services;
- Create measurable improvements in customer satisfaction by using the principles and practices proven by leading private sector organizations;
- Increase trust in the Federal Government by improving the experience citizens and businesses have with Federal services whether online, in-person, or via phone, and
- Leverage technology to break down barriers and increase communication between Federal agencies and the citizens they serve.

21st Century Integrated Digital Experience Act

Passed in Congress on December 20, 2018, the 21st Century IDEA Act aims to increase efficiencies by promoting data-driven, secure, personalized, and mobile-friendly Websites (United States Congress (2018a)). The law establishes minimum standards for Federal Websites and encourages agencies to digitize manual processes and accelerate the usage of electronic signatures.

Taxpayer First Act of 2019

Passed in the House of Representatives on April 9, 2019, the Taxpayer First Act of 2019 is an amendment to the Internal Revenue Code of 1986 to modernize and improve the Internal Revenue Service (United States Congress (2018b)). In sum, this Act provisions that the IRS develop a comprehensive customer service strategy within 1 year of the bill passing, and requires the IRS to implement an Internet platform for Form 1099 filings, a fully automated program for disclosing taxpayer information for third-party income verification using the Internet, and uniform standards and procedures for the acceptance of electronic signatures.

How We Integrated Customer Experience Processes at the IRS—Our Methodology

To achieve the goal of integrating customer experience into the agency, we followed IDEO's Human-Centered Design (HCD) process (IDEO.org (2015)). The process, coupled with Forester Research CX measurements, created a working methodology that we could easily socialize and scale. At a high-level, the HCD process includes moving through three phases: Immersion, Ideation, and Implementation.

In the Immersion Phase, which provided foundational research through primary and secondary sources, we focused on:

- Gathering and analyzing previously conducted research.
- Conducting a gap analysis, listing topics and questions for further research.
- Writing a research plan and obtaining approvals.
- Recruiting taxpayers to engage in user research activities.
- Listening and learning directly from taxpayers.

In the Ideation Phase, which sought to synthesize insights from the Immersion Phase, we focused on:

- Synthesizing feedback and insights.
- Identifying potential areas of opportunity.

- Updating/establishing user personas and journey maps.
- Conceptualizing potential solutions.

Finally, in the Implementation Phase, we sought to design and test solutions to meet taxpayer needs.

To integrate the CX Framework into a known service within the broader agency, it was imperative to deeply understand the various audiences that the IRS serves. This included listing potential and current audience segments, their needs, wants, goals, concerns, and frustrations with the agency. After reviewing previously conducted research and coming up with a prioritization strategy with stakeholders, we began conducting one-on-one user interviews and usability testing sessions of key content sections on IRS.gov. In addition, we analyzed Website and application data usage via Google Analytics and ForeSee reports and reviewed publicly-available statistics and reports issued from the IRS that included tax compliance rates, the number of taxpayers who visit TAC/VITA Centers, and other similar pieces of information.

Since the IRS serves several different audience segments, we needed to come up with a coordinated prioritization strategy with IRS stakeholders to efficiently and effectively analyze customer needs and goals. As a result, individual taxpayers within the following categories were identified as priority audience segments for this initial research phase.

- Taxpayers who identify as having straightforward tax situations
- Taxpayers who identify as having complex tax situations
- Taxpayers who owe back taxes
- Low-income taxpayers
- Atypical taxpayers who file using nonstandard processes

With these key audience segments prioritized and identified, we enlisted the assistance of IRS stakeholders and product owners to help us identify customers to interview and engage in user research activities. A critical but often forgotten step in the human-centered design process is the ability to recruit quality participants who can provide feedback and insight on a variety of relevant topics. Without seeking direct qualitative feedback and observing users completing key tasks, an organization risks creating unusable products for their customers. Therefore, identifying and recruiting targeted audiences is key for all organizations, especially as existing products get updated or enhanced and new products and services are created.

Although recruiting individual taxpayers to participate in user research activities can sometimes prove challenging in the Federal Government, our team was able to establish a repeatable method of identifying and scheduling potential taxpayers to obtain direct feedback:

1. **Target**—In collaboration with IRS project stakeholders, we identified the audience segment and criteria for the research study.
2. **Review/Approve**—In cases when more than nine participants were needed, we submitted the appropriate forms and documentation to the Office of Management and Budget (OMB) in accordance with the Paperwork Reduction Act (PRA).
3. **Recruit**—Locate and recruit participants who were representative of the targeted audience.
4. **Screen**—To ensure taxpayers met predefined criterion, we screened potential candidates to confirm qualification.
5. **Schedule**—Once screened, we confirmed a date and time for the research session to be conducted.

To ensure balanced and unbiased feedback, Federal Government employees, including IRS stakeholders and staff, were not eligible to participate in research studies. To locate potential taxpayers, calls to participate were posted on online forums and social media platforms, including LinkedIn and Facebook. Additionally, taxpayers were found in-person through "guerilla" research techniques at local libraries, museum cafeterias, and on the lawn of the National Mall in Washington D.C.

Case Study: Optimizing IRS.gov To Pay Taxes

Through foundational analysis with surveys and one-on-one interviews with over 1,000 taxpayers via the Immersion Phase, several potential challenges to the customer, or "pain points," were uncovered with the current IRS.gov payment experience and areas of opportunity for improvement were identified. After conducting this prestudy analysis, we engaged with IRS leadership and key stakeholders to learn about the context of the wholistic payment journey a taxpayer may go through when setting up a payment agreement. From those insights, the following goal and research focus was established.

Business Goal

Allow individual taxpayers to setup or revise installment agreements from within their existing online account so that they can manage multiple aspects of their interactions with the IRS from one user experience, while simultaneously decreasing telephone and paper requests.

Research Focus

To identify and understand the needs, challenges, and opportunities related to integrating an Online Payment Agreement (OPA) into an Online Account (OLA).

To understand the journey from the perspectives of IRS customers and staff, we had to test our assumptions about their journeys by engaging them in their own context. During these activities, we developed a hypothesis about the customer journey and validated it in the field.

Hypothesis

Enabling individual taxpayers to easily make payments and/or set up a payment plan online instead of via paper or over the phone, which would also help deflect the number of calls to the IRS call center, would ultimately help increase overall tax compliance.

The IRS uses a variety of notices and letters to contact taxpayers when addressing tax related matters, including collection of unpaid tax debts. In FY2016 alone, 150,595,689 pieces of mail were issued to taxpayers to address a range of tax account issues.

Study Objectives

1. Understand the points of greatest anxiety in the taxpayer journey when paying the IRS.
2. Understand the mindset of taxpayers who owe money to the IRS and gain insight into what may be hindering them from making a payment and/or setting up a payment plan online.
3. Obtain taxpayer feedback and behavioral data on select payment scenarios.
4. Validate if taxpayers who owe money to the IRS understand the concept of existing IRS nomenclature terms and phrases.

To improve the digital experience of making payments to the IRS, key audience segments of individual taxpayers were targeted. Appropriately, this study correlated with two existing IRS Office of Online Services (OLS) Personas: Susan, the "Exasperated Ower" who owes back taxes (Figure A1 in the Appendix), and Stephan, the "Entrepreneur" who has complex taxes (Figure A2 in the Appendix). The main qualifying criteria for recruitment was finding respondents who had filed taxes within the last year and who had experience making payments to the IRS within the past 5 years.

Using Payment Research Insights To Build Personas and Customer Journey Maps

A persona is a fictional character who represents the qualities of an average user within an audience segment. Personas are not "made up;" they are discovered as a by-product of the investigative user research process. In

essence, personas are the voices of our customers when they are not in the room with us. An industry leader in user experiences, Nielsen Norman Group (NN/G) states:

> *"When based on user research, personas support user-centered design throughout a project's life cycle by making characteristics of key user segments more salient."* [4]

A persona is a singular user and highlights specific details and important features of a group. Personas are intended to be living, breathing documents, and as such will be updated based on new research findings and additional groups we speak with. Since we are continually conducting research, and collecting research from other departments throughout the IRS, established personas will evolve over time.

A customer-journey map tells the story of the customer's experience: from initial contact, through the process of engagement and into a long-term relationship. It may focus on a part of the story or give an overview of the entire experience. What it always does is identify key interactions that the customer has with the organization. It talks about the user's feelings, motivations, and questions for each of these touchpoints. It often provides a sense of the customer's greater motivation. What do they wish to achieve, and what are their expectations of the organization?

A customer-journey map takes many forms but typically appears as some type of infographic. Whatever its form, the goal is the same: to teach organizations more about their customers and identify opportunities to improve the customer's overall experience.

Taken together, personas and customer-journey maps are core artifacts that are a direct outcome of deploying a Customer Experience framework and methodology. From several rounds of one-on-one interviews and surveys with taxpayers held between March 2017 and the present day, we were able to validate and learn more about current persona assumptions for Susan and Stephan and uncover qualitative explanations behind the analytics. These insights were incorporated into the existing personas and served as the basis for establishing three payment-specific-journey maps that encapsulate user data in a memorable, shareable story.

For example, two surveys were conducted to collect payment topic-specific information between May and August 2018. Statistical significance was used to quantify uncertainty so that we could evaluate the variances. In total, approximately 1,000 responses to a mix of both closed and open-ended questions were collected. By gathering a large set of user information on a wide range of payment topics, terms and phrases, we identified gaps and opportunities where further indepth research would be needed. One of the main takeaways from the surveys confirmed that when owing money to the IRS, taxpayers consider how much they can pay (the total amount) first before thinking of time and form of payment (e.g. bank account, credit card, check, etc.) in that order. Taxpayer insights such as this enables the IRS to create user experiences that take into account their existing thought process of how something should work when conceptualizing new and intuitive digital interactions with the IRS.

Insights and data collected to understand the knowns and unknowns of the payment experience were also incorporated into an overarching taxpayer-journey map. Understanding the lifecycle and key phases of filing taxes, getting a refund, or making a payment allows organizational decision makers to pinpoint the intersection points of customer needs and business goals. As we documented customer goals, activities, pain points, and opportunities, we were able to identify four phases in the taxpayer payment experience: Discover, Assess, Establish, and Pay. These phases are meant to highlight the overarching end-to-end relationship of the various paths a taxpayer can take when going through the payment lifecycle. From these phases, three key payment-journey maps were set up based on distinct paths a taxpayer may take based on a specific decision point. As a result, three specific payment-journey maps established were to highlight the experience of paying taxes in full (Figure A3 in the Appendix); setting up a payment plan and establishing estimated taxes.

[4] Harley (2015).

TABLE 1. Four Phases of the Payment Process

1. Discover	2. Assess	3. Establish	4. Pay
Taxpayer discovers they owe money to the IRS.	Taxpayer assesses how they plan to address payment to the IRS.	Taxpayer determines best payment option for their situation and proceeds down the journey of Pay.	Taxpayer determines form of payment and completes payment. Based on taxpayer situation, may elect to pay taxes in full, set up a payment plan or establish estimated taxes.

Findings and recommendations from this study are ow being implemented to improve content on payments-related pages on IRS.gov. For example, multiple participants who were interviewed used the word "overwhelming" to describe the Pay Your Taxes by Debit or Credit Card page, causing uncertainty on which option to select in the payment process. To address the issue of content overload, OLS in partnership Online Engagement, Operations & Media (OEOM) is actively working on design enhancements to better display debit and credit card processor information on the page.

Limitations and Future Research

Our research has only just begun. Due to resource and time constraints, a number of important research topics came up in this process that were not within our scope but could prove valuable in understanding additional IRS customer journeys and needs.

Customer Research Segments. Although we have created research deliverables that align to the most prevalent segments (several individual taxpayer subgroups and tax professional subgroups), we have identified additional user groups that need more attention, including those from large businesses, tax-exempt organizations, government groups, and informational groups. The table below highlights the audience segments that have been identified; the ones marked with asterisks need more of our attention to better uncover their needs and ascertain how to address them.

TABLE 2. IRS Audience Segments Identified for Future Research

Audience Segment	Definition of Audience Segment
Individual taxpayers	Includes several subgroups, including routine fliers, taxpayers who owe back taxes, complex taxpayers, low-income taxpayers, atypical taxpayers, international taxpayers*, non-English speaking taxpayers*
Tax professionals	Includes certified public accountants, enrolled agents, tax attorneys, and tax return preparers.
Businesses	Includes small business, large business, and international business.
Tax-exempt*	Includes 503(c) charitable organizations and other not-for-profit organizations.
Government*	Includes Federal, State, local, foreign, and Indian tribal nations.
Informational*	Includes media, trade/lobby/advocacy, and policy advisors.

*Indicates audience segments needing most attention.

Online Chat. Most people, when needing to contact the IRS, currently do so by phone. Efforts are currently underway to provide taxpayers with the ability to contact the IRS via a live chat feature integrated into IRS.gov. This feature could greatly improve the overall taxpayer experience by eliminating the need to stay on hold (often for multiple hours) to reach someone and also providing a written record of correspondence that a taxpayer can refer back to. Online Chat could also result in some cost savings for the agency, as customer service representatives could field multiple chats at a time (as opposed to just one phone call at a time), reducing staffing needs at IRS call centers.

Digital Notices. The IRS currently sends information to taxpayers in the form of paper notices that are sent through the mail via the U.S. Postal Service. Efforts have been ongoing at the agency to modernize and partially digitize these communications both to save money and improve the customer experience for those who receive these notices. Research will play a vital role in this process by ensuring that taxpayer feedback is incorporated into building a service that allows taxpayers to opt into receiving and viewing digital notices.

Conclusion

Regardless of organization or industry vertical, customer experience research is a constantly evolving, never ending endeavor. Our team has only begun to scratch the surface of all the customer interactions, user groups, touchpoints, and digital/nondigital product types that the IRS offers. Taken together, insights that we've already uncovered and insights that we will uncover in the future can and should be used across the agency to improve the customer experience for all current and potential IRS customer groups. Even teams that do not work directly on public-facing products for the agency should leverage these insights to ask themselves how what they are doing might impact the public and how they can ensure that those insights adhere to the Three Es and seven drivers of CX quality. As insights are gathered and relevant documentation is updated, product teams that review and engage in conversations often will ensure that both the taxpayer and the IRS organization benefit.

References

Harley, A. (2015). Personas Make Users Memorable for Product Team Members. Nielsen Norman Group. Retrieved from https://www.nngroup.com/articles/persona/.

IDEO.org. (2015). The Field Guide to Human-Centered Design.

Manning, H. (2010). Customer Experience Defined. Forrester Research. Retrieved from https://go.forrester.com/blogs/definition-of-customer-experience/.

Morgeson, F. (2018). ASCI: Citizen Satisfaction Improves for Federal Government—Except with Republicans. American Consumer Satisfaction Association. Retrieved from https://m.theacsi.org/news-and-resources/press-releases/press-archive/press-release-federal-government-2017.

Parrish, R., M. Rodriguez, H. Manning, W. Willsea, and R. Birrell. (2018) The U.S. Federal Customer Experience Index, 2018. Forrester Research. Retrieved from https://www.forrester.com/report/The+US+Federal+Customer+Experience+Index+2018/-/E-RES142378.

President's Management Council. (2018). President's Management Agenda. Retrieved from https://www.whitehouse.gov/wp-content/uploads/2018/03/Presidents-Management-Agenda.pdf.

United States Congress. (2018a). 21st Century Integrated Digital Experience Act. (H.R. 5759). Retrieved from https://www.congress.gov/bill/115th-congress/house-bill/5759/text.

United States Congress. (2018b). Taxpayer First Act. (H.R.R. 5444). Retrieved from https://www.congress.gov/bill/115th-congress/house-bill/5444.

Appendix

FIGURE A1. Persona—Susan

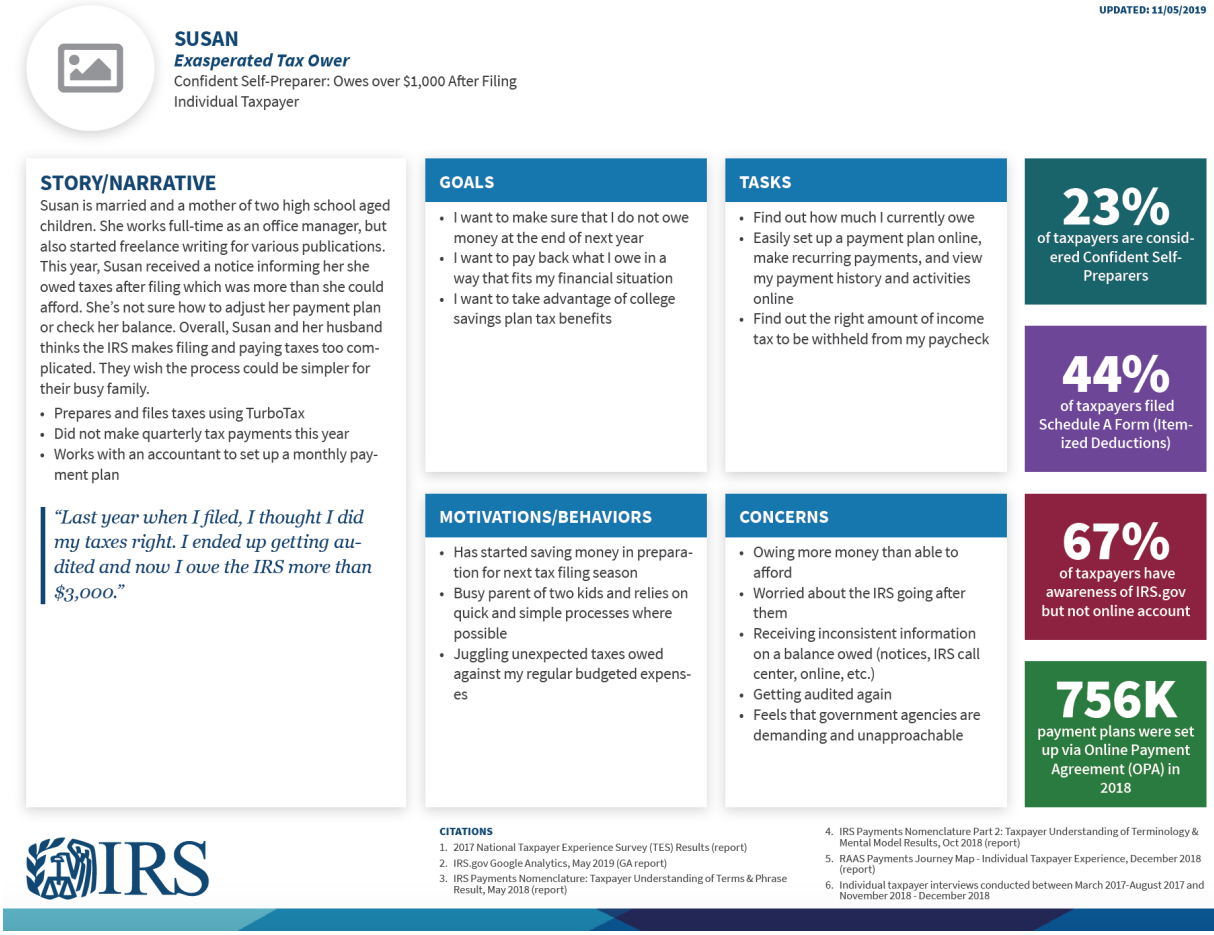
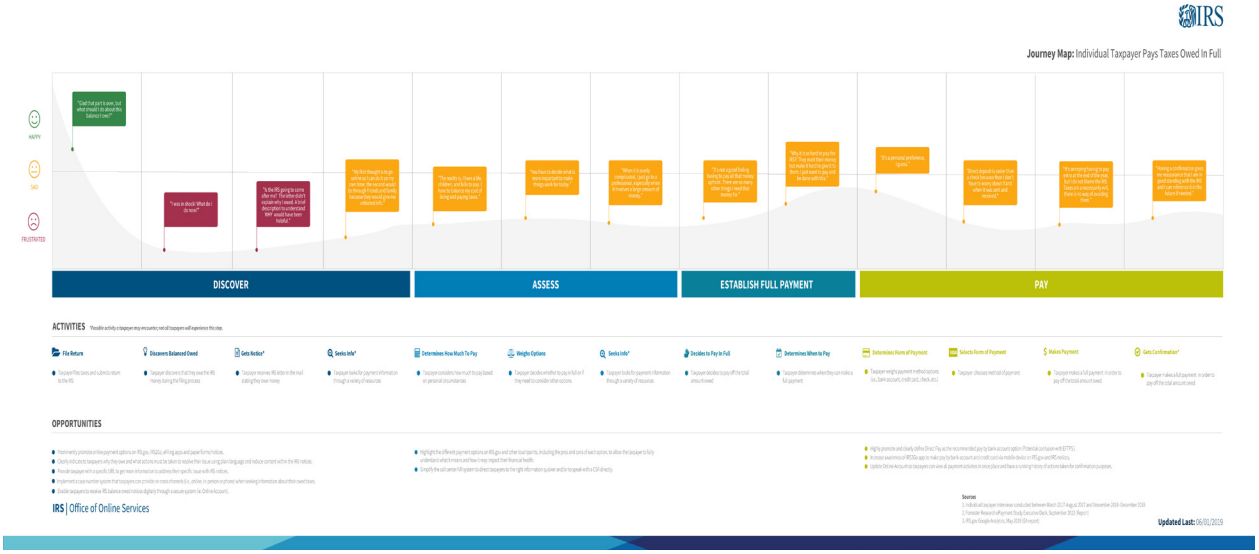


FIGURE A2. Persona—Stephan

FIGURE A3. Journey Map: Pay IRS Bill in Full



4

Understanding the Drivers of Taxpayer Behavior

Bryant ◆ Collins ◆ Li ◆ Miller ◆ Turk
Wind ◆ Orlett

Hageman ◆ LaMothe ◆ Marshall

Horvath ◆ Tikekar

Underpayment of Estimated Tax: Understanding the Penalized Individual Taxpayer Population

Victoria Bryant, Brett Collins, Janet Li, Alicia Miller, Alex Turk, and Tomás Wind (IRS, Research, Applied Analytics, and Statistics), and Stacy Orlett (IRS, Small Business/Self-Employed Division)

1. Introduction

Prepaying income taxes throughout the year is not just a requirement, it helps taxpayers meet their annual tax obligations and is also important for supporting the fiscal planning of the U.S. Treasury. The estimated tax penalty is assessed on a Federal income tax return when a taxpayer with certain levels of tax liability does not submit sufficient prepayments by the prepayment deadlines before filing.

In this paper, we present an analysis of individual tax filers who incur estimated tax penalties and attempt to answer several questions about this penalized population. How prevalent is this problem, and is it growing or shrinking? What types of taxpayers incur the estimated tax penalty? What taxpayer characteristics, including types of income earned and demographic variables like age, filer type, and income group, are linked to the penalty? Is the penalty associated with other types of noncompliance, and what happens after a taxpayer gets the penalty: do they change their prepayment behavior? The paper concludes with an introduction of proposed research to improve understanding of underlying behavioral drivers among penalized taxpayers, which could be engaged to improve future compliance.

2. Background: Prepaying Taxes and the Estimated Tax Penalty

The United States' Federal income tax system is a pay-as-you-go system. This means that taxpayers need to prepay this tax throughout the year as they receive income, rather than only paying their taxes when they file their return. Withholding and estimated tax payments are the two types of prepayment options available to taxpayers. The IRS instructs that taxpayers who expect to owe at least $1,000 in taxes after credits should make estimated tax payments (IRS (2018)).

Withholding allows a taxpayer to set a portion of their income to be diverted from their paycheck for income taxes and have it sent directly to the IRS by (in most cases) their employer or pension administrator. Withholding is mandated on wage or salary income (set up via Form W-4) and on some retirement income (set up via Form W-4P).[1] Withholding on Form W-4 is currently done through the claiming of withholding allowances generally aligning with taxpayer exemptions and deductions.[2] Withholding on Form W-4P is also done in this way for periodic payments but is set as a flat rate for nonperiodic payments and rollover distributions.[3] If a taxpayer does not fill out one of these forms or submits one with an invalid Taxpayer Identification Number, withholding is mandatory and is set at a rate for a default tax filer situation.[4] Taxpayers can opt into withholding for taxable Social Security benefits and unemployment insurance through Form W-4V, which

[1] For retirement income, taxpayers can opt out of withholding for periodic payments (e.g., periodic payments of pensions and annuities) and nonperiodic payments (e.g., Individual Retirement Account distributions that are payable on demand) but cannot opt out for eligible rollover distributions (e.g., 401(k) plans and 457(b) plans) or for any payments delivered outside the U.S. See 26 U.S.C. § 3405.

[2] Form W-4 is undergoing a redesign to be implemented starting in Tax Year 2020 which forgoes withholding allowances in lieu of calculation of the total withholding amount by the IRS (IRS (2019a)).

[3] Nonperiodic payments and eligible rollover distributions are withheld at a flat 10 and 20 percent rate, respectively, without the option to claim allowances but with the option to withhold an extra flat amount. See 26 U.S.C. § 3405.

[4] Without a Form W-4, wage or salary income is withheld at the rate for a single taxpayer with zero allowances. For periodic payments, withholding defaults to the rate for a married taxpayer with three withholding allowances if no W-4P is submitted and to the rate for a single taxpayer with zero allowances if an incorrect Social Security Number is submitted. See 26 U.S.C. § 3401(e) and Treas. Reg. § 31.3402(f)(2)–1(e).

provides several flat-rate options, or for sick pay paid by a third party (such as an insurance company) through Form W-4S, which calculates a flat level of withholding.[5]

Withholding places the burden of payment on the provider of income, e.g., the employer or pension administrator, instead of on the income recipient. Taxpayers need to file new W-4 forms if they start new jobs, but are not prevented from making adjustments during the course of employment. Taxpayers can change their allowances or ask for additional amounts to be withheld and may choose to do so when they experience life transitions, such as getting married or divorced or becoming eligible for new credits or deductions, which can affect the amount they need to pay in taxes. Withholding provides the lowest overall burden on the taxpayer and is associated with higher tax compliance (IRS (2016)).

TABLE 1. Prevalence of All Income Types by Ability To Withhold, TY 2017

Income type	Withholding	Number of returns* (millions)	Number of returns with estimated tax penalties* (millions)
Wages	Default	124.17	6.42
Interest	None	43.58	5.04
Dividends	None	27.79	3.61
Pension or annuity	Default	27.70	2.87
Capital gains	None	25.24	2.68
Self-employment (Schedule C)	None	25.21	3.67
Social Security benefits	Opt-in	20.53	2.94
Rental real estate, royalties, partnership, S corporation, estate, or trust (Schedule E)	None	17.10	2.64
Individual Retirement Account distributions	Default	14.87	1.93
Other income**	None	8.40	1.04
Unemployment compensation	Opt-in	5.10	0.23
Farming	None	1.77	0.07
Alimony***	None	0.45	0.07

*Income types are nonexclusive, i.e., a tax return with both wage and interest income is counted in both rows.
**Other income is income listed on line 21 of Form 1040.
***Alimony received will no longer be counted towards income for divorces or separations that are finalized after December 31, 2018. See 26 U.S.C. § 71.
SOURCE: IRS Compliance Data Warehouse, August 2019.

Estimated tax payments are prepayments of tax that a taxpayer must calculate and submit on their own for tax liability generated by income not subject to withholding, or if a taxpayer's withholding is insufficient. As listed in Table 1, this includes (in order of prevalence) interest and dividends (filed on Schedule B); capital gains (filed on Schedule D); self-employment income (filed on Schedule C); rental real estate, royalties, partnerships, S corporations, estates, trusts (all filed on Schedule E); "other" income (reported on line 21 of Form 1040); alimony;[6] and farm income (filed on Schedule F). Estimated tax payments must be paid to the IRS roughly quarterly throughout a calendar year (on April 15, June 15, September 15, and January 15 of the following year).[7]

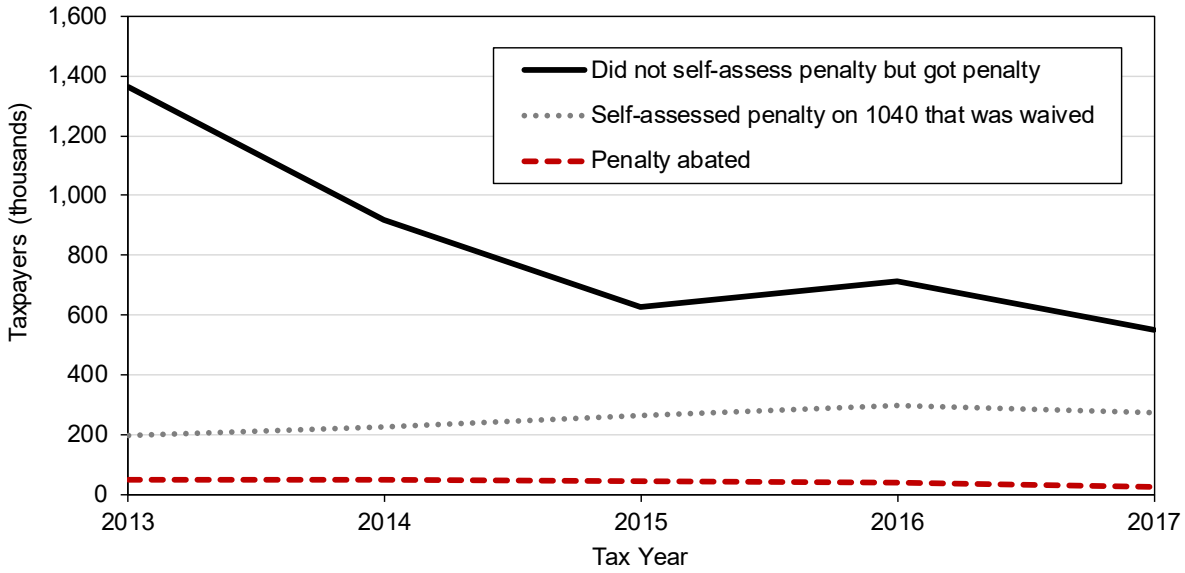
[5] Form W-4V flat withholding rate choices are 7, 10, 12, and 22 percent. For unemployment insurance, 10 percent is the only available withholding rate. See 26 U.S.C. § 3402(o).

[6] Payments made to a spouse or former spouse under a divorce or separation instrument (including a divorce decree, a separate maintenance decree, or a written separation agreement) may be alimony for Federal tax purposes. For divorce or separation agreements settled after December 31, 2018, alimony will no longer be included under gross income. See 26 U.S.C. § 71.

[7] Taxpayers may also submit Form 2210 to follow a fiscal year schedule for paying estimated taxes or to waive penalties for certain quarters if they are earning income seasonally. See 26 U.S.C. § 1.6655.

For most taxpayers, the estimated tax penalty, also called the underpayment of estimated tax penalty, is calculated for each quarter a taxpayer's total prepayments (withholding + estimated tax payments) sum to less than 22.5 percent of their tax liability for the current year; in total, the prepayments must sum to at least 90 percent of the current year's tax liability or a "safe harbor" of 100 percent of the previous year's liability (110 percent for higher-income taxpayers) for the taxpayer to avoid the penalty (26 U.S. Code § 6654).[8] The estimated tax penalty is assessed at a debit interest rate of 3 percent of outstanding tax plus the Federal short-term rate, which is determined by the Secretary of the Treasury at the beginning of each quarter (26 U.S. Code § 6621).

FIGURE 1. Number of Tax Returns in Special Estimated Tax Penalty Situations, TYs 2013–2017



SOURCE: IRS Compliance Data Warehouse, April 2019.

The estimated tax penalty is the only penalty for individual taxpayers that deals with prefiling tax behavior. At close to 10 million assessments a year, it is the most common IRS penalty by number of filers charged and the second most common IRS penalty by number of assessments, after the failure-to-pay penalty (IRS (2019b)). It makes up almost a quarter of all penalties assessed; most are self-assessed and are incurred by individual taxpayers, rather than business taxpayers (Fiscal Year 2018). The average penalty size is around $160 and, in total, about 1.5 billion dollars is assessed per year, with close to 90 percent actually collected by the IRS in Fiscal Year 2016. The percentage of returns that are assessed an estimated tax penalty has remained relatively consistent over the last 5 years, at around 6.5 percent in Tax Year (TY) 2017, compared to a high of 6.8 percent in TY 2014. A nontrivial number of returns were assessed a penalty that the filer did not self-assess upon filing, but this number has dropped significantly since 2013, as shown in Figure 1. The number of penalties waived or abated has remained small and relatively consistent during this timeframe. Penalties can be waived (through Form 2210) for reasonable causes such as a casualty event, bankruptcy, natural disaster, retirement, or new disability, and can be abated (fully or partially) in special situations.

While the majority (59 percent) of taxpayers with an estimated tax penalty are withholding, this only reflects the fact that the majority of taxpayers overall report income from withholdable sources, or sources subject to withholding. As noted in Table 2, only 4 percent of solely withholding taxpayers incurred an estimated tax penalty in Tax Year 2016. On the other hand, over 25 percent of taxpayers who were making solely estimated tax payments or were withholding and making estimated tax payments incurred a penalty. Among penalized taxpayers, it is more common to make no prepayments than to make both types of prepayments, or

[8] There are lower thresholds for farmers and fishers.

to solely make estimated tax payments. The majority of taxpayers who make no prepayments remain unpenalized because their income and credits result in tax liabilities lower than $1,000.

TABLE 2. Prepayment Distribution Among Tax Returns with an Estimated Tax Penalty, TY 2016

Prepayment type	N (thousands)	Percent of all penalized returns	Percent of all filers making these prepayments who are penalized
Returns with an estimated tax penalty, TY 2016	9,686	100.0	6.4
Withholding only	5,686	58.7	4.4
Estimated tax payments only	647	6.7	27.3
Both	1,392	14.4	25.0
No withholding or estimated tax payments	1,961	20.2	12.9

SOURCE: IRS Compliance Data Warehouse, September 2019.

3. The Estimated Tax Penalty and IRS Collection

Now that we have covered the basics of the estimated tax penalty, including when it is assessed, overall trends, and trends in tax prepayment methods, we investigate what happens after individual filers are penalized. Specifically, we explore the relationship between insufficient prepayment of tax and other compliance issues, including discrepancies in reporting, filing late, paying late, general payment noncompliance, and other penalties.

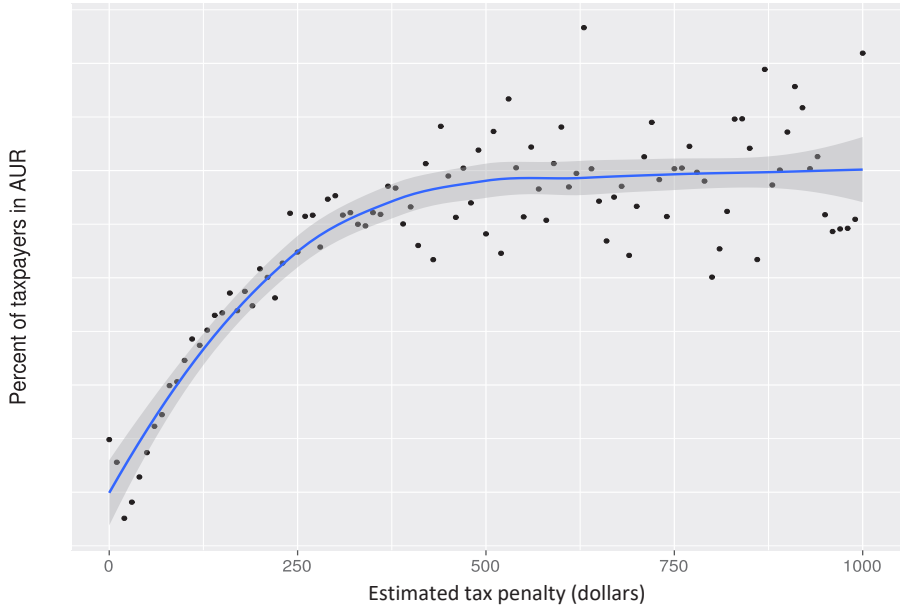
A. Late Filing

Examining filing times shows that taxpayers with an estimated tax penalty are more likely to request an extension and/or file late. While returns with an estimated tax penalty make up 6.4 percent of all returns, they make up 6.9 percent of returns that are filed late without an extension (TY 2017). More telling, they make up 15 percent of all returns that are filed late with an extension. Overall, 3.2 percent of filers with estimated tax penalties file late without an extension, compared to 2.8 percent among those without a penalty; 2.0 percent file late with an extension, compared to 0.7 percent among the nonpenalized (TY 2017). The percentage of filers filing late with an extension increases steadily with larger penalty sizes, whereas filing late without an extension occurs less with larger penalty sizes. Taxpayers filing an extension must still pay their taxes on time; although the taxpayer may file their return later, they will still be assessed a failure-to-pay penalty if they do not pay their taxes by the filing deadline.

B. Accurate Reporting

There is also evidence pointing to a positive relationship between insufficient prepayment of tax and underreporting of income. This is of critical concern for the IRS as underreporting of individual income is the single largest source of the U.S. Government's overall tax gap (IRS (2016)). The Automated Underreporter program (AUR) uses an algorithm to automatically flag filers whose income reported on their tax returns is lower than income reported through third-party sources, such as the W-2, 1099-MISC, 1099-K, etc. (IRS (2019d)). In total, in TY 2015, AUR flagged 36.6 percent more filers with estimated tax penalties than filers without. As seen in Figure 2, the percentage of returns with an estimated tax penalty flagged by AUR increases logarithmically as the size of the penalty increases.

FIGURE 2. Detection by Automated Underreporter System of Tax Returns with Estimated Tax Penalties by Size of Penalty, TY 2015



NOTE: Y-axis values suppressed for disclosure reasons.
SOURCE: IRS Compliance Data Warehouse, August 2019.

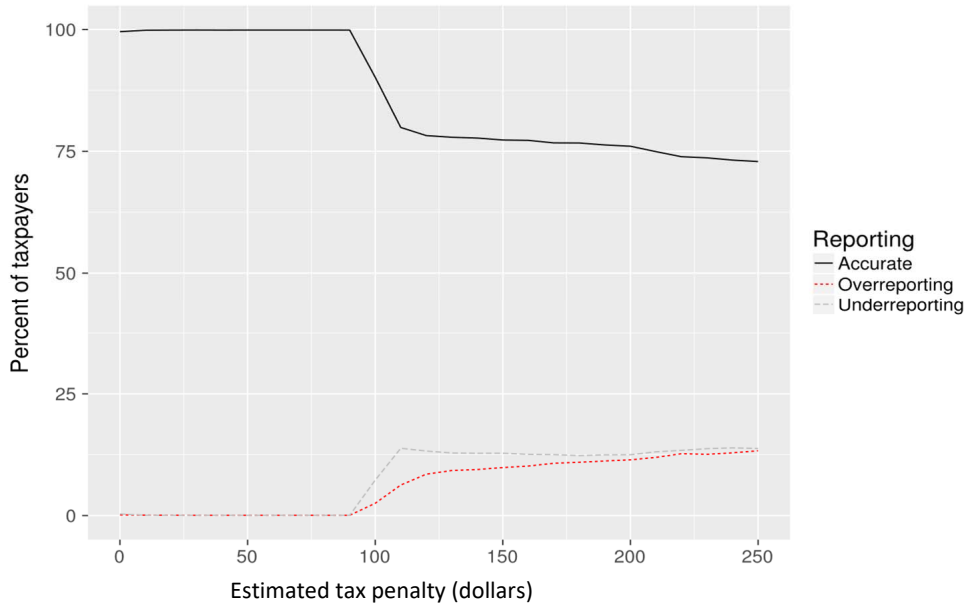
Table 3 shows that for almost every type of income, except wage income, partnership income, and Social Security benefits, AUR flags more underreporting among returns with estimated tax penalties than among returns without the penalty. The largest reporting differences between penalized and nonpenalized taxpayers are for self-employment income (66-percent difference), interest income (66-percent difference), Individual Retirement Account distributions (65-percent difference), and real estate income (61-percent difference). Conversely, more taxpayers without estimated tax penalties underreport wage income and Social Security benefits (a difference of 79 percent and 23 percent, respectively).

TABLE 3. Underreporting by Income Type and Estimated Tax Penalty, TY 2015

Reported income type	Percent difference in tax returns flagged by AUR (Returns with penalty—Returns without penalty)
Self-employment	66.3
Interest	66.0
Individual retirement account distributions	64.7
Real estate	61.0
Other income	56.0
Dividends	52.0
Pension/Annuities	47.3
Capital gains	43.0
Withholding	8.7
Unemployment compensation	7.2
Rental income	5.9
Partnership	-4.8
Social Security	-23.0
Wage	-79.0

NOTE: Percentages are based on the universe of tax returns with the corresponding income type but may vary depending on which income variables are used.
SOURCE: IRS Compliance Data Warehouse, August 2019.

FIGURE 3. Discrepancies in Self-Assessment of Estimated Tax Penalty by Size of Penalty, TY 2017



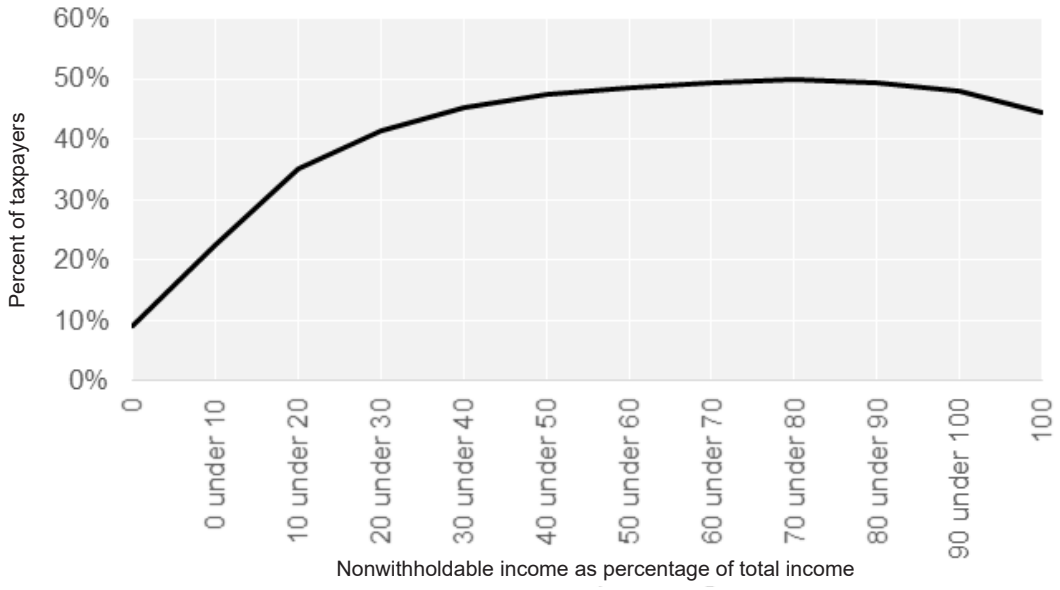
SOURCE: IRS Compliance Data Warehouse, August 2019.

In addition to the various incomes reported on a tax return, another area which can face inaccuracies is the calculation of the estimated tax penalty itself. Tax preparers and taxpayers typically self-assess estimated tax penalties on individual tax returns. As depicted in Figure 3, the vast majority of tax filers with estimated tax penalties correctly calculate and report their penalties. However, there are still significant levels of both undercalculating and overcalculating that seem directly related with the penalty size. The vast majority of estimated tax penalties are self-assessed accurately when they are under $100, which encompass approximately 69 percent of estimated tax penalties in TY 2017. Thirteen percent of taxpayers with an estimated tax penalty between $100 and $499 undercalculate their penalty, while 11 percent overcalculate penalty amounts. This trend switches for taxpayers with estimated tax penalties of at least $500. Here we observe slightly more overcalculating of the penalty.

C. Owing Taxes

Arguably, the most obvious outcome of insufficient prepayment of tax is having a balance due upon filing. The vast majority of taxpayers with an estimated tax penalty (91 percent) owe taxes upon filing, compared to just 15 percent of taxpayers without the penalty (TY 2017). The 9 percent of penalized taxpayers who file an even return or receive a refund when they file do not represent the norm: they incurred the estimated tax penalty for underpaying during one or more quarters but wound up correcting their prepayments at some point over the course of the year to end the year with no balance due. Among those with a balance due, penalized taxpayers owe on average almost $11,000, compared to just $4,000 for nonpenalized taxpayers.

FIGURE 4. Percent of Taxpayers with Balance Due Upon Filing, By Ratio of Nonwithholdable Income to Total Income, TY 2017



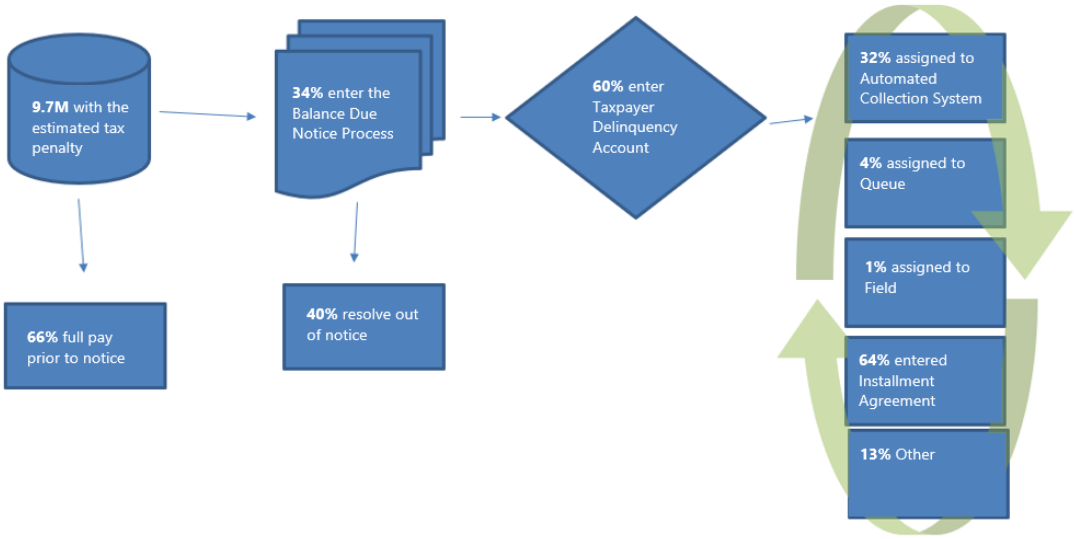
SOURCE: IRS Compliance Data Warehouse, August 2019.

While it may appear from these statistics that the estimated tax penalty is an indicator for higher balances due at filing, the real driver is perhaps more deeply rooted in the income situation of the taxpayer prior to ever incurring the penalty. As presented in Figure 4, this proportion of taxpayers with nonwithholdable income that have a balance due is rather consistent between 40 and 50 percent of all filers once nonwithholdable income reaches at least 20 to 30 percent of total income. This suggests that once a taxpayer's nonwithholdable income reaches even just a fifth of total income, the likelihood that she will report a balance due versus receive a refund are nearly even. The highest share reporting a balance due are among taxpayers who make 70 to 80 percent of their income from nonwithholdable sources. Among taxpayers who make 100 percent of their income from nonwithholdable sources, the proportion who owe a balance when they file drops slightly to 44 percent. While the underlying reasons for the drop are unclear, one explanation could be that this represents a less complex tax situation, as estimated tax payments serve as the only available prepayment option and the taxpayer does not need to be concerned about calculating both estimated tax payments and withholding. What is clear is that there is a positive correlation between having nonwithholdable income and reporting a balance due upon filing. Among taxpayers with solely withholdable income, only 9 percent report a balance due. Overall, there remains a strong relationship between incurring the estimated tax penalty and owing taxes when filing.

D. Collection Streams

What happens to taxpayers who do not resolve their tax balances? A large number of taxpayers incurring estimated tax penalties enter downstream workstreams of IRS collection processes. Figure 5 illustrates the flow of taxpayers post penalty assessment. We find that 66 percent of taxpayers who report a balance due after incurring an estimated tax penalty still fully pay their taxes upon filing. However, 34 percent of these balances due become unpaid assessments and enter IRS Collection's balance due notice process. This compares to just 6 percent of overall taxpayers who enter unpaid assessments. Moving downstream, the majority, or 60 percent, of unpaid assessments associated with estimated tax penalties become taxpayer delinquent accounts that may be assigned to stricter action streams for collection. Among all penalized taxpayers entering the balance due notice process, 64 percent enter installment agreements at some point in their lifecycle.

FIGURE 5. Issue Resolution for Taxpayers with an Estimated Tax Penalty, TY 2017



NOTE: TDA assignments are not mutually exclusive, i.e., a taxpayer can be assigned to multiple collection functions throughout their lifecycle.
SOURCE: IRS Compliance Data Warehouse, June 2019.

E. Additional Penalties

Many taxpayers who are charged estimated tax penalties also incur additional penalties. As detailed in Table 4, as many as 36.5 percent of taxpayers with estimated tax penalties are charged with other penalties in the same tax year. Failure-to-pay penalties make up the vast majority of these add-ons but not all: other concurrent penalties include failure to file, civil penalties, and bad check penalties (TY 2017).

TABLE 4. Other Penalties Among Taxpayers with Estimated Tax Penalties, TY 2017

Percent of all taxpayers with estimated tax penalties with another penalty	
Any other penalty	36.5
Failure to pay	34.4
Civil penalty	4.8
Failure to file	4.7
Bad check	0.8

SOURCE: IRS Compliance Data Warehouse, August 2019.

4. The Estimated Tax Penalty and Recidivism

We have seen that the estimated tax penalty has linkages to other compliance issues, including filing late, discrepancies in income reporting, owing taxes, unpaid assessments and other collection streams, and additional penalties. On top of these other compliance issues, we now explore recidivism of the estimated tax penalty itself.

Behavioral literature establishes that individuals are motivated by loss aversion, which would lead us to believe that people would change their behavior to avoid losses, including penalties (Kahneman and Tversky (1979); Engstrom *et al.* (2015)). We also know that penalties may be unavoidable due to other factors, such as income variability that can hinder a taxpayer's ability to correctly predict annual income and calculate

estimated payments if earning self-employment income. Nonetheless, in this section we explore whether estimated tax penalties work as deterrence against future noncompliance in tax prepayments. Specifically, what happens to an individual's behavior after they incur the estimated tax penalty—do they adjust their prepayment behavior or do they continue to incur the penalty?

Using the most recent full 6 years of data, we take tax returns with an estimated tax penalty in Tax Year 2012 and examine whether the filers incurred the penalty again in the following 5 years, between Tax Year 2013 and Tax Year 2017. As shown in Table 5, nearly a third (31 percent) of these returns with the penalty in TY 2012 did not get this penalty again in the following 5-year period. However, 69 percent were penalized again at least once, and 22 percent were penalized at least four more times.

TABLE 5. Estimated Tax Penalty Occurrence Following Estimated Tax Penalty in TY 2012

Penalty behavior	N (millions)	Percent of returns with estimated tax penalty in TY12
Penalty in TY12	7.81	
Never penalized again	2.40	31
Penalized at least once more	5.41	69
Penalized at least 2 more times	3.82	49
Penalized at least 3 more times	2.64	34
Penalized at least 4 more times	1.72	22
Penalized every year (TY12-TY17)	0.96	12

SOURCE: IRS Compliance Data Warehouse, July 2019.

TABLE 6. Prepayment Distribution and Prepayment Rates in TY 2017, Among Returns with an Estimated Tax Penalty in TY 2016

Returns with an estimated tax penalty, TY 2016	Withholding rate, TY16 (%)	ETP rate, TY16 (%)	Withholding only, TY17 (%)	WH rate (%)	ETP only, TY17 (%)	ETP rate (%)	Both, TY17 (%)	WH rate (%)	ETP rate (%)	Neither, TY17 (%)
Withholding only	8	-	89	10	1	22	8	9	6	3
Estimated tax payments only	-	19	4	9	72	41	11	4	15	13
Both	6	9	18	10	4	37	77	7	13	1
No withholding or estimated tax payments	-	-	15	17	8	52	1	4	10	75

SOURCE: IRS Compliance Data Warehouse, July 2019.

We now look at changes in prepaying behavior after incurring the estimated tax penalty. We separate out tax filers in Tax Year 2017 who incurred an estimated tax penalty in Tax Year 2016, based on the types of prepayments they made. As seen in Table 6, the majority continue the same prepayment behavior, with the most consistency among taxpayers who were solely withholding. Specifically, 89 percent continue to solely withhold the next year, and their withholding rate on average increases from 8 percent to 10 percent. Twenty-five percent of taxpayers who were not making any types of prepayments do start to withhold, make estimated tax payments, or do both in the next year. Taxpayers solely making estimated tax payments who were penalized increase their average estimated tax payment rate from 19 percent to 41 percent in the next year. The majority of those who were penalized, who were both withholding and making estimated tax payments, increase both types of prepayments but increase their estimated tax payment rates more on average. That some taxpayers do

self-correct suggests that the penalty might help motivate future compliance to avoid losses to at least some extent. However, it is still unclear whether the primary behavioral driver is related to the penalty or a nonpenalty motivator such as awareness of tax requirements or a desire to avoid a tax bill.

5. The Estimated Tax Penalty and Income Characteristics

The association between estimated tax penalties and other compliance issues highlights the importance of understanding this filing population and motivates future research on how best to encourage their compliance. As we have seen, estimated tax penalties are associated with such things as filing late, underreporting income, owing taxes when filing, incurring other penalties like the failure-to-pay and the failure-to-file penalty, and becoming unpaid assessments that impact IRS collection resources. This all builds the case that it is in the IRS' interest to improve prepayment compliance, in order to avoid more costly and burdensome resolution and potentially uncollectible statuses downstream. In this section, we explore the relationship between income factors and incurring the estimated tax penalty.

A. Prepayment Rates

We first examine the total level of tax prepayments made by taxpayers who incur estimated tax penalties and taxpayers who avoid penalties, as a proportion of their adjusted gross income (AGI). As shown in Table 7, the average level of tax prepayments as a rate of their AGI is considerably lower for taxpayers with estimated tax penalties than for those without a penalty. For example, the average nonpenalized taxpayer who prepays taxes through withholding pays a rate of 18 percent of their AGI, compared to a penalized taxpayer who only pays a rate of 8 percent of their AGI. This is not surprising, as those assessed a penalty have been identified as not paying sufficient taxes throughout the year. Nonpenalized taxpayers making estimated tax payments and no withholding pay a prepayment rate of 59 percent of their AGI. This is considerably higher than the highest tax bracket and may reflect the disproportionate effect of certain deductions and expenses on taxable income for taxpayers making less income; their prepayment rates on income before deductions and expenses would be lower than their prepayment rates after deductions and expenses are factored in.

TABLE 7. Prepayment Rates by Prepayment Type and Incurrence of the Estimated Tax Penalty, TY 2016

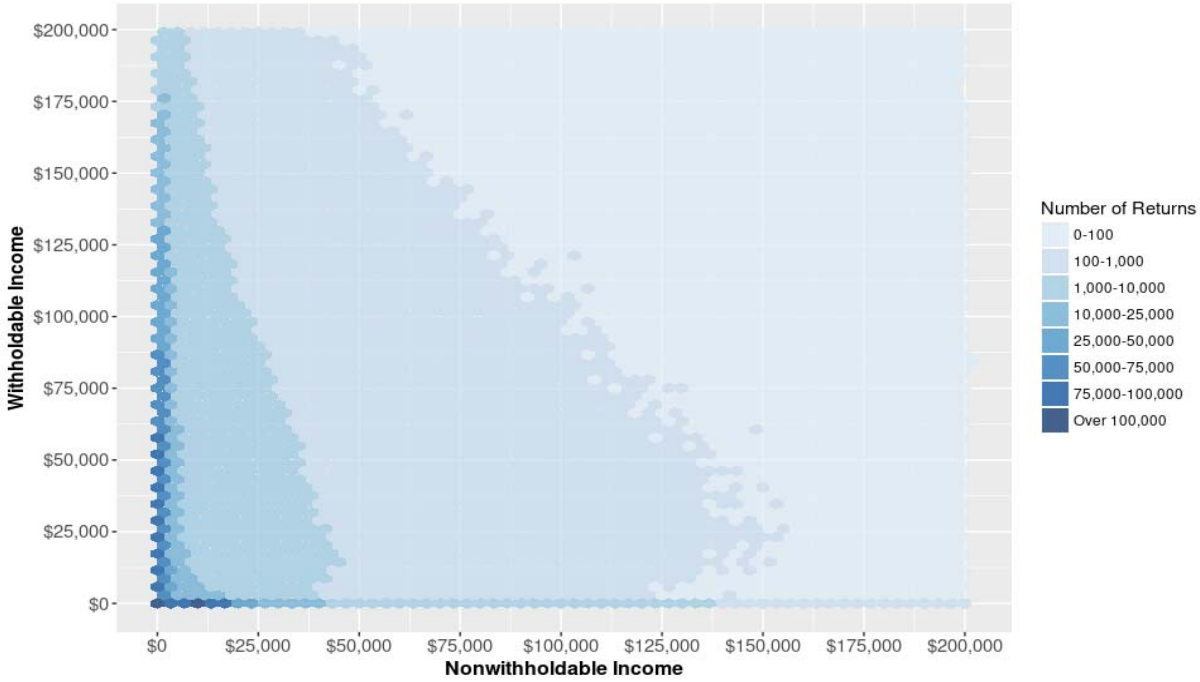
Type of tax prepayments made	Estimated tax penalty		No penalty	
	Percent of taxpayers making this prepayment type	Average prepayment rate (Prepayments over AGI)	Percent of taxpayers making this prepayment type	Average prepayment rate (Prepayments over AGI)
Withholding only	58.3%	8.0%	85.8%	17.9%
Estimated tax payments only	7.1%	19.1%	1.6%	59.0%
Both	15.0%	15.0%	3.7%	16.6%
No withholding or estimated tax payments	19.7%	-	8.9%	-

SOURCE: IRS Compliance Data Warehouse, August 2019.

B. Withholdable and Nonwithholdable Income

Understanding the factors behind why taxpayers incur estimated tax penalties requires understanding the tax prepayments they make, and this requires looking at the types of income they earn. As can be seen by the concentration of colors in Figure 6, the vast majority of taxpayers report earning all or a portion of their income through withholdable sources. In TY 2017, about 81 million people reported only withholdable income, while 8 million reported only nonwithholdable income, and 60 million reported a combination of the two.

FIGURE 6. Heat Map of Withholdable and Nonwithholdable Income Levels Among Population of All Taxpayers (TY 2017)

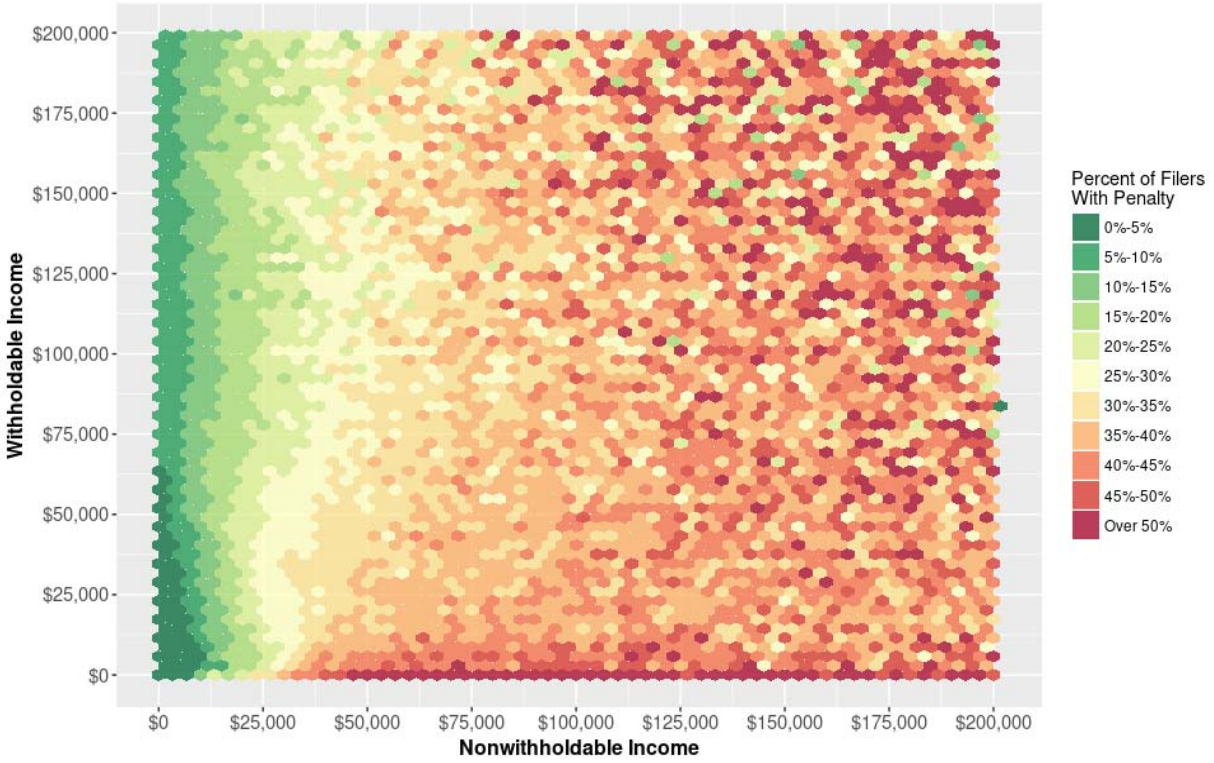


NOTE: Figure 6 is a hex-plot heat map, where each hexagon represents a set of tax returns with a certain level of withholdable and nonwithholdable income. Its color reflects the number of returns within that hexagon.
SOURCE: IRS Compliance Data Warehouse, April 2019.

While the majority (54 percent) of returns report solely withholdable income, the average level of reported income for this group is not as high as the average reported income for the minority (5 percent) of returns reporting solely nonwithholdable income: $42,000 and $59,000 respectively. When looking at the remaining 60 million returns (40 percent of all returns) that report a combination of withholdable and nonwithholdable income, the imbalance reverses, with an average of $89,000 withholdable income and $35,000 nonwithholdable income reported. Overall, levels of withholdable income tend to be larger than levels of nonwithholdable income, when ignoring income type splits.

Given the automated nature of withholding from the taxpayer's perspective, we expect that taxpayers having withholdable income sources are more likely to avoid estimated tax penalties. Building off Figure 6 to examine estimated tax penalty occurrence at varying levels of withholdable and nonwithholdable income, we confirm this expectation. The vertical gradients of the heat map in Figure 7 show that penalty occurrence increases with higher levels of nonwithholdable income. The percent of the population incurring a penalty increases from only 5 to 10 percent of taxpayers with around $10,000 in nonwithholdable income to closer to 30 percent of taxpayers with around $50,000 in nonwithholdable income. The bands of estimated tax penalty occurrence appear to be solely related to the size of nonwithholdable income; as we move up the withholdable income distribution but hold nonwithholdable income fixed, there is no correlated increase in penalty occurrence. This supports the reasonable conclusion that nonwithholdable income levels are a more important factor for incurring the estimated tax penalty than the amount of withholdable income.

FIGURE 7. Heat Map of Estimated Tax Penalty Occurrence by Withholdable and Nonwithholdable Income Levels, TY 2017



NOTE: Figure 7 is a hex-plot heat map, where each hexagon represents a set of tax returns with a certain level of withholdable and nonwithholdable income. Its color reflects the percentage of returns within that hexagon that are assessed estimated tax penalties.
SOURCE: IRS Compliance Data Warehouse, April 2019.

Examining specific types of withholdable and nonwithholdable income shows that the most common income situation among taxpayers with the estimated tax penalty is having solely Schedule C, or self-employment, income, as ranked in Table 8. About 8 percent of penalized taxpayers have this income situation, compared to 3 percent of all taxpayers, meaning their rate of penalty incurrence is much higher than average. Indeed, 19 percent of taxpayers with solely Schedule C income incur an estimated tax penalty, compared to just 7 percent of taxpayers overall. The next most common income situation is having solely wage income. About 8 percent of penalized taxpayers have this income situation; however, 42 percent of all taxpayers are in this situation, meaning the majority of taxpayers with solely wage income (99 percent, in fact) are in compliance and do not incur the penalty. Nine of the top ten most common income situations include a type of income that is withholdable. Eight of the top ten most common income situations involve more than one types of income, including both withholdable and nonwithholdable income. The most common nonwithholdable income types reported by a taxpayer earning primarily wage income are interest income and dividends.

TABLE 8. Ten Most Common Income Situations by Types of Income Among Taxpayers with an Estimated Tax Penalty, TY 2017

	Income situation	N (thousands)	% Rate of taxpayers in income situation with estimated tax penalty	% Share of all taxpayers with estimated tax penalty	% Share of all taxpayers
1	Schedule C only	809	19.4	8.2	2.7
2	Wage income only	803	1.3	8.2	41.6
3	Wage income and Schedule C	499	6.7	5.1	4.9
4	Wage income, interest, dividends, and Schedule D (capital gains)	229	8.2	2.3	1.8
5	Wage income, interest, dividends, Schedule D (capital gains), & Schedule E	227	18.4	2.3	0.8
6	Wage income and Schedule E	185	12.2	1.9	1.0
7	IRA, pension, Social Security, interest, dividends, and Schedule D (capital gains)	175	12.8	1.8	0.9
8	Wage income and interest	168	2.3	1.7	4.7
9	Wage income, interest, and Schedule C	151	10.7	1.5	0.9
10	Wage income, interest, and Schedule E	140	12.8	1.4	0.7

SOURCE: IRS Compliance Data Warehouse, July 2019.

C. Income Complexity

Now that we have established the connection between earning nonwithholdable income and incurring the estimated tax penalty, we look at the relationship between penalties and income complexity, or the number of income types. For simplicity, here we define income types by the lines with income reported on the 1040 return; for instance, a tax return may be associated with two W-2 forms, but we only count this as one type of income, wage income, despite the multiple sources. On average, taxpayers with estimated tax penalties have more types of income (e.g., wages, self-employment income, interest income, etc.) and have one more nonwithholdable type of income than the average taxpayer. The average number of income types for taxpayers with estimated tax penalties is 3.4 (TY 2017), compared to 2 income types for the overall filing population. Having multiple sources of income can make the accurate calculation of prepayments, both withholding and estimated tax payments, more difficult and thereby lead to higher risk of incurring the estimated tax penalty.

FIGURE 8A. Percent of All Filers with Estimated Tax Penalty by Number of Income Types, TY 2017

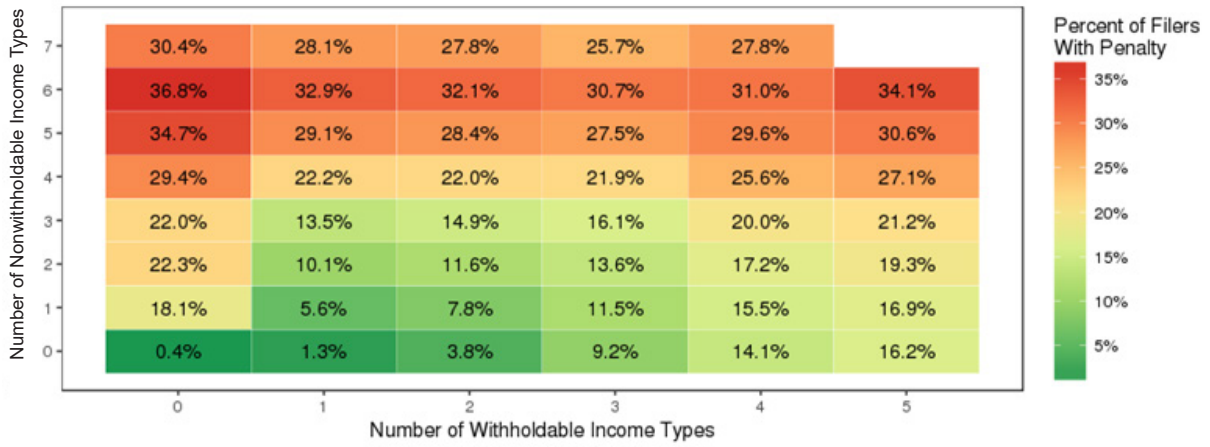
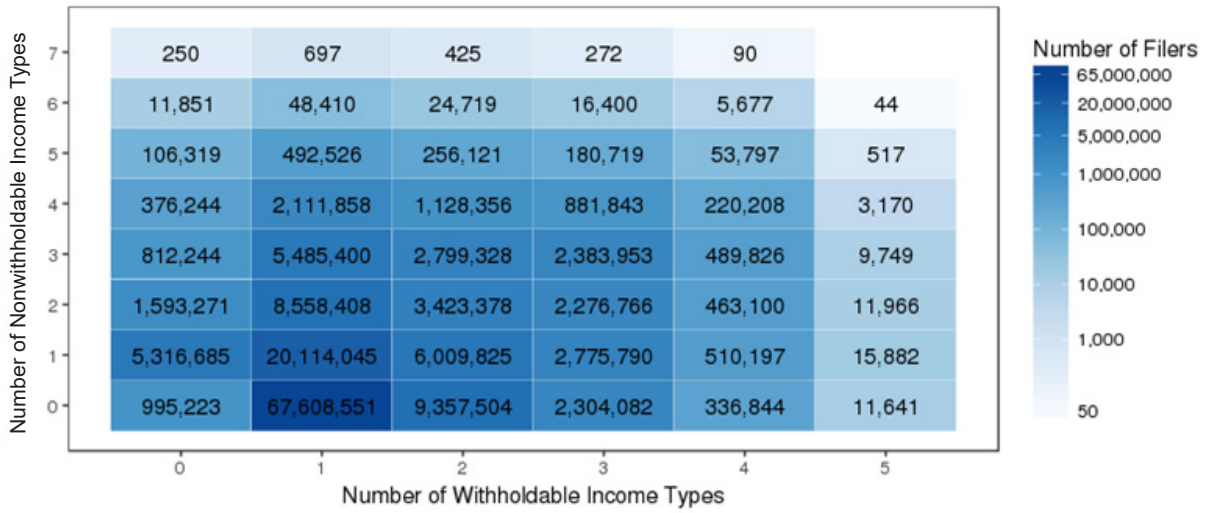


FIGURE 8B. Number of Filers by Number of Income Types, TY 2017



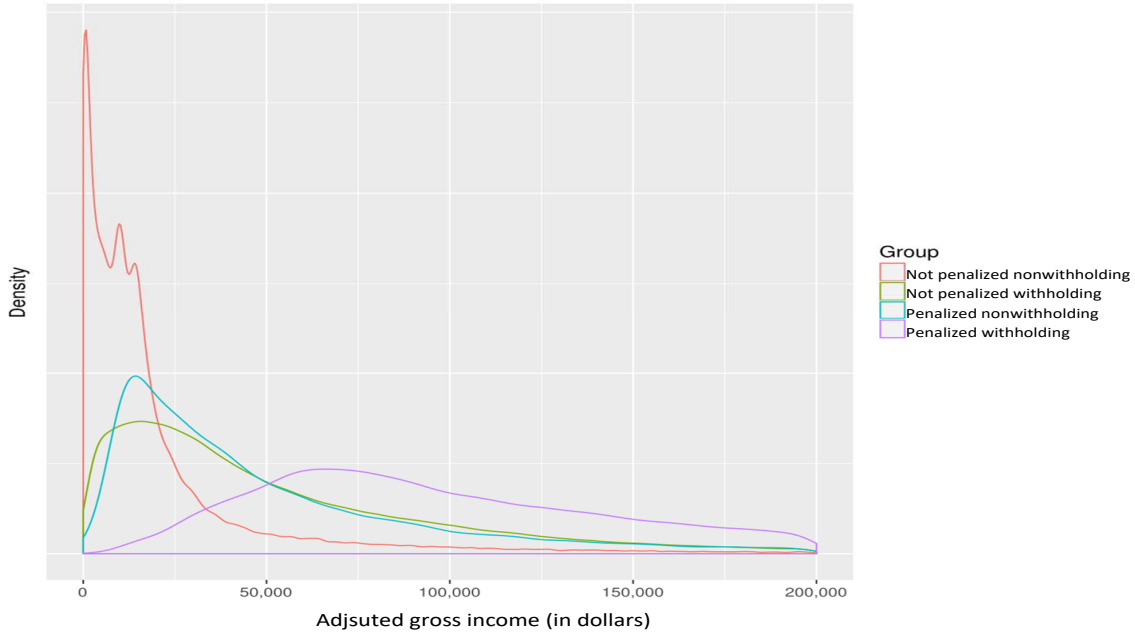
NOTE: An income type is counted if a taxpayer has positive income of that type. Taxpayers with zero income types may have zero or negative income (i.e., losses as reported on Schedules C, D, E, or F).
SOURCE: IRS Compliance Data Warehouse, May 2019. The numbers of filers are rounded to the nearest thousand or to the nearest hundred for cells under 1,000 filers.

As seen in Figure 8A, the earning of one additional type of nonwithholdable income increases penalty prevalence, regardless of withholdable income. Earning solely one type of withholdable income is the most common situation, shared by 68 million taxpayers, as seen in Figure 8B. With this group, penalty incidence, which is only 1.3 percent, increases exponentially with each addition of another type of nonwithholdable income, to over 30 percent for taxpayers with 6 types of nonwithholdable income. This same trend applies regardless of the number of types of withholdable income, although penalty occurrence increases most drastically for each additional nonwithholdable income type when the taxpayer does not have access to withholding. The relationship between complexity and penalty prevalence also holds true to a lesser degree for taxpayers without any nonwithholdable income types: the percentage of these taxpayers with a penalty increases steadily with each addition of another type of withholdable income, from 1 percent for those with 1 withholdable income type to 16 percent for those with 5.

D. Income Levels and Penalty Size

In addition to the types of income earned, how does a household's overall income level affect its probability for incurring an estimated tax penalty or the size of the penalty? The size of an estimated tax penalty assessment is directly related to the amount of outstanding tax a taxpayer owes. We expect that larger incomes, which are generally associated with larger tax liabilities, would be associated with larger penalties. For these taxpayers, minor discrepancies in setting withholding or calculating estimated tax payments could lead to larger underpayments. As established in Section 5B, the incurrence of estimated tax penalties is largely related to levels of nonwithholdable income, so we expect to see a difference between those who are withholding and those who are not. We thus separate these groups and look at how income levels correlate with whether someone is assessed a penalty and the size of the assessment.

FIGURE 9. Adjusted Gross Income of Estimated Tax Penalty Population by Withholding, TY 2017



SOURCE: IRS Compliance Data Warehouse, July 2019.

A distribution analysis shows that penalized taxpayers report more income on average than taxpayers who are not assessed estimated tax penalties. Excluding the top 5 percent and bottom 5 percent of the distribution, the average adjusted gross income (AGI) of a taxpayer with an estimated tax penalty is about $130,000 for those who are withholding and $72,000 for those who are not withholding (TY 2017), not accounting for the split in income types or sources. This is compared to an average AGI among nonpenalized taxpayers of about $66,000 for withholding taxpayers and $23,000 for nonwithholding taxpayers. Figure 9 shows the density distributions of the AGI of each of these groups.

What about the size of estimated tax penalties? Table 9 shows that average penalty amounts do increase for each income quartile, suggesting that there is a positive correlation between income and penalty size. However, estimated tax penalties overall are small, with about 69 percent of assessments under $100 (TY 2017). Further, penalties only marginally increase for withholding and nonwithholding taxpayers in the first three income quartiles, yet there is a marked increase in the average penalty in the top quartile. This nevertheless could be due to a few outliers among the wealthiest taxpayers.

TABLE 9. Estimated Tax Penalty Assessments by Withholding and Income Quartile, TY 2017

Adjusted gross income quartile	Withholding		Nonwithholding	
	Quartile thresholds	Average penalty	Quartile thresholds	Average penalty
1	Less than $67K	$46	Less than $20K	$41
2	$67K to $109K	$62	$20K to $40K	$70
3	$109K to $189K	$95	$40K to $83K	$119
4	Over $189K	$432	Over $83K	$498

SOURCE: IRS Compliance Data Warehouse, September 2019.

Overall, we see that taxpayers earning larger amounts of income not subject to withholding and more types of income have a higher occurrence of estimated tax penalties. Taxpayers with estimated tax penalties make lower prepayment rates, and taxpayers who do not withhold tend to be charged larger penalty assessments.

6. The Estimated Tax Penalty and Additional Demographic Factors

In this section, we move beyond income to examine a variety of other taxpayer demographic factors, such as filing status, socioeconomic status, age, and industry, to determine whether underlying demographics are associated with estimated tax penalties.

A. Filing Status

TABLE 10. Percent of Estimated Tax Penalties by Filing Status, TY 2017

Filing status	Percent with estimated tax penalty	Filing status of taxpayers with estimated tax penalties	Filing status of all filers
Single	4.5%	33.0%	45.6%
Married filing jointly	10.6%	58.7%	32.1%
Married filing separately	10.9%	3.6%	1.9%
Head of household	2.1%	4.6%	13.8%
Widower	4.6%	0.04%	0.05%
Married filing separately, spouse exemption	5.4%	0.02%	0.03%

SOURCE: IRS Compliance Data Warehouse, April 2019.

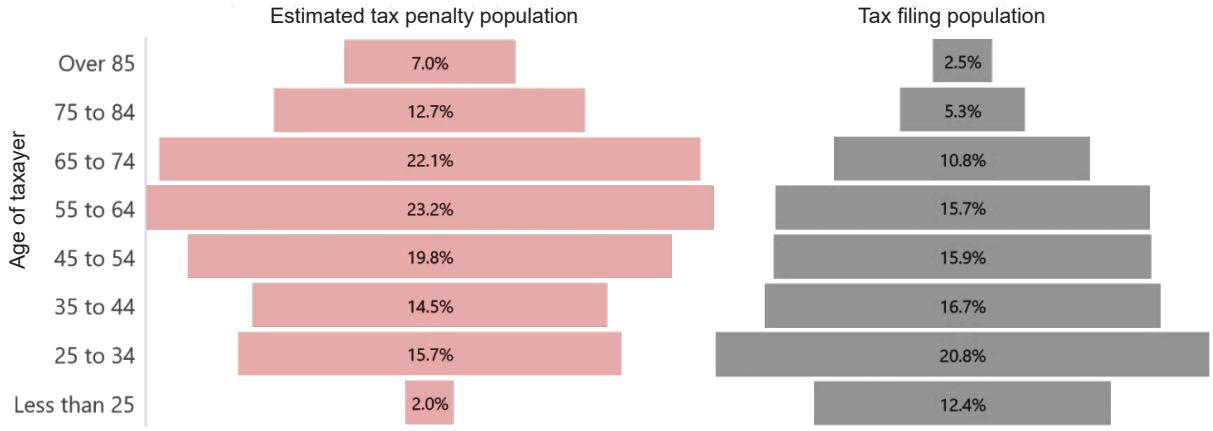
In Table 10, we break down estimated tax penalties by filing status. Married filers have a higher prevalence of the penalty than single filers. In fact, more than 10 percent of married filers, filing either jointly or separately, incur an estimated tax penalty, compared to less than 5 percent of single or widowed households. Individuals who file as a head of household have the lowest penalty prevalence at 2 percent. Note, however, that these rates do not control for income level and may reflect the high correlation between said level and penalty incurrence. It may be the case that income level is the main driver of the differences in rates.

B. Age

Taxpayers with an estimated tax penalty tend to be older than the average taxpayer, as seen in Figure 10. Almost half (42 percent) of tax returns with estimated tax penalties are filed by seniors aged 65 or older, although this demographic comprises only 19 percent of all individual taxpayer returns. Taxpayers younger than 35 comprise only 18 percent of the estimated tax penalty population, but 33 percent of overall taxpayers. Again, one explanation for the increased prevalence of the penalty with age could be the positive correlation between age and income and the positive correlation between income and incurrence of the penalty. Another explanation

could be the unique income situations of taxpayers who are retired from traditional wage or salary jobs. While retirement income is subject to a similar withholding process as wages and salaries, taxable Social Security benefits are not, and neither are other major sources of income for many retirees, like investment income.

FIGURE 10. Age Distribution of Taxpayers with an Estimated Tax Penalty, TY 2017



SOURCE: IRS Compliance Data Warehouse.

C. Industry

We investigate estimated tax penalties by industry among Schedule C filers, as reporting Schedule C income alone represents the largest single group in the penalty population and also the highest rate of penalty incurrence among any income type (as detailed previously in Table 8). Industry identification, through the North American Industry Classification System (NAICS) reported on the Schedule C, is challenging because NAICS codes are self-reported, resulting in some missing, incomplete, or inaccurate data. About 6.2 million (22 percent) of the 29.3 million Schedule C's filed in Tax Year 2017 had missing or invalid NAICS codes. Table 11 shows data for the remaining 78 percent of Schedule C filers. Nineteen percent of taxpayers with solely Schedule C income overall incur an estimated tax penalty. Finance and insurance, and real estate and rental and leasing, are the industries that see the highest prevalence of the estimated tax penalty.

TABLE 11. Top Five 2-digit NAICS Industries with Most Estimated Tax Penalties from Schedule C, TY 2017

Industry	NAICS (2-digit)	Estimated tax penalties (1000s)	Schedule C filers (1000s)	Percent with estimated tax penalty	Rank by size of industry
Finance and Insurance	52	188	659	28.6%	11
Real Estate and Rental and Leasing	53	392	1,376	28.5%	9
Professional, Scientific, and Technical Services	54	756	3,368	22.4%	2
Mining, Quarrying, and Oil and Gas Extraction	21	22	99	22.4%	19
Construction	23	479	2,170	22.1%	3

SOURCE: IRS Compliance Data Warehouse.

7. Comparing Estimated Tax Penalty Factors Through Regression Analysis

Thus far, we have ostensibly observed that estimated tax penalties are more highly correlated with married filers, higher-income households, older taxpayers, and self-employment in finance and real estate. However, due

to the collinear nature of many of these demographic variables, we cannot say for sure if any one relationship in isolation drives estimated tax penalties. To address this, we conducted a probit regression on a 1-percent sample of returns from Tax Year 2017 to better understand which variables have the highest associative relationship with incurrence of the estimated tax penalty. In the equation below, P represents incurrence of the estimated tax penalty, Φ represents the Cumulative Distribution Function of the standard normal distribution, and X represents a vector of regressors which include variables from the 1040 return: filing status, number of income types, age, adjusted gross income, and proportion of income coming from withholdable sources.

$$\Pr(P_i = 1|X_i) = \Phi(\beta_i X_i)$$

Table 12 reports our findings. Dummies for filing status are compared against filing as a single filer, and continuous regressors are standardized for better effect comparison.[9] The regression results reflect the findings of the descriptive demographic analyses in that, holding other regressors constant, there are still statistically significant correlations between incurrence of the estimated tax penalty and variables like filing status and age. In particular, filing separately as a married couple is associated with a higher probability of incurring the estimated tax penalty than filing as a single taxpayer, while the opposite is true for heads of households. Increasing income is associated with a higher probability of incurring the penalty, and so is age. Reporting a higher percentage of income from sources subject to withholding is linked to lower penalty incurrence, corroborating our earlier analysis. In fact, this seems to be the variable with the largest one-unit effect on the probability of incurring the penalty compared to the other variables included. The direction of these relationships are robust to the inclusion of other regressors. Overall, the estimated tax penalty is associated with a wide range of factors beyond income type, which can only explain so much of a taxpayer's prepayment behavior.

TABLE 12. Probit Model Coefficients

Regressor	Estimate	Standard Error	
(Intercept)	-9.4704	0.0310	***
Filing status married filing jointly	-0.0003	0.0046	ns
Filing status married filing separately	0.3985	0.0114	***
Filing status head of household	-0.3685	0.0083	***
Filing status widow	-0.4893	0.1043	***
Filing status married filing separately, spouse exemption	-0.0496	0.1250	ns
Proportion of total income subject to withholding (z-value)	-814.667	3.2023	***
Log of adjusted gross income† (z-value)	0.5636	0.0033	***
Age (z-value)	0.1241	0.0025	***
Number of income types (z-value)	0.0505	0.0022	***
Negative total income (dummy)	-1.6192	0.0313	***

NOTE: ***p<0.001; nsp>0.10; †Natural log of the absolute value of AGI; set to zero if AGI is zero.
SOURCE: IRS Compliance Data Warehouse, August 2019.

8. Estimated Tax Payment Behaviors

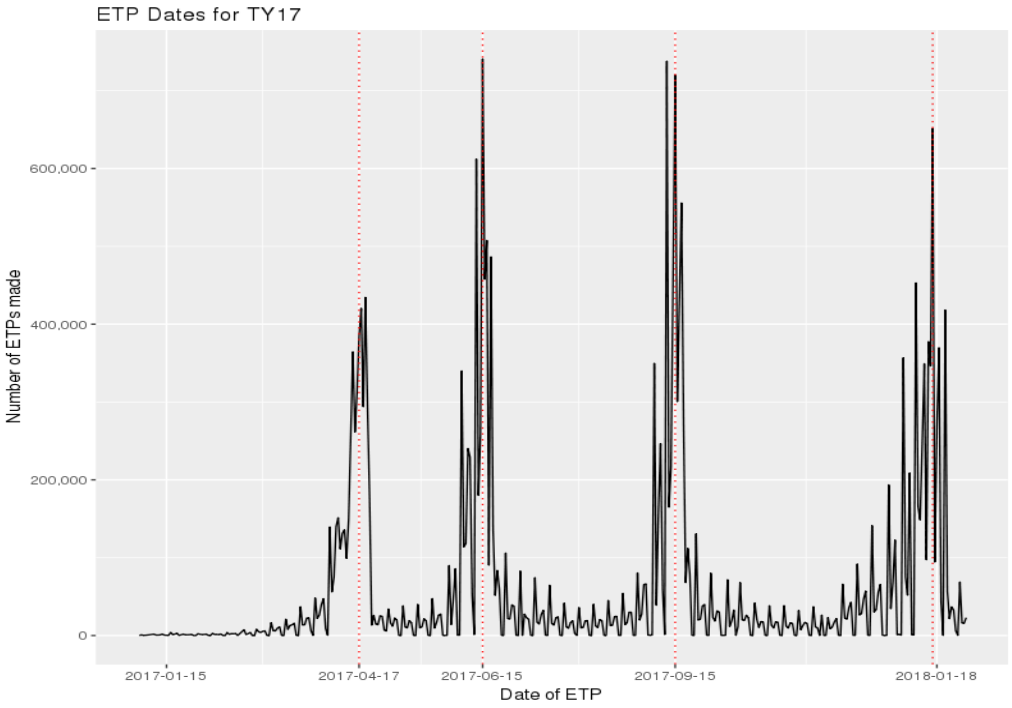
Our analysis has shown that one of the major contributing factors to the estimated tax penalty is a taxpayer's estimated tax obligation generated by income not subject to withholding. We thus investigate behavior around making estimated tax payments. First, we investigate the timing of estimated tax payments across the full universe of taxpayers who make estimated tax payments. Figure 11 and Table 13 show that the number of payments made for the first estimated tax payment deadline (April 15th) is lower than for all other quarters. One challenge is that this first estimated tax due date coincides with Tax Day. Taxpayers with estimated tax obligations may need to both file their tax return and pay any tax bill owed for the previous year, as well as make an

[9] Continuous regressors are standardized using the scale function in R, which subtracts the population mean from individual elements and divides them by the population standard deviation to produce dimensionless z-values whose levels can be compared.

estimated payment for income earned in the current calendar year. Table 13 further shows that the first quarter sees the fewest payments and that the total number of payments made per quarter increases steadily as the year goes on. Of note is also the fact that a significant number of taxpayers make more than one estimated tax payment in the same quarter, which might suggest that some taxpayers are making late 'make-up' payments later in the year.

We now look at estimated tax payments made by taxpayers who made payments but incurred an estimated tax penalty. As shown in Figure 12, the plurality of taxpayers, 24 percent, in this situation made their first estimated tax payment during the second quarter of the year. This is consistent with the findings that estimated tax payment volumes over the calendar year are lowest for the first quarter. In the year after incurring the penalty, 36 percent made their first estimated tax payment during the second quarter, a 12-percentage point increase from the previous year. However, the majority are still making estimated tax payments late and missing the first quarter's payment. These behaviors likely contribute to the incurrence and size of estimated tax penalties.

FIGURE 11. Estimated Tax Payment Volume Over a Calendar Year, TY 2017



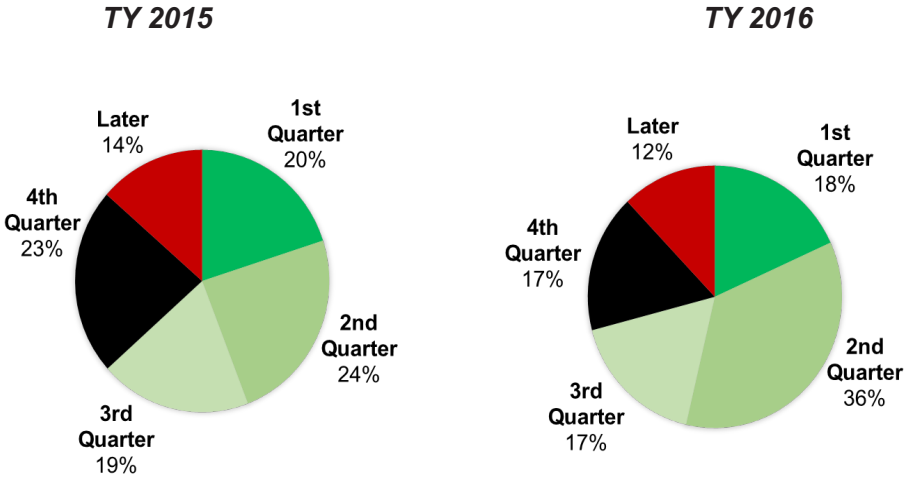
SOURCE: IRS Compliance Data Warehouse, May 2019.

TABLE 13. Aggregate Estimated Tax Payments Made by Quarter, TY 2017

Quarter	Number of payments (millions)	Number of taxpayers (millions)	Percent of taxpayers making more than 1 payment
1	3.72	3.60	1.9
2	4.89	4.40	11.4
3	6.55	5.58	18.7
4	7.83	6.27	27.9
Late	2.00	1.85	5.2

SOURCE: IRS Compliance Data Warehouse, September 2019.

FIGURE 12. Timing of First Payment Among Taxpayers with an Estimated Tax Penalty Making Estimated Tax Payments, TYs 2015–16



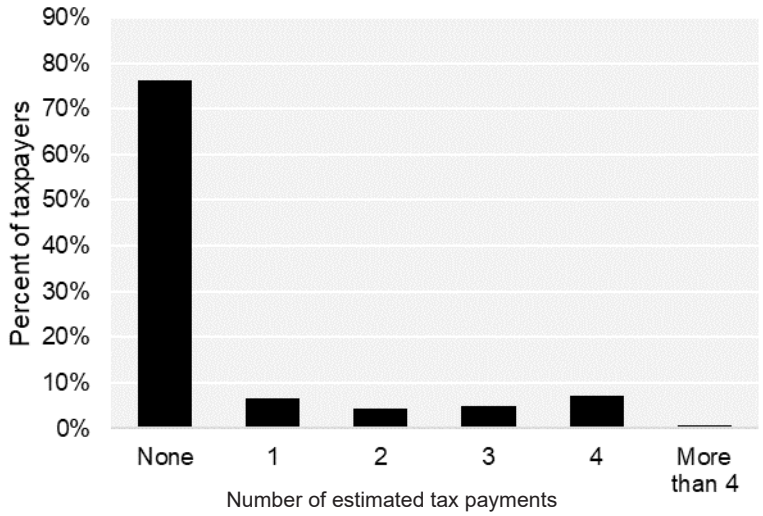
SOURCE: IRS Compliance Data Warehouse, 2018.

The universe of taxpayers who are making estimated tax payments is not the full universe of those who should be making payments. While the presence of nonwithholdable income does not necessarily mean that a taxpayer should be making estimated tax payments, as they may not earn sufficient income to generate tax liabilities, or could meet their tax liabilities through additional withholding, this population still provides a good starting point. The number of taxpayers with any amount of positive nonwithholdable income is 68.2 million (TY 2017), or 45 percent of all tax filers. The majority (81 percent) of these taxpayers are withholding, while only 13 percent make estimated tax payments,[10] and 15 percent make no prepayments at all. As can be seen in Figure 13, among taxpayers with estimated tax penalties who do not withhold or have withholdable income, only 24 percent made any estimated payments in TY 2017. This provides further evidence that the population of estimated tax payers understates the population of taxpayers with estimated tax liabilities.

While taxpayers without any withholdable income may be clear candidates for estimated tax payments, the majority of taxpayers earning nonwithholdable income also have access to withholding. To better explore when a taxpayer or household with a mix of income types should be making estimated tax payments, we begin by looking at payment behaviors based on two income constructions: 1) nonwithholdable income levels and 2) the fraction of total income that a taxpayer's nonwithholdable income comprises. Figure 14 shows that the percentage of taxpayers making estimated tax payments ranges from 7 percent to 72 percent at various levels of nonwithholdable income, with increasing likelihood of making payments as income increases. Over 50 percent of taxpayers are making estimated tax payments by the time their nonwithholdable income reaches around $120,000. The proportion making estimated payments remains about 5 percentage points higher for nonwithholding versus withholding taxpayers at most levels of nonwithholdable income. This is to be expected, as those who are withholding have an avenue for making their obligated prepayments throughout the year without needing to make estimated tax payments.

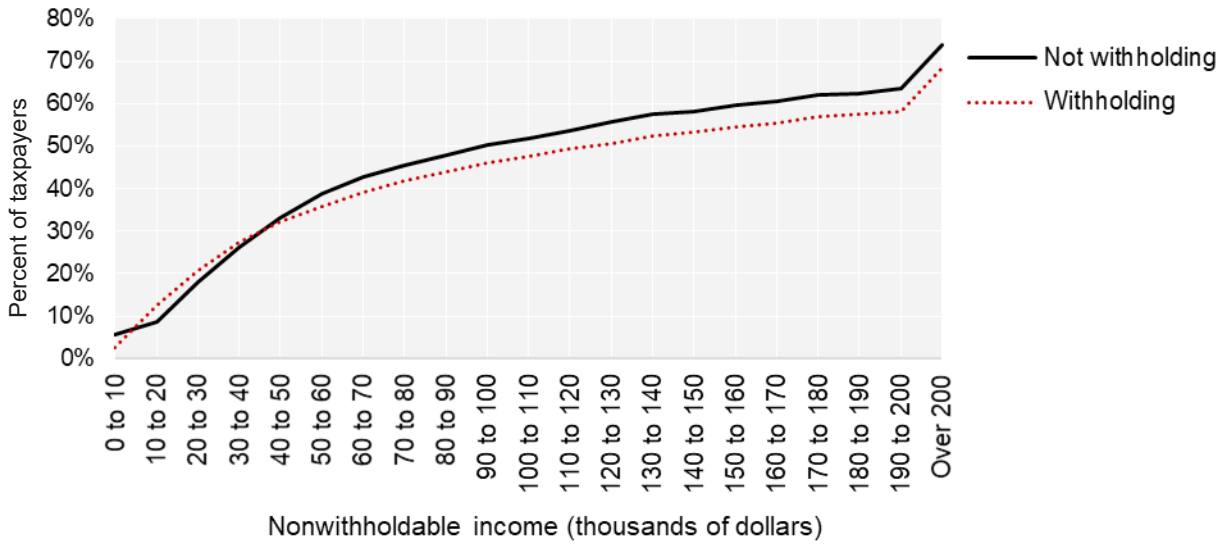
[10] Taxpayers who are making estimated tax payments may also be withholding.

FIGURE 13. Number of Estimated Tax Payments Submitted by Taxpayers with Estimated Tax Penalties Who Have No Withholding or Withholdable Income, TY 2017



SOURCE: IRS Compliance Data Warehouse, September 2019.

FIGURE 14. Percent of Taxpayers Making Estimated Tax Payments by Nonwithholdable Income Level and Withholding, TY 2017

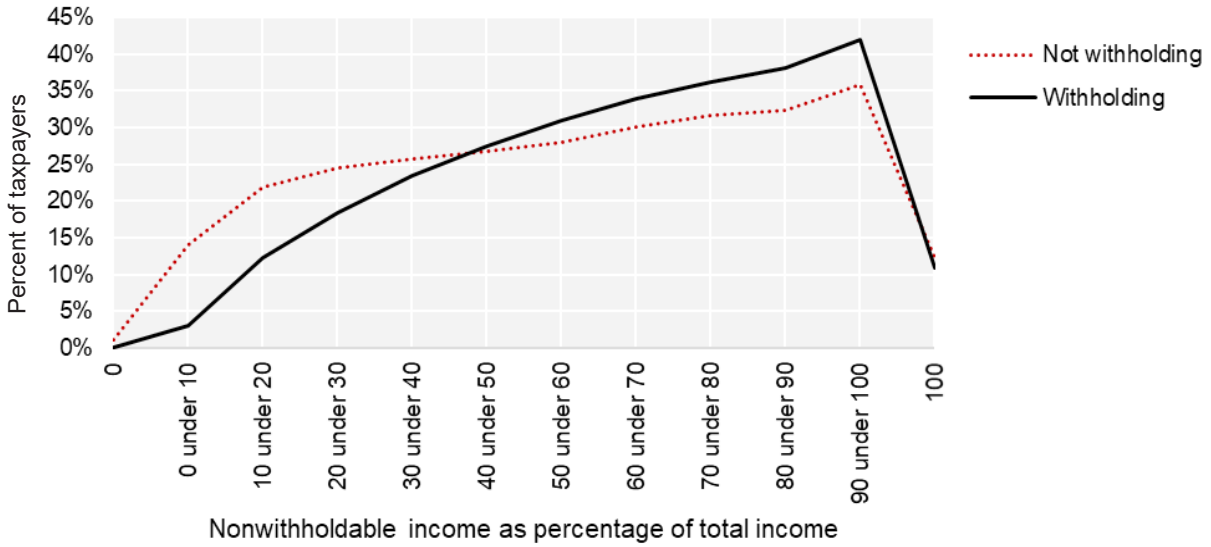


SOURCE: IRS Compliance Data Warehouse, August 2019.

How does the percent of total income coming from nonwithholdable sources affect whether a taxpayer makes estimated payments? As seen in Figure 15, based on fractions alone, without conditioning on level of income, no single decile of nonwithholdable income as a percentage of total income sees more than 50 percent of taxpayers making estimated tax payments. The closest is when a taxpayer's nonwithholdable income comprises 90 to 100 percent of their total income; there are close to 690,000 withholding taxpayers in this category and about 42 percent of them make estimated tax payments. Taxpayers who make 100 percent of their income from nonwithholdable sources tend to make less income overall. Less than 30 percent of taxpayers whose

nonwithholdable income comprises a minority of their total income make estimated tax payments, showing a significant preference against making estimated tax payments.

FIGURE 15. Percent of Taxpayers Making Estimated Tax Payments by Ratio of Nonwithholdable Income to Total Income and Withholding, TY 2017



SOURCE: IRS Compliance Data Warehouse, August 2019.

9. Discussion

The estimated tax penalty continues to ensnare a significant proportion of taxpayers every year, some of whom incur the penalty repeatedly. It is the most common penalty affecting individual taxpayers. Taxpayers may not know that they are required to prepay their taxes throughout the year and may not know which types of income are subject to withholding and which require self-made estimated tax payments. This is especially true among taxpayers with first-time or one-off income generating estimated tax liability. Withholding plays a key role in mitigating estimated tax penalties; the penalty is much more common among taxpayers who have significant sources of nonwithholdable income like self-employment income or investment income. Thus, there may be policy questions around which types of income should mandate or provide opt-in capacity for withholding, and whether the process of making estimated tax payments can be optimized for tax compliance.

Another factor affecting the prevalence of the estimated tax penalty is its size. The estimated tax penalty was not designed to change taxpayers' behavior. The Internal Revenue Manual states: "The purpose of the estimated tax penalties is not so much to penalize the taxpayer as it is to compensate the United States for the use of money that should have been paid over to the U.S. Treasury" (IRM § 20.1.3.1.1). General deterrence theory, the theory that a sentencing objective can discourage other than the offender from committing an offense, can apply here, in that the punishment factor is so trivial that it may be minimizing deterrence and limiting any impact on compliance. Higher-income taxpayers in particular may not find the penalty very salient. In addition, a clever taxpayer acting rationally under the classical economic theory of tax compliance (CRS (2019)) could eschew making estimated tax payments for nonwithholdable income and plan to incur the penalty, if they are able to get a return on investment on these would-be payments that is higher than the rate of the penalty they will pay later. Alternatively, they may be cash-constrained and choose not to prepay taxes if their opportunity cost of credit is higher than their penalty rate. This idea of the opportunity cost of capital is something that is being explored by other research, including Boning's (2018) paper on the effect of withholding on tax-filing behavior.

Literature has found that taxpayers think "primarily in terms of the out-of-pocket gains and losses at the time of filing a return," meaning that the level of refund or taxes owed upon filing is correlated with a taxpayer's

voluntary compliance (Yaniv (1998)). Some of this research utilizes prospect theory rather than traditionally applied expected utility theory to model the relationship between prepayments and a taxpayer's decision to evade taxes through underreporting income or overreporting expenses; risk-averse taxpayers who make overpayments have fewer incentives for noncompliance. Other research on taxpayer preferences for overwithholding includes a simulated laboratory experiment with 132 participants which found that roughly half of the group preferred overpayments beyond what was needed to avoid an underpayment penalty, regardless of whether they were selected to withhold (presented as a "net gain" frame) or make estimated tax payments (presented as an "explicit loss" frame) (Ayers *et al.* (1999)). Uncertainty in tax liability led to a preference for higher payments beyond what was needed in the worst-case scenario to avoid the penalty, while more experienced taxpayers were less likely to overpay and were less sensitive to uncertainty. This study corroborates other research on taxpayers' seemingly irrational preference for receiving refunds and leads to further questions about what latent or behavioral characteristics are different among taxpayers with estimated tax penalties (Feltham and Paquette (2002)). One possibility is an inability to afford prepayments, which could supersede any positive preferences for making overpayments.

The Tax Cuts and Jobs Act (TCJA), implemented in Tax Year 2018, presented an opportunity for both taxpayers and the IRS to reexamine their approaches to prepayments. Treasury and the IRS decided to relax the safe harbor criteria for incurring estimated tax penalties for TY 2018 to help accommodate taxpayers to the new tax schedules.[11] As of August 2019, this temporary rule led to a 28-percent drop in the number of estimated tax penalties assessed in TY 2018 as compared to TY 2017. This is the equivalent of 2.7 million fewer penalties, to an overall count of 7 million penalties, down from 9.6 million. In addition, TCJA enacted cross-population tax cuts and did away with personal exemptions. The new law did not require taxpayers to update their withholding despite new caps and removals of certain itemized deductions and the doubling of the standard deduction, which reduced the overall liability of many taxpayers and in turn set them up to overclaim withholding allowances and to underwithhold taxes, if they did not adjust their withholding during TY 2018.

TCJA represents a natural experiment that can be studied further. One idea would be to pursue a regression discontinuity model where taxpayers close to the penalty threshold on either side could be compared to see how being charged the penalty affected their prepayment behavior in subsequent years. For withholding taxpayers, we should be able to see who updated their withholding for the tax year and who did not, and the characteristics associated with each population.

10. Conclusion

A consistent share of close to 7 percent of all individual tax returns each year are assessed an estimated tax penalty for failing to prepay enough tax throughout the year through either tax withholding from the source and/or self-initiated estimated tax payments. This corresponds to a population of almost 10 million tax returns annually, in recent years. Taxpayers incurring estimated tax penalties on average have higher levels of income, more types of income, and higher levels of income not subject to withholding; they are also more likely to be older and married. Although the estimated tax penalty could be incurred even by a taxpayer without a balance due upon filing (for instance, if they had made late prepayments), the estimated tax penalty most frequently appears with higher balances due upon filing and a higher incidence of unpaid assessments, which are costlier and more time-consuming to resolve for both taxpayers and the IRS. More than a third of taxpayers receiving this penalty incur multiple penalties for the same tax year and are more likely to have discrepancies between information reports and reported income, which suggest underreporting of income. While the penalty works to change some taxpayers' prepaying behavior, many are not affected, suggesting there could be room for policy changes to create better pathways for prepayment compliance through increased access to withholding or improvements to the process of making estimated payments. The IRS is proposing research around this latter component of prepayment, which is presented in the Future Research section of this paper. Maximizing voluntary compliance continues to be a win-win goal for the IRS and the public; optimizing prepayment behavior and minimizing the prevalence of the estimated tax penalty are foundational to this objective.

[11] The safe harbor rule to prepay 90 percent of one's taxes by the last estimated tax payment deadline was relaxed to 85 percent and then 80 percent for Tax Year 2018 (IRS (2019c)).

11. Future Research on Estimated Taxes

Overall, the underpaying of estimated taxes on income that is not subject to withholding remains a key area of prepayment noncompliance. Additional analysis and research are needed to fully understand this taxpayer behavior, which may differ even for taxpayers of similar income or social demographics. Our in-process and proposed follow-on research focuses on understanding and encouraging taxpayer compliance in this area with compliance challenges—making estimated tax payments. These include: A) a Web survey sent to compliant and noncompliant taxpayers to estimate the burden of making estimated payments and to understand the drivers of noncompliance, and B) a randomized controlled experiment testing a variety of behaviorally informed reminder letters to encourage the timely submission of estimated payments.

A. Estimated Tax Payment Survey

The Web survey samples two groups of taxpayers: those making estimated tax payments for Tax Year 2018 and those not making payments who probably should have been making them.[12] The 20,000-taxpayer sample is segmented by taxpayers making any payments, to form a final split of 70-percent payers and 30-percent nonpayers. The survey includes questions about awareness of estimated tax requirements, sources for estimated tax information, income variability and predictability, preferences around payment timing and scheduling, time and money spent calculating and submitting payments, penalty salience and utility, changes to behavior after incurring the estimated tax penalty, recordkeeping behavior, income report frequency, saving and planning for taxes, attitudes towards taxes, self-employment identification and motivation, and customer service feedback. Among other functions, the survey will be used to calculate the individual burden of making estimated tax payments and to understand which barriers are most salient in hindering taxpayers from making payments, and which, if any, can be addressed operationally or via proposed policy changes.

B. Estimated Tax Payment Reminder Letter Experiment

Our proposed experimental outreach would test four different letters within seven different treatment options against a selected control group. The letters include a baseline reminder letter, a letter describing an option to make monthly payments, a letter stating that making payments can help the taxpayer avoid a penalty, and a letter stating that making payments can help the taxpayer avoid a high tax bill. The idea for the monthly option treatment letter is supported by other experimental research, which found that the majority of self-employed taxpayers given the option to make monthly payments would choose to do so and would end the tax year with less delinquency (Chambers and Curatola (2012)). A range of other preexisting behavioral research by the IRS and other tax authorities was reviewed to prioritize the letter designs to test (Meiselman (2018); Hallsworth (2016); Guyton *et al.* (2017); Chirico *et al.* (2016); Orlett *et al.* (2017); Ariely and Wertenbroch (2002)). Treatments include sending each letter once and sending the three letters aside from the monthly payment option in multiple waves before estimated tax payment deadlines to measure the impact of repeated reminders versus general awareness.

Proposed sampling criteria could include taxpayers who incurred an estimated tax penalty in Tax Year 2017, made fewer than four estimated tax payments in Tax Year 2018, had at least $10,000 of taxable income in TY 2017, and were either not withholding or were a married filing jointly household whose nonwithholdable income made up at least 50 percent of their overall income. These criteria would reduce the overall taxpayer population down to a universe of approximately 2 million returns, from which a sample of 60,000 could be randomly drawn. The sample could be segmented by taxpayers making any payments in TY 2018, oversampling those making any payments. The sample would be limited to taxpayers residing in U.S. States and the District of Columbia with a complete mailing address from a recent tax return.

[12] The non-paying group consists of taxpayers who had an estimated tax penalty in Tax Year 2017, were not making payments for Tax Year 2018, and were either nonwithholding or were filing jointly and reported a majority of their income from nonwithholdable sources. The first group of taxpayers making payments consists of any who made at least one estimated tax payment in Tax Year 2018, regardless of how fully compliant their payments were or whether they incurred a penalty in Tax Year 2017.

This pilot experiment could measure the cost-effectiveness of a soft notice approach for estimated tax payment compliance. Costs, aside from printing and mailing letters, include potential impact to call center volumes from taxpayers with questions about payment reminder letters. Findings from other IRS research indicate that the prevention of downstream compliance problems would likely outweigh these upfront operational costs.

References

Ariely, D., and Wertenbroch, K. 2002. Procrastination, Deadlines, and Performance: Self-Control by Precommitment. *Psychological Science*, 13(3): 219–224.

Ayers, B. C., Kachelmeier, S. J., and Robinson, J. R. 1999. Why Do People Give Interest-Free Loans to the Government? An Experimental Study of Interim Tax Payments. *The Journal of the American Taxation Association*: Fall 1999, 21(2): 55–74.

Boning, W. C. 2018. Does Employer Withholding Affect Tax Compliance, and Why? *2018 IRS Research Bulletin*, 8th Annual Joint Research Conference on Tax Administration, 3–7.

Chambers, V., and Curatola, A. 2012. Could Increasing the Frequency of Estimated Tax Payments Decrease Delinquency Rate Among the Self-Employed? *Advances in Taxation, Vol. 20*. Bingley: Emerald Group Publishing Limited, 1–28.

Chirico, M., Inman, R. P., Loeffler, C., MacDonald, J., and Sieg, H. 2016. An Experimental Evaluation of Notification Strategies To Increase Property Tax Compliance: Free-Riding in the City of Brotherly Love. NBER, *Tax Policy and the Economy*, 30(1), 2016.

Congressional Research Service Insight. 2019. Behavioral Economics, IRS Letter Campaigns, and Tax Compliance. IN11151. Retrieved from https://crsreports.congress.gov/product/pdf/IN/IN11151, August 1, 2019.

Engstrom, P., Nordblom, K., Ohlsson, H., and Persson, A. 2015. Tax Compliance and Loss Aversion. *American Economic Journal: Economic Policy*, 7(4) (November 2015): 132–164.

Feltham, G. D., Paquette, S. M. 2002. The Interrelationship between Estimated Tax Payments and Taxpayer Compliance. *The Journal of the American Taxation Association: Supplement 2002*, 24(s-1): 27–45.

Guyton, J., Langetieg, P, Manoli, D., Payne, M., Schafer, B., and Sebastiani, M. 2017. Reminders and Recidivism: Using Administrative Data To Characterize Nonfilers and Conduct EITC Outreach. *American Economic Review: Papers & Proceedings 2017*, 107(5): 471–475.

Hallsworth. 2016. Reducing Fraud, Error and Debt. The Behavioral Insight Team's Update Report: 2015–2016. Retrieved from http://www.behaviouralinsights.co.uk/wp-content/uploads/2016/10/BIT_Update_Report_2015-16_ReducingFraudErrorDebt.pdf, March 2019.

Internal Revenue Service. 2016. Tax Gap Estimates for Tax Years 2008-2010. Publication 1415 (5–2016), https://www.irs.gov/newsroom/the-tax-gap. Retrieved August 1, 2019.

Internal Revenue Service. 2018. Publication 525: *Taxable and Nontaxable Income: For Use in Preparing 2018 Returns*.

Internal Revenue Service. 2019a. IRS, Treasury unveil proposed W-4 redesign for 2020. 2019, May 31. News Release. Retrieved from https://www.irs.gov/newsroom/irs-treasury-unveil-proposed-w-4-design-for-2020.

Internal Revenue Service. 2019b. *IRS Data Book 2018*. Table 17: Civil Penalties Assessed and Abated, by Type of Tax and Type of Penalty, Fiscal Year 2018 (XLS). Retrieved from https://www.irs.gov/statistics/soi-tax-stats-civil-penalties-assessed-and-abated-by-type-of-tax-and-type-of-penalty-irs-data-book-table-17, July 2019.

Internal Revenue Service. 2019c. Notice 2019–25: Relief from Addition to Tax for Underpayment of Estimated Income Tax by an Individual. Office of the Associate Chief Counsel.

Internal Revenue Service. 2019d. Topic No. 652: Notice of Underreported Income—CP2000. Retrieved from https://www.irs.gov/taxtopics/tc652. Updated August 23, 2019.

Kahneman, D., and Tversky, A. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica*, 47(2) (March 1979): 263–292.

Meiselman, B. 2018. Ghostbusting in Detroit: Evidence on nonfilers from a controlled field experiment. *Journal of Public Economics*, Elsevier, 158(C): 180–193.

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. Retrieved from https://www.irs.gov/pub/irs-soi/17resconorlett.pdf, August 1, 2019.

Yaniv, G. 1998. Tax Compliance and Advance Tax Payments: A Prospect Theory Analysis. The National Insurance Institute Research and Planning Administration. Discussion Paper No. 68.

The Positive and Negative Effects of Burdensome Audits

Amy Hageman (Kansas State University), Ethan LaMothe (Oklahoma State University), and Mary Marshall (Louisiana Tech University)[1]

The purpose of this study is to examine the effect of audit burden on compliance behavior subsequent to the audit. Specifically, we define audit burden as the monetary and non-monetary costs of responding to and complying with an income tax examination ("audit") that are independent of an individual's chosen level of compliance (i.e., the costs of being audited as opposed to the costs of being noncompliant). Prior research suggests experiencing an audit influences subsequent compliance behavior (Boylan (2010); Kastlunger *et al.* (2011)) and higher penalty rates should lead to higher levels of compliance (Alm *et al.* (1992)). As a result, a higher level of audit burden may be viewed as beneficial to the extent it reinforces the perceived costs of noncompliance.

However, audit burden is not a penalty for noncompliance per se because it is experienced by all audited taxpayers, irrespective of their chosen compliance level. Whereas only noncompliant individuals (hereafter "evaders") incur penalties for noncompliance, both compliant individuals (hereafter "compliers") and evaders can experience audit burden. Prior research indicates not all individuals who undergo an audit are evaders (Beer *et al.* (2015); Gemmell and Ratto (2012)). In fact, Internal Revenue Service (IRS) data suggest at least 13–15 percent of audited taxpayers made appropriate compliance choices or even overpaid their tax liability (IRS (2018)). Audit burden may therefore have unintended effects on the choices of compliers after an audit. Accordingly, we examine the effect of audit burden with a particular interest in understanding how burden differentially influences compliers and evaders.

Mental accounting theory suggests audit burden will be perceived differently depending on whether an individual complied or evaded on the audited return due to differences in cost-loss framing. Specifically, prior research finds individuals frame an expenditure as a cost when it is associated with some benefit and as a loss when it is not (Kahneman and Tversky (1984); Thaler (1985); Lipe (1993)). In relation to tax audits, we predict evaders view audit burden as an additional cost attributed to the same mental account as their choice to evade, which in turn reinforces the effect of other costs. Accordingly, we expect evaders will increase compliance subsequent to experiencing a burdensome audit (as compared to an audit without burden). In contrast, compliers incur audit burden despite making truthful compliance choices. As no benefit is associated with the audit burden endured, we expect compliers will perceive audit burden as a loss and will subsequently become increasingly risk-seeking (Kahneman and Tversky (1979)). As a result, we predict compliers will reduce compliance subsequent to experiencing high audit burden (as compared to an audit without burden).

Our results are consistent with our hypotheses. Specifically, in an online experimental setting using actual U.S. taxpayers, we find evaders report a greater amount of income after experiencing a high burden audit than after experiencing an audit without any burden. In fact, we find evaders experiencing an audit without burden do not change their compliance after being audited, suggesting traditional evasion penalties alone may not be sufficient to deter evasion. This finding is particularly interesting given the penalties associated with noncompliance in our experiment (150 percent of taxes evaded) are twice the size of the civil penalty for fraudulently

[1] We are very grateful for the feedback and helpful comments on previous drafts of this manuscript from Paul Black, Cynthia Blanthorne, Donna Bobek Schmitt, Billy Brink, Bonnie Brown, Natasha Bernhardt (discussant), Brian Erard (discussant), Diana Falsetta, Sarah Judge, Susan Jurney (discussant), Marlys Lipe, Jason Rasso, Tim Rupert, Jason Schwebke, Shane Stinson (discussant), and Scott Vandervelde, as well as participants at the 2018 AAA Annual Meeting, the 2018 ABO Research Conference, the 2019 PhD Project ADSA meeting, the 2018 ATA Midyear Meeting, the 2018 Behavioral Tax Symposium, 2019 IRS-TPC Joint Research Conference, the Spring 2019 Texas Tech University Tax Seminar, and workshop participants at the University of South Carolina.

underreporting tax liability in the U.S. (75 percent of taxes evaded). In contrast, compliers reduce the amount of income they report to a greater extent after experiencing a high burden audit than after experiencing an audit without burden.

In a supplementary experiment, we also examine whether an intervention by the tax authority can offset the perceived loss and prevent the increase in noncompliance among compliers. Informed by research on rewards in tax compliance (Kastlunger *et al.* (2011)) and the apology literature (Davidow (2003); Kim *et al.* (2004); Doyle *et al.* (2009); Roschk and Kaiser (2013)), we predict audited compliers are less likely to reduce subsequent compliance when the taxing authority apologizes for the undue burden and acknowledges the individual taxpayer's compliance. Consistent with our expectation, we find no evidence of an effect of audit burden on compliers who received an apology. This finding suggests a simple apology may mitigate the negative effect of audit burden on compliers.

This study makes several important contributions. First, we introduce audit burden as an additional antecedent and deterrent of evasion, which has not previously been incorporated into models of compliance choices. Traditional models of tax evasion (e.g., Allingham and Sandmo (1972); Yitzhaki (1974)) view compliance decisions as an economic gamble in which taxpayers maximize expected utility based *only* on tax rates, audit rates (probability of detection), and penalty rates. More recent research finds noneconomic factors such as tax morale (Torgler (2007)), trust in government (Kirchler (2007)), and perceptions of fairness (Wenzel (2003)) influence compliance decisions. We bridge these literatures by examining audit burden, a construct potentially containing both monetary and nonmonetary elements.

Furthermore, our results have important implications for regulators and enforcement agencies. The level of audit burden experienced by an individual being audited is, to some extent, under the control of the auditor as the degree of audit burden is likely influenced by the approach an auditor takes (e.g., a holistic field audit may be more burdensome than a narrow scope correspondence audit). These same choices also likely influence the extent to which an audit will effectively uncover any noncompliance. Auditors must balance the benefits and costs of using higher burden audits, as burdensome audits may not only be necessary to detect and deter noncompliance, but also have negative effects. Our study provides evidence of the trade-offs evident in these audits.

References

Allingham, M. G., and A. Sandmo. 1972. Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics* 1 (3–4): 323–338.

Alm, J., B. Jackson, and M. McKee. 1992. Estimating the Determinants of Taxpayer Compliance with Experimental Data. *National Tax Journal* 45 (1).

Beer, S., M. Kasper, E. Kirchler, and B. Erard. 2015. Audit Impact Study. *National Taxpayer Advocate 2015 Annual Report to Congress*, 2, 68–98.

Boylan, S. J. 2010. Prior Audits and Taxpayer Compliance: Experimental Evidence on the Effect of Earned Versus Endowed Income. *The Journal of the American Taxation Association* 32 (2): 73–88.

Davidow, M. 2003. Organizational Responses to Customer Complaints: What Works and What Doesn't. *Journal of Service Research* 5: 225–250.

Doyle, E., K. Gallery, and M. Coyle. 2009. Procedural Justice Principles and Tax Compliance in Ireland: A Preliminary Exploration in the Context of Reminder Letters. *Journal of Finance and Management in Public Services* 8 (1): 49–62.

Gemmell, N., and M. Ratto. 2012. Behavioral Responses To Taxpayer Audits: Evidence From Random Taxpayer Inquiries. *National Tax Journal* 65 (1): 33–58.

Internal Revenue Service (IRS). 2018. *Data Book, 2017*. Washington, DC.

Kahneman, D., and A. Tversky. 1979. Prospect Theory: An Analysis of Decision Under Risk. *Econometrica* 47 (2): 263–291.

———. 1984. Choices, Values, and Frames. *American Psychologist* 39 (4): 341–350.

Kastlunger, B., S. Muehlbacher, E. Kirchler, and L. Mittone. 2011. What Goes Around Comes Around? Experimental Evidence of the Effect of Rewards on Tax Compliance. *Public Finance Review* 39 (1): 150–167.

Kim, P. H., D. L. Ferrin, C. D. Cooper, and K. T. Dirks. 2004. Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence-Versus Intergrity-Based Trust Violations. *Journal of Applied Psychology* 83 (1): 104–118.

Kirchler, E. 2007. *The Economic Psychology of Tax Behaviour*. Cambridge University Press.

Lipe, M. G. 1993. Analyzing the Variance Investigation Decision: The Effects of Outcomes, Mental Accounting, and Framing. *The Accounting Review* 68 (4): 748–764.

Roschk, H. and S. Kaiser. 2013. The Nature of an Apology: An Experimental Study on How To Apologize After a Service Failure. *Marketing Letters* 24: 293–309.

Thaler, R. 1985. Mental Accounting and Consumer Choice. *Marketing Science* 4 (3).

Torgler, B. 2007. *Tax Compliance and Tax Morale: A Theoretical and Empirical Analysis*. Edward Elger Publishing.

Wenzel, M. 2003. Tax Compliance and the Psychology of Justice: Mapping the Field. *Taxing Democracy*: 41–70.

Yitzhaki, S. 1974. Income Tax Evasion: A Theoretical Analysis. *Journal of Public Economics* 3 (2): 201–202.

Using a Graph Database To Analyze the IRS Databank

Ririko Horvath and Rahul Tikekar (IRS Research, Applied Analytics, and Statistics)

Introduction and Motivation

The Research, Applied Analytics, and Statistics (RAAS) organization of the Internal Revenue Service (IRS) is the primary research arm of the agency. RAAS provides researchers and analysts with a wide range of data, tools, and infrastructure to analyze tax data. The IRS Research Conference is testament to the efforts and functions of this organization. The Compliance Data Warehouse (CDW), hosted in RAAS, forms the main source of data for most analysts and researchers (RAAS (2019)). It is a collation of several IRS upstream data sources like individual returns, business returns, information returns (W-2, 1099, etc.), and other relevant and useful data, including records of births and deaths from the Social Security Administration.

The IRS Databank (Raj Chetty (2018)) is a panel dataset within CDW where data from different CDW datasets are assembled and arranged by individual taxpayer and tax year. What makes the Databank unique is that it provides a longitudinal view of taxpayers over generations. The Databank is built by first generating a list of all Social Security Numbers (SSNs) from the records of births and deaths (for all individuals not reported to be dead before 1996–the first year of the Databank). This is termed the spine of the Databank. The spine is then augmented with data from the CDW for each of the SSNs. As a result, if a certain individual did not file a return in a year, and there weren't any information returns for them, all the fields of the record for that individual's SSN for that year will be blank. Along the same lines, if an individual did not file a tax return for a year but did receive information returns (like a 1099), then many of the fields for that record will be blank, but the fields corresponding to the information returns will be populated.

Conceptually, the Databank can be visualized as linked data: taxpayers connected to their spouses and to their dependents. The dependents, in turn, connected to their spouses and their dependents, and so on. Such a longitudinal dataset can provide researchers and analysts with unique opportunities to study the behavior and evolution of the American taxpayer. At the database level, however, these data are stored in a relational database. Relational databases store data in tables, and to extract data one must employ a query language called SQL (not an acronym). Querying longitudinal or linked data via SQL is not intuitive. Also, in terms of performance, queries pursuing links can become highly CPU-intensive and memory-intensive, because they employ an expensive operation called "the join."

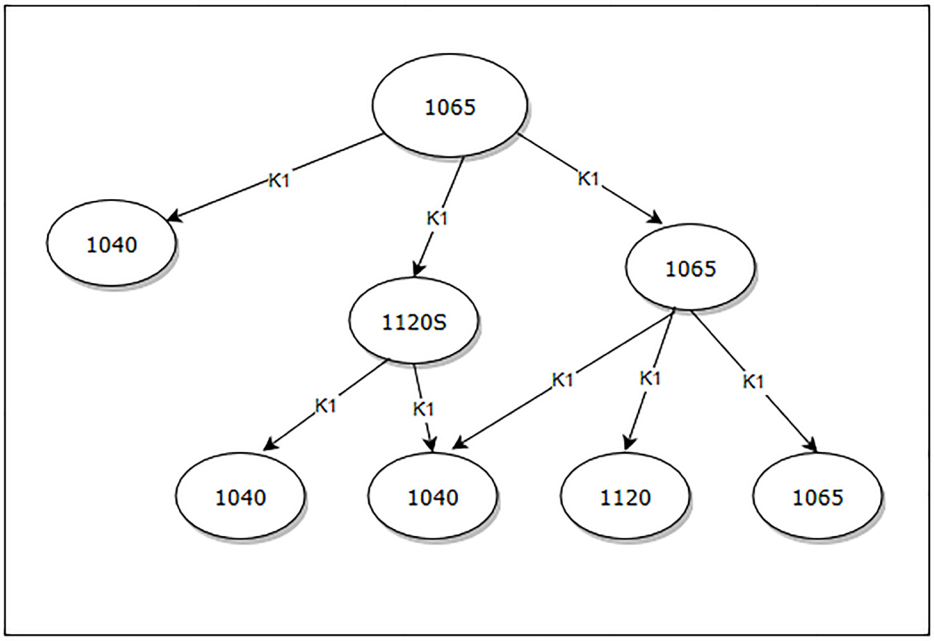
This paper proposes an alternative to storing the Databank in a relational database: the graph database. The Databank then can not only be conceptually visualized as a linked dataset (or graph) but can also be implemented as one.

Graph Databases

Just as the table (or relation) is the basic unit in a relational database, the graph is the conceptual unit in a graph database. A graph consists of "nodes" (also called vertices) and "edges" (also called links); an edge symbolizes a relationship between nodes (Joshi (2017)). Generally, each node represents an entity (like a person or taxpayer), and the relationship can represent an association like "files return" or "is married to." Whereas a relational database design requires knowledge of functional dependencies and normal forms, a graph database design is more intuitive. An example of a graph database that models tax entities such as Partnerships (1065), S-corporations (1120S), and Individuals (1040) is shown in Figure 1. Notice, the simplicity of expressing the data and their connections. The example graph models flowthrough (or passthrough) tax entities and their relationships with other tax entities (upcounsel (2019)). This relationship is the form K1 that all flowthrough

entities send to their partners (or shareholders). The simple example in Figure 1 shows only one type of relationship (K1) among nodes. In a real-world situation there can exist many types of relationships among nodes (e.g., parent/subsidiary links, tax preparer links, address links, etc.).

FIGURE 1. Illustration of Typical Nodes



It is easy to see that graphs can get very deep (e.g., a partnership can partner with other partnerships and so on) and can get very bushy (e.g., a partnership can have thousands of partners). This is where graph databases shine and can outperform relational databases while performing traversals. Relational databases do very well in a transaction-based environment (like in a department store database), while graph-based databases are very efficient when following links in a linked dataset where data don't change very often. This paper proposes a linked paradigm to store the Databank instead of its current relational database format.

For the example in Figure 1, consider the task of finding all the owners of a given partnership no matter how many levels deep. In a relational database such a query is expressed as a series of joins (recursive joins). Not only is such a query difficult to imagine and express, it is also very expensive in terms of CPU time and memory usage. In a graph database such an operation is a matter of following the links from one node to another and is very fast.

In addition to storing nodes and edges, a graph database can also store one or more properties on each node and edge. For example, the node 1065 in the graph can store properties such as a taxpayer ID (TIN or SSN), name, address, etc., while the K1 edges can store properties such as gain, loss, tax year, etc. Such a graph, with nodes and edges along with their properties, is termed a property graph. It is possible to query a property graph database based on node and edge properties. An overview of property graphs along with modeling data with graphs is given in this reference (Frisendal (2017)).

Querying a Graph Database

Querying a property graph database is more natural than querying a relational database as will be shown next. Without going into all the gory details, this section presents a rudimentary overview of the concepts of querying a graph database using a popular graph query language called Cypher (Neo4j (2019)).

A query is constructed by representing nodes and relationships using a straightforward syntax:

```
(node) -[relationship]-> (node)
```

For example, `(1065) - [:K1] -> (1040)` would indicate to follow the K1 relationship from a 1065 node to a 1040 node. This is combined with keywords and variables to forms complete queries. Here are some examples:

a. Show all partnership (1065) nodes in the graph. In this example, MATCH and RETURN are keywords while p is a variable.

```
MATCH (p:1065)
RETURN p
```

b. Show all partnership (1065) nodes located in the 20224 ZIP Code. This example shows how properties on a graph can be queried. This assumes that there is a zip property on the node.

```
MATCH (p:1065 {zip: '20224'})
RETURN p
```

c. Find all individuals receiving a K1 from partnerships. This involves following the K1 relationship from a 1065 node to a 1040 node.

```
MATCH (p:1065) - [:K1] -> (i:1040)
RETURN i
```

d. Find all individuals receiving a K1 from partnerships located in the 20224 ZIP Code area. This is a simple extension of the previous query.

```
MATCH (p:1065 {zip: '20224'}) - [:K1] -> (i:1040)
RETURN i
```

e. Show the names of individuals receiving a K1 from partnerships located in the 20224 ZIP Code area. Note the subtle difference between the previous query and this. In the previous query we requested the entire 1040 node whereas this query requests the name property of the 1040 node.

```
MATCH (p:1065 {zip:'20224'}) - [:K1] -> (i:1040)
RETURN i.name
```

f. Show names of partnerships who issued K1s to an individual with SSN nnnnnnnnn: to answer this query, look for sets of nodes such that their K1 links point to the given individual: `(1065) - [:K1] -> (1040) <- [:K1] - (1065)`

```
MATCH  p1:1065)-[:K1]->(i:1040{SSN:'nnnnnnnnn'})<-[:K1]-(p2:1065)
RETURN p1.name, p2.name
```

g. To query variable length pattern matching, the syntax provides for specifying how deep you want the relationship to be traversed: Show all partnerships receiving a K1 from "Acme Partnership" two levels deep. This would mean that Acme issues K1 to an entity, that in turn issues a K1 to a partnership.

```
MATCH (p:1065 {name: 'Acme Partnership'} - [:K1*2] -> (p:1065)
RETURN p
```

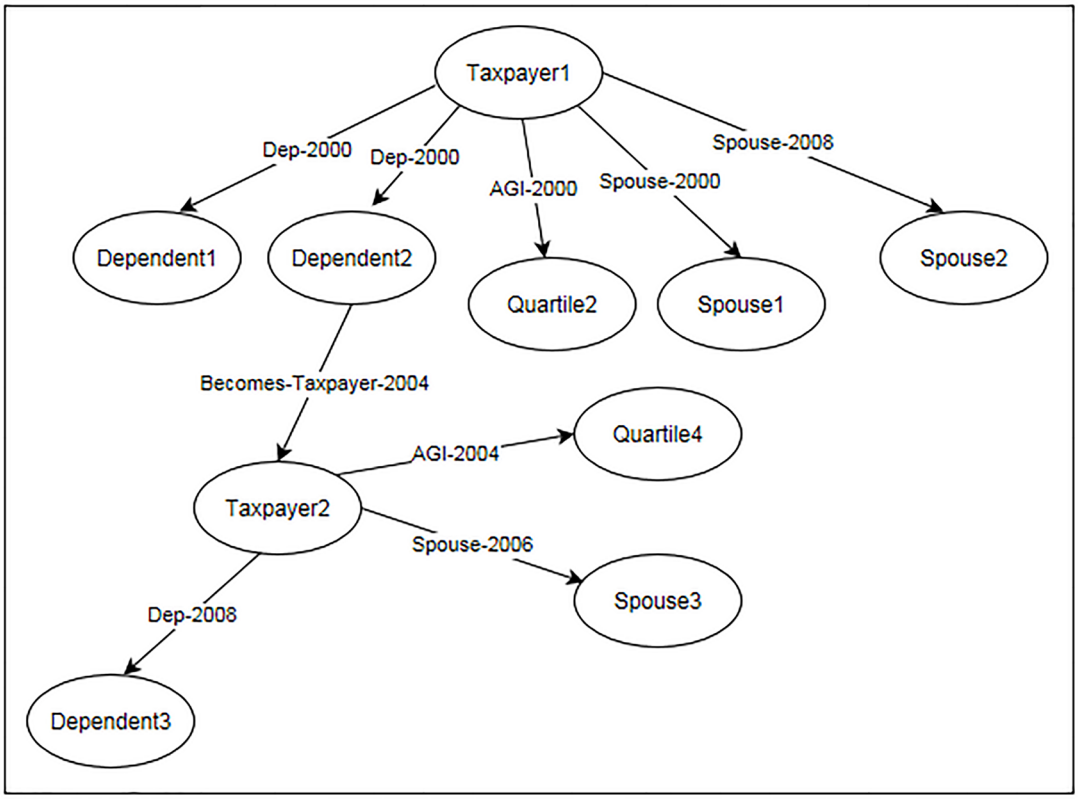
Modeling the Databank as a Graph Database

This section presents the proposed graph database model of the Databank. There is generally more than one way to model a given problem, and the design depends on how one wishes to use the data. In this case, the proposed model can be used to follow taxpayers and their dependents. The Databank provides several attributes

of a taxpayer, and, for this exercise, a small subset of those attributes is chosen to demonstrate the concept. When a production-level database is created there are two choices for the use of the attributes: a. all the attributes can be used as properties in the graph database, or b. a hybrid design is used where some properties remain in a relational database while the relevant ones are moved into the graph database.

In the design presented here, the attributes chosen to model are: taxpayer identification number (TIN), taxpayer name, ZIP Code of the taxpayer, their adjusted gross income (AGI) expressed as a quartile, and their dependent and spouse information. The model is shown in Figure 2.

FIGURE 2. Illustrative Model Design



The model design can be summarized this way: For a tax year, a taxpayer can claim one or more dependents and/or a spouse. Because the exact AGI may not be of importance, reported AGI is then assigned to a quartile. Note that it is entirely possible that a taxpayer may not claim a dependent or a spouse, in which case there will not be any links between the taxpayer node and a dependent node or a spouse node. Further note that if a taxpayer claims multiple dependents, there will be multiple links from the taxpayer node to multiple dependent nodes. To avoid clutter and to focus on the concept, the graph model in Figure 2 is a highly simplified version of a real graph.

Specifically, in the example in Figure 2, for the Tax Year 2000, a taxpayer (Taxpayer1) claims two dependents Dependent1 and Dependent2. In that year, Taxpayer1's AGI falls in the 2nd quartile. For the same tax year (2000) Taxypayer1 also reports a spouse (Spouse1). In Tax Year 2008, Taxpayer1 claims a different spouse (Spouse2), presumably because of remarrying.

At some point in time, a dependent will cease being a dependent and, hopefully, become a taxpayer. This is captured by the relationship "Becomes-Taxpayer." They can, in turn, also claim their own dependents and spouse. Eventually the dependent's dependents will also become taxpayers thus resulting in a longitudinal

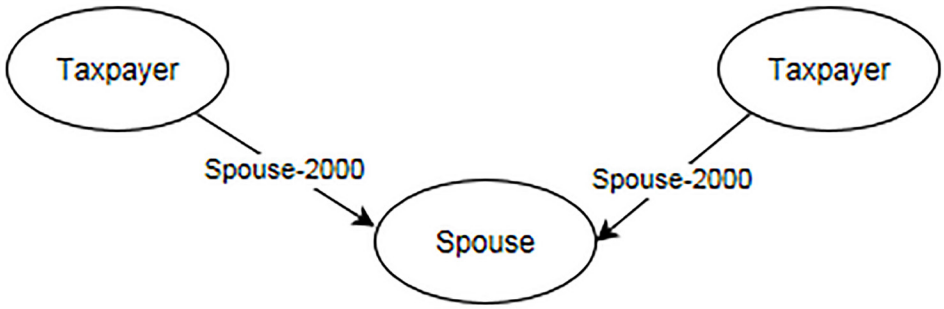
dataset captured by this property graph. In the example shown in Figure 2, Dependent2 becomes a taxpayer in 2004 during which time their AGI falls in the 4th quartile. Dependent2 goes on to claim a spouse (Spouse3) in 2006 and a dependent (Dependent3) in 2008.

Querying the Databank Graph Database for Analysis

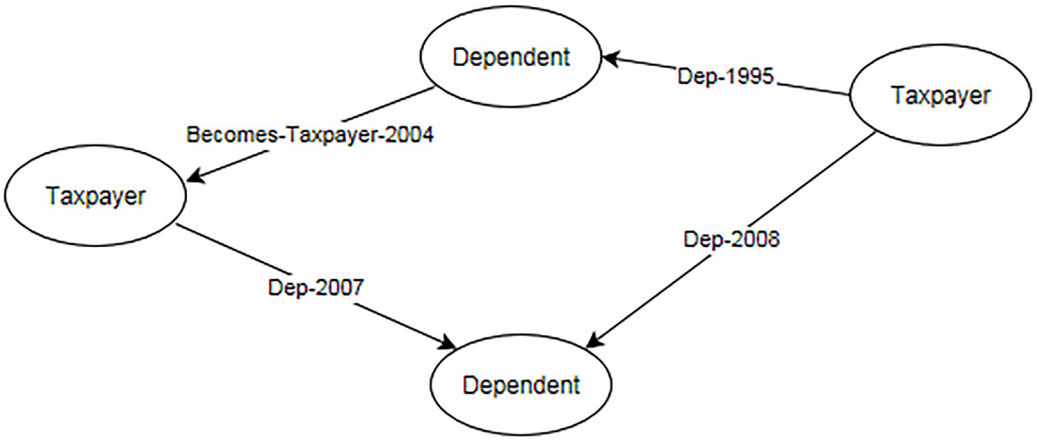
This section will show the value of a graph database by taking a few use cases for the graph database model of the Databank. The introduction above provided a quick primer on how queries need be structured to extract information. That will be applied to the examples here.

a. Find taxpayers who claimed the same spouse during a given tax year (say 2000). This is very simple to conceptualize: look for two or more taxpayer nodes whose spouse links point to the same spouse node:

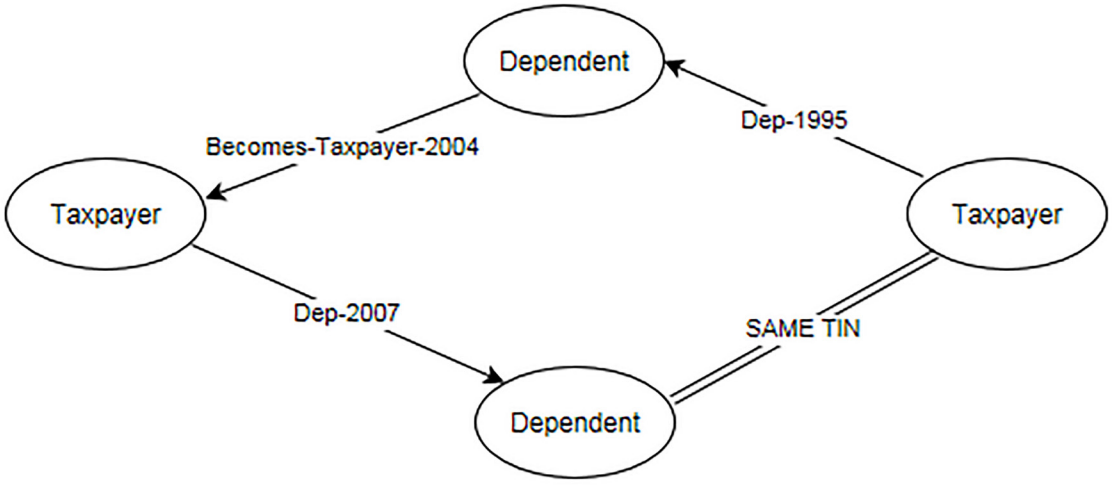
```
MATCH (t1:Taxpayer) - [:Spouse-2000] -> (:Spouse) <- [:Spouse-2000]
- (t2:Taxpayer)
RETURN t1, t2
```



b. Show taxpayers who claimed their grandchildren as dependents during a given tax year (say 2008). This query can very easily be conceptualized on a graph database. Look for a dependent link between a taxpayer (say T1) and dependent (say D1) in a year before the year in question. Then look for a point in time when taxpayer T1 was claimed as a dependent by their parent (say T2). Finally, look for a dependent link between T2 and D1 for the given tax year.



c. For a given tax year, show taxpayers who claim their parents as dependents. This is a very simple variation on the previous query. In this case D1 and T2 are the same (they will have the same TIN).



d. Identify children who have migrated from their home state. In this instance, follow the dependent links when they become taxpayers and compare the ZIP Codes of the parents and dependents.

e. Identify cases where children's income has dropped from their parents'. This is very similar to the previous use case—here instead of comparing ZIP Codes compare the AGIs.

f. Identify low income taxpayers with many dependents. Here we look for taxpayers in a lower quartile and follow the links to their dependents and count those links.

Conclusion and Future Work

Graph databases can provide significant advantages over relational databases where it is important to traverse links in a data set. In relational databases these links are traversed via multiple joins that be very costly especially if the size of the database is large. If the data don't undergo frequent updates, then modeling the data as a graph can make link traversals very efficient and intuitive.

This paper has shown how certain problems that have traditionally been the domain of relational databases, can be modeled as graphs and then queried for insights. One approach to modeling a large dataset like the IRS Databank—that is longitudinal in nature—as a graph database was presented. Some use cases of data analysis were considered where the power of traversing links was demonstrated. Future continuing work will involve the loading of the entire databank into a working graph database that can be made available to analysts and researchers.

References

Frisendal, T. (2017). *Property Graphs: The Swiss Army Knife of Data Modeling*. Retrieved from Dataversity: https://www.dataversity.net/property-graphs-swiss-army-knife-data-modeling/.

Joshi, V. (2017). *A Gentle Introduction to Graph Theory*. Retrieved from Medium: https://medium.com/basecs/a-gentle-introduction-to-graph-theory-77969829ead8.

Neo4j. (2019). *The Neo4j Cypher Manual v3.5*. Retrieved from Neo4j Website: https://neo4j.com/docs/cypher-manual/current/.

Research, Applied Analytics, and Statistics. (2019). *Compliance Data Warehouse*. Retrieved from IRS RAAS Website (IRS Intranet): https://cdw.web.irs.gov/.

Raj Chetty, E. S. (2018). The SOI Databank: ACase Study in Leveraging Administrative Data in Support of Evidence-Based Policymaking. *Statistical Journal of IAOS*.

upcounsel. (2019). *Pass Through Entity: Everything You Need to Know*. Retrieved from Upcounsel Website: https://www.upcounsel.com/pass-through-entity.

5

Appendix

Conference Program

9th Annual IRS-TPC Joint Research Conference on Tax Administration
Urban Institute, 2100 M Street, NW, Washington, DC
June 20, 2019

Program

8:30 – 9:00 Check in

9:00 – 9:15 Opening Remarks

Welcome Eric Toder (Co-Director, Tax Policy Center) and
Barry Johnson (Director, Statistics of Income, IRS, RAAS)
Charles Rettig (IRS Commissioner)

9:20 – 10:50 Estimating the Effects of Tax Administration on Compliance

Moderator: *Robert McClelland (Urban-Brookings Tax Policy Center)*

- Estimating the Specific Indirect Effect for Multiple Types of Correspondence Audits
  Ben Howard, Lucia Lykke, Leigh Nicholl (MITRE Corporation), and Alan Plumley (IRS, RAAS)
- Enforcement vs. Outreach - Impacts on Tax Filing Compliance
  Anne Herlache, Stacy Orlett, Rizwan Javaid, Ishani Roy, and Alex Turk (IRS, RAAS)
- Assessing the Impact of Exchange of Information
  Pierce O'Reilly (OECD)

Discussant: *Michael Udell (District Economics Group)*

10:50 – 11:00 Break

11:00 – 12:30 The Influence of External Factors on Compliance

Moderator: *George Contos (IRS, Communications & Liaison)*

- Recent Changes in the Paid Return Preparer Industry and EITC Compliance
  Emily Y. Lin (US Treasury, Office of Tax Analysis)
- Taxpayer Responses to Third-Party Income Reporting: Evidence from Spatial Variation across the US
  Bibek Adhikari, Timothy F. Harris (Illinois State University), and James Alm (Tulane University)
- Effect of Recent Reductions in the Internal Revenue Service's Appropriations on Revenues
  Janet Holtzblatt (Tax Policy Center) and Jamie McGuire (Joint Committee on Taxation)

Discussant: *Alan Plumley (IRS, RAAS)*

12:30 – 1:00 Lunch

1:00 – 1:30 Keynote Speaker

Richard Rubin (US Tax Policy Reporter, Wall Street Journal)

1:30 – 3:00 Improving the Digital Taxpayer Experience

Moderator: *Alcora Walden (IRS, Office of Online Services)*

- Online Account User Testing
  Heather Gay (Mediabarn Inc.)
- Accessible Authentication for All: An Evaluation Framework for Assessing Usability and Accessibility of Authentication Methods
  Becca Scollan and Ronna ten Brink (MITRE Corporation)
- Understanding the Voice of the Taxpayer through the User-Centered Design Paradigm
  Nikki Kerber, Kristen Papa, and Jake Sauser (Booz Allen Hamilton)

Discussant: *Courtney Rasey (IRS, Wage & Investment Division)*

3:00 – 3:10 Break

3:10 – 4:40 Understanding the Drivers of Taxpayer Behavior

Moderator: *Melissa Vigil (IRS, RAAS)*

- Underpayment of Estimated Tax: Understanding the Penalized Taxpayer Population
  Victoria Bryant, Brett Collins, Janet Li, Alicia Miller, Alex Turk, and Tomás Wind (IRS, RAAS), and Stacy Orlett (IRS, SB/SE)
- The Effect of Audit Burden on Subsequent Tax Evasion
  Amy Hageman (Kansas State University), Ethan LaMothe (University of South Carolina), and Mary Marshall (Louisiana Tech University)
- Using a Graph Database to Analyze the IRS Databank
  Ririko Horvath and Rahul Tikekar (IRS, RAAS)

Discussant: *Brian Erard (Brian Erard & Associates)*

4:40 – 4:50 Wrap-up

Eric Toder (Co-director, Urban-Brookings Tax Policy Center)