The IRS Research Bulletin

Proceedings of the 2020 IRS / TPC Research Conference

Research, Applied Analytics & Statistics

Publication 1500 (Rev. 5-2021) Catalog Number 11546J Department of the Treasury **Internal Revenue Service** www.irs.gov

Papers given at the

10th Annual Joint Research Conference on Tax Administration

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

Held Virtually
June 18, 2020

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

*Prepared under the direction of Barry W. Johnson, Acting IRS 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 virtually on June 18, 2020. 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. Many people participated in this, our first fully virtual conference. Videos of the presentations are archived on the Tax Policy Center website to enable additional participation. Attendees 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. The remainder of the conference included sessions on behavioral responses to audits, new insights on taxpayer behavior, advances in taxpayer service, and doing more with less. The keynote speaker was former IRS Commissioner Charles Rossotti, who explained his proposals for improving tax compliance.

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, Brett Collins, Tom Hertz, Lindsay Schrock, and John Weaver (RAAS); Fran Cappelletti (Taxpayer Advocate); Melissa Hayes (Large Business & International Division); Quinton Anderson and Alexis Kinikin (Small Business / Self-Employed Division); Alcora Walden (Online Services); and Rob McClelland (Tax Policy Center). In addition, Ann Cleven and Hailey Roemer from the Tax Policy Center oversaw numerous details to ensure that the conference ran smoothly.

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

10th Annual IRS-TPC Joint Research Conference on Tax Administration

Contents

4. Doing More With Less

5. Appendix

1

Behavioral Responses to Audits

Erard ◆ Kirchler ◆ Olsen

Lykke ◆ McGill ◆ Nicholl ◆ Plumley

Alm ◆ Kasper

The Specific Deterrence Implications of Increased Reliance on Correspondence Audits

Brian Erard (B. Erard & Associates), and Erich Kirchler and Jerome Olsen (University of Vienna)[1]

Introduction

Tax administrations rely on audits as a key tool for promoting and enforcing tax compliance. In addition to bringing in revenue directly through additional tax assessments, audits potentially raise revenue indirectly by improving "voluntary" tax reporting. This can happen in two ways. First, the perceived risk of an audit may discourage taxpayers from underreporting their tax liability, thereby inducing a "general deterrent" effect. Second, the actual experience of an audit may encourage taxpayers to become more (or less) compliant in their future tax-reporting behavior, thereby eliciting a "specific deterrent" effect. In this paper, we focus on the latter effect.

The specific deterrent effect of an audit is likely to depend on the nature of one's audit experience. Over time, the Internal Revenue Service (IRS) has increasingly shifted away from performing face-to-face examinations in favor of conducting audits by mail. Whereas face-to-face audits accounted for the majority (62 percent) of all examinations of returns filed in 1990, the lion's share (81 percent) of all audits of returns filed in 2017 were conducted through correspondence.[2] In comparison with face-to-face examinations, correspondence audits tend to be more narrowly focused and impersonal. At the same time, they are less costly and burdensome for both the taxpayer and the tax agency. In this paper, we summarize our recent research on the implications of the administrative shift towards audits by mail for the future tax-reporting behavior of self-employed taxpayers.[3]

Estimation Methodology

To control for differences in characteristics among taxpayers who have experienced a correspondence audit, a face-to-face audit, or no audit when estimating specific deterrent effects, we rely on the inverse probability of treatment weighting (IPTW) methodology. This estimation methodology requires no assumptions about the functional relationship between the determinants of audit selection and taxpayer reporting behavior. Under this approach, one begins by estimating the propensity scores (predicted probabilities), π_i^c, π_i^f, and π_i^{na}, associated with a correspondence audit, a face-to-face audit, and no audit, respectively. Define the indicator variable for taxpayers in the sample who did not experience an audit as I^{na}, and denote the outcome variable as *y*. The Horvitz-Thompson estimator of the expected counterfactual outcome among taxpayers receiving a correspondence audit had they not been audited is then defined as:

$$\left(\frac{1}{\sum I_i^{na}}\right)\sum\left[I_i^{na}\, y_i \left(\frac{\pi_i^c}{\pi_i^{na}}\right)\right].$$

[1] This paper is based on research conducted for the National Taxpayer Advocate (NTA) under contract TIRNO-14-E-00019 with technical support from NTA Technical Advisor Jeff Wilson. Any opinions expressed in this document are those of the authors and do not necessarily reflect the views of the National Taxpayer Advocate. We are grateful to Sebastian Beer and Matthias Kasper for their technical guidance. We also thank our discussant, Janet Holtzblatt, and the other participants at the 2020 IRS-TPC Research Conference for their many helpful comments.

[2] Authors' calculations based on Internal Revenue Service (1992, Table 11, p. 24) and Internal Revenue Service (2019, Table 9a, p. 23).

[3] A more detailed presentation of our research findings is provided in Erard *et al.* (2020).

Similarly, the Horvitz-Thompson estimator of the expected counterfactual outcome among taxpayers receiving a face-to-face audit had they not received an examination is defined as:

$$\left(\frac{1}{\sum I_i^{na}}\right)\sum\left[I_i^{na}\, y_i \left(\frac{\pi_i^f}{\pi_i^{na}}\right)\right].$$

These counterfactual outcome estimates are thus computed as a weighted average of the outcomes observed for the unaudited taxpayers in the sample, where the weights are computed as the ratio of the relevant propensity scores.[4] Intuitively, greater weight is applied to unaudited taxpayers with a relatively high predicted probability of selection for the specified type of audit, as taxpayers with their characteristics will tend to have greater representation among the sample of filers who received that type of audit. The estimated specific-deterrent effect is computed as the difference between the actual mean outcome for taxpayers who received the specified type of audit and the estimated counterfactual outcome for these taxpayers.

In our analysis, we rely on propensity scores derived from a multinomial logit model of audit selection.[5] This analysis allows for three possible audit selection outcomes: (1) no audit; (2) correspondence audit; or (3) face-to-face audit. We have constructed a set of over 60 candidate explanatory variables for the audit selection process. Included among these covariates are measures of the current and prior year DIF-scores that are relied upon by the IRS to help identify high-risk returns for face-to-face examination. A sequential selection process is employed to choose the final set of covariates.

Some taxpayers in our sample were audited prior to filing the next year's tax return, and others were audited after doing so. To account for differences in the audit selection process for these two groups, a separate multinomial logit analysis is performed for each group. The estimation results are employed to predict the odds of a correspondence audit and the odds of a face-to-face audit (relative to no audit) for each taxpayer in the estimation sample.

For taxpayers who were audited prior to filing the next year's tax return, our outcome variable is the difference between the natural log of reported tax for a subsequent tax year (either of the next two filed tax returns) and the natural log of reported tax on the audited return. Effectively, then, this approach produces "difference-in-differences" estimates to account for unobserved time-invariant differences between the audited and unaudited taxpayers in our sample. A one-year ahead impact estimate is derived using the very next year as the subsequent tax year, while a two-year ahead impact estimate is obtained using the following year as the subsequent tax year.

In the case of taxpayers who were audited only after filing their tax return for the following year, we rely on our one-period ahead impact estimate as a "placebo test." Since these taxpayers were not aware of the audit until after they had filed a return for the following year, the one-period ahead impact estimate has an expected value of zero. Therefore, this placebo test provides a useful check on the quality of the matching process. The two-period ahead impact estimate calculated as described above is effectively a one-period ahead estimate for this group, since the return filed 2 years later was the first return that was filed subsequent to the initiation of the audit.

Data

The data for this study include detailed line-item information from returns filed by audited and unaudited self-employed taxpayers. The audit sample consists of Schedule C filers who experienced an audit of the return they filed for the relevant tax year (2010 or 2014). A stratified random sample of taxpayers who did not experience an audit was also drawn. The sampling for each of these groups was subject to certain restrictions meant to help isolate the audit impact, such as a documented history of filing returns over a period of years and no recent prior examinations before the audit year in question.

[4] In our application, we follow the conventional approach to stabilizing the weights used in our analysis.

[5] We employ sample weights in estimation to account for the choice-based nature of our data sample.

The members of the unaudited taxpayer sample were selected to provide a means for developing counterfactual estimates of behavior for the members of the audit sample. To this end, unaudited taxpayers with relatively high audit risk scores ("DIF scores") were oversampled. Sample weights were introduced so that the weighted sample was representative of the more general population of unaudited taxpayers.

For each taxpayer in the two data samples, detailed line item information was collected from each tax return filed for the reference audit year, the two prior years, and the two subsequent years. Table 1 presents the numbers of audited and unaudited taxpayers that were sampled.

TABLE 1. Sample Count of Taxpayers by Audit Type

Reference Audit Year	Correspondence	Face-to-Face	Unaudited
Tax Year 2010	40,359	12,541	421,309
Tax Year 2014	13,629	3,274	377,168

Results

The estimated one-period and two-period ahead specific deterrent effects for correspondence and face-to-face audits of self-employed taxpayers are presented in Table 2. These estimates reflect the predicted percentage change in reported tax liability as a result of the examination.

For audits that were initiated after the return for the following year was filed (but before the 2nd subsequent return was filed), a placebo impact estimate is provided for the next year's return. The expected audit impact is equal to zero in this year because the taxpayer would not have been aware of the audit when that return was filed. The estimated impact for each audit type is in fact small and (with the exception of correspondence audits for Tax Year 2014) statistically insignificant, consistent with expectations. This finding helps to substantiate the validity of the estimation methodology.

The estimation results indicate that face-to-face audits have a very large specific deterrent effect. For Tax Year 2010 audits that began prior to the filing of the Tax Year 2011 return, reported tax liability is estimated to have increased by 40.8 percent for Tax Year 2011 and 27.3 percent for Tax Year 2012 as a result of the examination. For audits that began after the Tax Year 2011 return was filed, reported tax liability is estimated to have increased by 37.5 percent in Tax Year 2012. The estimated impacts are even larger for Tax Year 2014 audits. For audits that began prior to the filing of the Tax Year 2015 return, reported tax liability is estimated to increase by more than 95 percent in Tax Year 2015 and remain around that level the following tax year. For Tax Year 2014 audits that began after the Tax Year 2015 return was filed, reported tax liability is estimated to increase by nearly 62 percent on the first return filed since the audit was initiated (Tax Year 2016).

The estimation results for correspondence audits are more nuanced. For audits that began prior to the filing of the Tax Year 2011 return, there is evidence of a marked counter-deterrent effect. Reported tax liability is estimated to have declined by 7.3 percent in Tax Year 2011 and 8.3 percent in Tax Year 2012 as a result of the examination. On the other hand, reported tax liability is estimated to have been 37.5 percent higher in Tax Year 2012 for taxpayers whose Tax Year 2010 audits were initiated later in the examination cycle.

TABLE 2. Predicted Percentage Change in Reported Tax Liability by Audit Type and Tax Year

Audit Type	Audited Before Next Return Filed		Audited After Next Return Filed	
	1st Year Impact	2nd Year Impact	Placebo Impact	1st Year Impact
Tax Year 2010 Audit Results				
Correspondence	-7.3%*	-8.3%*	-1.2%	37.5%*
Face-to-Face	40.8%*	27.3%*	4.1%	37.5%*
Tax Year 2014 Audit Results				
Correspondence	-5.7%*	-15.0%*	9.4%*	61.1%*
Face-to-Face	95.3%*	97.3%*	9.5%	61.8%*

*Statistically significant at the 5% level.

The disparity among the findings within the correspondence audit group may reflect differences in the types of issues or taxpayers that are addressed over the correspondence audit cycle. Based on a preliminary analysis of audit findings, approximately half of all correspondence audits involving self-employed taxpayers are initiated before the taxpayer has filed a return for the following tax year. A very substantial share (over 70 percent) of these early audits involves taxpayers who claim the Earned Income Tax Credit (EITC).[6] In contrast, only about 19 percent of the audits initiated later in the cycle (after the return for the following tax year has been filed) involve EITC claimants. To investigate whether the estimated counter-deterrent effect from early audits is associated with EITC claims, we extended our analysis to develop separate impact estimates for EITC claimants and nonclaimants. The results suggest that the findings are *not* attributable to such claims. A remaining possibility is that the heterogeneity in outcomes may be attributable to differences in the characteristics of the taxpayers selected at different points in the examination cycle. Alternatively, it could have to do with the amount of time that lapses between filing a return and being notified of an audit. Further research is needed to evaluate these possibilities.

In addition to extending our model to separately investigate EITC claimants and nonclaimants, we have performed some sensitivity analyses involving alternative estimation methodologies and additional tax years. The results of these supplemental analyses corroborate our main findings.

Conclusion

An important purpose of audits beyond immediate revenue generation is to discourage future reporting noncompliance. In this paper, we have conducted a preliminary analysis of how risk-based operational audits impact future reporting behavior, and we have paid special attention to how correspondence and face-to-face examinations may differ in this regard. Our estimation results indicate that correspondence audits that take place later in the examination cycle (after the subsequent tax return has been filed) are comparable to face-to-face audits in terms of their impact on future reporting behavior. Both types of audits have a substantial pro-deterrent effect when they are initiated after the following year's tax return has been filed. In contrast, however, correspondence audits that take place early in the audit cycle are actually associated with a counter-deterrent effect. Reported tax liability is estimated to fall by 6 to 15 percent in the first two tax years following the initiation of the audit. This is an important finding, because approximately half of all correspondence examinations take place early in the audit cycle.

Overall, then, the results of this study suggest that correspondence audits are not a perfect substitute for face-to-face examinations. Not only do they tend to be more narrowly targeted and impersonal, they also appear to be less consistent in terms of improving future taxpayer reporting behavior. This raises concerns about

[6] Some of these correspondence audits involve issues beyond the EITC.

IRS' increasing reliance on this form of enforcement. The disparate findings for correspondence audits that take place at different points in the audit cycle may reflect differences in the types of issues or taxpayers that are addressed over the cycle. Alternatively, the amount of time that lapses between filing a return and notification of an audit may have a direct impact on future reporting behavior. Further research is needed to understand this result. More generally, the findings suggest that further study on the proper balance between face-to-face and correspondence audits is warranted.

References

Advani, A., W. Elming, and J. Shaw. 2015. How long-lasting are the effects of audits? TARC Discussion Paper 011–15. Exeter, UK: Tax Administration Research Centre, University of Exeter.

Allingham, M. G., and A. Sandmo. 1972. Income tax evasion: A theoretical analysis. *Journal of Public Economics,* 1(3–4), 323–338.

Beck, P., and W. O. Jung. 1987. An economic model of taxpayer compliance under complexity and uncertainty. *Journal of Accounting and Public Policy,* 8, 1–27.

Beer, S., M. Kasper, E. Kirchler, and B. Erard. 2015. Audit impact study. *National Taxpayer Advocate 2015 Annual Report to Congress,* 2, 68–98.

DeBacker, J., B. T. Heim, A. Tran, and A. Yuskavage. 2018. Once bitten, twice shy? The lasting impact of IRS audits on individual tax reporting. *The Journal of Law and Economics,* 61(1), 1–35.

Erard, B. 1992. The influence of tax audits on reporting behavior, in *Why people pay taxes: Compliance and enforcement,* ed. Joel Slemrod, Ann Arbor: University of Michigan Press, 95–114.

Erard, B., M. Kasper, E. Kirchler, and J. Olsen. 2019. What influence do IRS audits have on taxpayer attitudes and perceptions? Evidence from a national survey. *National Taxpayer Advocate 2018 Annual Report to Congress,* 2, 77–130.

Erard, B., E. Kirchler, and J. Olsen. 2020. Audit impact study: The specific deterrence implications of increased reliance on correspondence audits. *National Taxpayer Advocate 2019 Annual Report to Congress,* 257–268.

Feld, L., and B. S. Frey. 2003. Deterrence and tax morale: How tax administrations and taxpayers interact, *European Review,* 11(3), 385–406.

Francesco, G., and L. Mittone. 2005. Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology,* 12, 495–515.

Frey, B. S., M. Benz, and A. Stutzer. 2004. Introducing procedural utility: Not only what, but also how matters. *Journal of Institutional and Theoretical Economics,* 160, 377–401.

Frey, B. S. 2011. Punishment—and beyond, *Contemporary Economics,* 5 (2), 90–99.

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. 1992. *IRS Annual Report,* 1991. Retrieved from https://www.irs.gov/pub/irs-soi/91dbfullar.pdf.

Internal Revenue Service. 2019. *IRS Data Book,* 2018. Retrieved from https://www.irs.gov/pub/irs-prior/p55b--2019.pdf.

Kastlunger, B., E. Kirchler, L. Mittone, and J. Pitters. 2009. Sequences of audits, tax compliance, and taxpaying strategies. *Journal of Economic Psychology,* 30, 405–418.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez. 2011. Unwilling or unable to cheat? Evidence from a randomized tax audit experiment in Denmark. *Econometrica,* 79(3), 651–692.

Maciejovsky, B., E. Kirchler, and H. Schwarzenberger. 2007. Misperceptions of chance and loss repair: On the dynamics of tax compliance. *Journal of Economic Psychology,* 28(6), 678–691.

Mittone, L., F. Panebianco, and A. Santoro. 2017. The bomb-crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology,* 61, 225–243.

Scotchmer, S., and J. Slemrod. 1989. Randomness in tax enforcement. *Journal of Public Economics,* 38(1), 17–32.

The Specific Indirect Effect of Correspondence Audits: Moving from Research to Operational Application

Lucia Lykke, Max McGill, 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 return that was the subject of the contact. These enforcement actions undoubtedly also have an indirect effect on revenue: a change in the current or future behavior of taxpayers who either have experienced an enforcement contact (the "specific" indirect effect) or have some knowledge or perception about others' tax enforcement experiences (the "general" indirect effect). In a 2012 report, the Government Accountability Office (GAO) called for the Internal Revenue Service (IRS) to account for these indirect effects when allocating resources to different types of tax enforcement actions (GAO (2012)). The IRS already allocates resources to enforcement activities on the basis of the expected direct revenue effect, which differs across activities; however, current practices do not account for indirect revenue, and it is unknown whether and how indirect revenue varies across enforcement activities. To account for this requires estimating how audits affect taxpayers' future contributions to IRS revenue across many types of enforcement activities and translating those estimates in such a way that they are usable in a resource allocation context. This is the topic of this study.

This study presents estimates of the specific indirect effects on taxpayers following one of five types of correspondence audits, using longitudinal taxpayer data obtained by the IRS through operational audits conducted on individual tax returns filed within the Tax Year (TY) 2006 through TY 2012 period. There are more than five types of correspondence audits conducted by the IRS, each with its own candidate audit population and specific procedures; we present five here, however, to demonstrate how indirect revenue, as measured by total tax reported in the years after audit, can be used in conjunction with direct revenue collected from the audit in the context of making budget decisions in the correspondence audit program. To our knowledge, this is the first study in the U.S. tax enforcement context that explicitly aligns estimating the specific indirect effect with audit operations in such a way that our estimates can be used when deciding how many of which types of correspondence audit will yield the greatest overall return on investment.

Comparing the subsequent reporting behavior of taxpayers who experienced different types of audits has research value as well as operational value. Much of the prior literature on the indirect effect of audits 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.* (2018a); 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 respond differently to different types of audits. We expand estimation of the indirect effect of a correspondence audit to include many disparate populations that are audited for different individual tax return line items and reasons. In this way, we advance the state of the field of audit impacts to better understand how audits affect different taxpayer groups' subsequent reporting.

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 controlling for the specific operational metrics applied to each return to determine a return's audit eligibility within each category and the priority given to it among all eligible returns in that category. This 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. Notably, the priority variable we can derive for each historical return can be considered the best available surrogate for that return's propensity to have been selected. Although we are not claiming a causal effect, the ability to control for selection priority is an important advantage to our modeling.

Additionally, applying indirect effects to resource allocation decision-making requires operationalizing the indirect effect in a way that is usable for this purpose. This is a challenge because audited taxpayers are not habitually re-audited to determine changes in compliance after the initial audit. We address this challenge by using differences in changes to estimated total tax between our two taxpayer groups—audited and not audited—over time as a proxy for indirect revenue, and we translate model estimates into tax dollar values that can be compared with direct revenue.

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 one of five types of correspondence audits. We do this by comparing audited taxpayers' post-audit tax reporting to the tax reporting of not-audited taxpayers who were eligible during the same tax year.
2. Translate estimates of total tax reporting for the audited and not-audited populations into dollar values that can be compared with audit direct revenue for each type of correspondence audit.

Literature Review

Types of Indirect Effect

Much of the literature and research conducted on taxpayer compliance behavior rests 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.

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 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 Different Taxpayer Populations

In this study, we build on prior work that suggests that variations in population characteristics, and the nature of dissimilar categories or types of audits, may be differently associated with subsequent-year reporting. 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. This points to the fact that we should not expect all taxpayer populations to respond in the same manner to the experience of being audited. 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, and increases reported Schedule C income by 14.2 percent, on average. 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[1] after an audit (Kleven *et al.* (2011)). Confirming the conclusion about the importance of third-party reporting for an indirect effect,[2] U.K. taxpayers audited at random increased reported tax liability substantially over a 4-year period for taxpayers who filed self-assessed[3] 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 an 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 IRS discriminant function (DIF) computer 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 and found evidence that the indirect effect among Schedule C filers varied depending on the audit outcome. 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,[4] 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)). This suggests that the indirect effect is stronger and longer lasting when the taxpayer's audit experience has stronger consequences (additional tax liability) than simply the experience of being audited regardless of outcome.

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 taxpayers who report capital gains and losses, list 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

[1] 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.

[2] 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)).

[3] 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)).

[4] 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).

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 EITC claiming decreases after experiencing an audit; after a random NRP audit, taxpayers who claimed EITC decreased their future EITC claiming (DeBacker *et al.* (2018b)), and taxpayers who were audited operationally for EITC credit validity also reduced EITC claiming in subsequent years, especially within the first year after audit (Guyton *et al.* (2018)).

Operational Context for this Study: Correspondence Audits

This study focuses on correspondence audits as a first step toward a broader model of how the IRS can incorporate indirect effects into resource allocation decision-making. Correspondence audits are one of many types of audit activities conducted by the IRS's Small Business/Self-Employed (SBSE) Division (Rettig (2016)). In a typical correspondence audit, SBSE identifies narrowly defined individual taxpayer populations for examination, and corresponds with the taxpayer by mail to examine specific line items. As such, a correspondence audit is *not* a comprehensive examination of the entire tax return.

Correspondence Audit Resource Allocation

Currently, the correspondence audit inventory (that is, the number of correspondence audits of different types that are performed) is based on the direct revenue—the tax adjustments—resulting from audits performed in the past several years. In this study, we present a method for incorporating taxpayers' subsequent year tax reporting into the revenue estimates for a type of correspondence audit that can be used to make decisions about the correspondence audit inventory. Currently, there are approximately 40 types of correspondence audits, each with its own unique populations of interest and selection criteria. We start with five categories for the purposes of this study.

Current resource allocation decisions are not made based on whether different types of audits are likely to result in an adjustment. That is, some audits will not result in evidence of noncompliance (overstating deductions or credits, or otherwise underreporting tax liability), but instead will result in *no* adjustment to the return and therefore no direct revenue. There is evidence that this type of audit, known as a "no-change" audit, has a different effect on subsequent taxpayer reporting than a "change" audit (e.g., Beer *et al.* (2015)), which suggests that it could be fruitful, in principle, to consider accounting for changes versus no-changes in resource allocation. However, if the IRS knew in advance which audits would result in no change, they would generally not conduct those audits at all, so changing resource allocations to account for a different (especially a smaller) indirect effect among no-change audits would be pointless. Therefore, we focus on estimates of the indirect effect for audited versus not-audited taxpayers in this study. However, we present for academic interest in Appendix 2 a supplementary analysis that includes separate estimates for audits that resulted in a change versus no-change outcome.

Categories of Correspondence Audits

In this study, we compare taxpayers from five distinct audit categories. For each category, we compare all audited taxpayers to a sample of eligible, but not-audited taxpayers (our two-group comparison). We selected categories of correspondence audit that were active for the majority of the study period (TYs 2006–2012),[5] 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 control group 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.

[5] Audit Category 4 was active from TYs 2008–2012.

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

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.

Audit Category 4: Examines some education-related credits on Form 1040 among taxpayers who met certain category-specific eligibility criteria. Note that this Audit Category started in TY 2008, so our data do not include audits or eligible taxpayers for TYs 2006 and 2007.

Audit Category 5: Examines Schedule E passive activity loss reporting among taxpayers who met certain category-specific eligibility criteria.

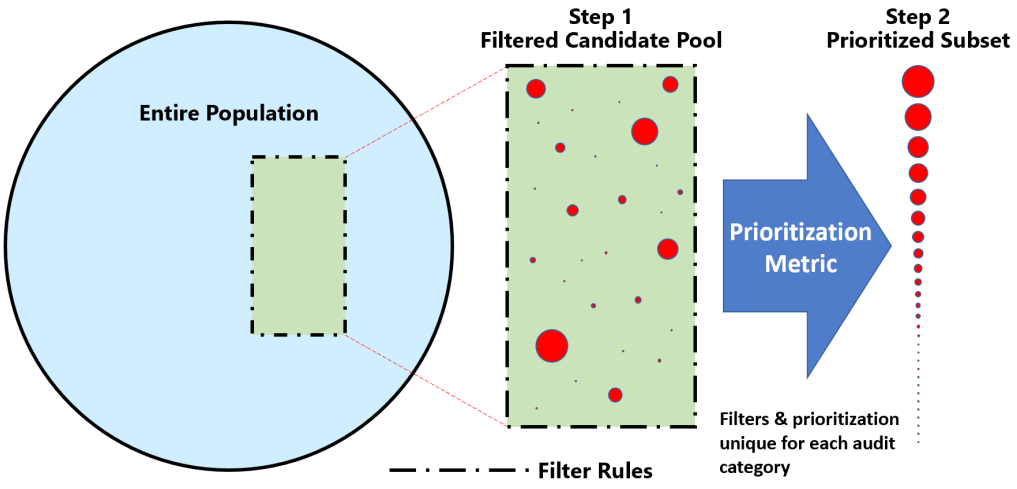
Correspondence Audit Selection

To identify narrowly defined populations of eligible taxpayers for correspondence audits, SBSE applies what Rettig (2016) calls "user-developed criteria" to identify eligible taxpayers. Each category of correspondence audit has its own user-developed criteria to identify the eligible pool, and to prioritize this candidate pool for examination. Figure 1 shows this identification and selection process. This is crucial to our study because we can account for the two key parts of this process to reduce endogeneity and selection bias by operationalizing two features:

1. **Filter rules:** We operationalize the user-developed criteria that SBSE applies to identify a candidate population for examination. This allows us to compare audited taxpayers to a comparable "eligible" group, rather than comparing audited taxpayers to the full universe of individual taxpayers.

2. **Prioritization metric:** 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 sensitivity. 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]

FIGURE 1. The Two-Step Correspondence Audit Selection Process



[7] We have access only to IRS operational documents from the most recent 1 to 3 tax years. As such, we assume that operational criteria stayed relatively stable over time for each correspondence audit type. This is one potential limitation of our study.

Research Questions and Hypotheses

In this study, we address the following research questions.

Total Tax Reporting

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

   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.

Dollar Value Estimates of the Indirect Effect

Research questions 2 and 3 are aimed at generating estimates that can be compared across audit categories. We do not conduct statistical significance testing to determine whether these estimates differ across categories (because this is not how they would be used operationally for resource allocation); therefore, we do not formulate hypotheses.

2. What is the average specific indirect effect of a given audit category, as measured in total attributable tax dollars paid by an individual taxpayer over the 5 years following the tax year of the audited return?

3. What is the ratio of average direct revenue assessed from an audit category for TYs 2006–2012 to the estimated average indirect revenue, as defined by total tax dollars?

Data and Methods

Data

In this study, we combine data on the five 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 TYs 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 whom our queries returned multiple times because they were captured as "eligible" multiple times and were not audited in TYs 2006–2012, we declare the most recent eligibility year as the "baseline" year.

2. For any taxpayers whom 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 whom our queries returned as being eligible in 1 or more years and audited in 1 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 taxpayers, we collected tax return information from

the Form 1040, Schedule A, Schedule C, Schedule SE, and Schedule E for the baseline tax year and up to 8 tax years after (up to TY2018). For example, for baseline TY 2006, we compiled return data up through TY 2014; for baseline year TY 2012, we compiled return data up to TY 2018. We chose to examine up to 8 years after the baseline year 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 non-audited taxpayers available in CDW. We restricted the returned records to a random sample of up to 25,000 taxpayers from the eligible population in each of TYs 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 were fewer than 25,000 eligible taxpayers; in this case we selected all eligible taxpayers regardless of the population size. For these taxpayers, we collected tax return information from the Form 1040, Schedule A, Schedule C, Schedule SE, and Schedule E for the tax year of the baseline year and up to 8 tax years after (up to TY2018).

Dependent Variable

Total Tax. Our primary dependent variable is the 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 (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.

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.[9] 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.[10] 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.[11] 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 the presence of deducting *mortgage interest*. For Audit Categories 1, 3, 4, and 5, we adjust our

[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] In Appendix 2, we present a supplementary analysis that treats the audited group separately as a three-group variable capturing whether the taxpayer was audited with a change, audited with no change, or not audited. Because these estimates are not aligned to current SBSE resource allocation processes, we do not use them for the purpose of illustrating how the indirect effect can be translated into revenue estimates to determine inventory across types of correspondence audit.

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

[11] U.S. Census Bureau, Urban Area Relationship Files: https://www.census.gov/geographies/reference-files/2010/geo/relationship-files.html#par_textimage_470670252.

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 TYs 2006–2018. A binary indicator for whether a paid preparer was used is included as a time-varying covariate. 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, 3, 4, and 5, this is measured in 2018 USD with the interpretation that higher priority is more likely to be audited. This variable is distinct for each audit category and is time-invariant, meaning that it is the taxpayer's assigned priority in the baseline year. Lastly, we control for the taxpayer's reported total tax in the tax year prior to the designated baseline year. This allows the model results to be interpreted as the estimated effects on the j^{th} year's total tax *for two taxpayers with the same tax reporting before they were eligible/selected for audit*.

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 a form of linear regression allowing for repeated measurements on subjects and 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 each taxpayer, which allows them 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. In each model, we interact the audit variable with the number of years since the audit to investigate whether the indirect effect varies over time. Within-taxpayer correlation is modeled with an autoregressive structure, as is common with evenly spaced repeated measures over time. 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 3.5.3, using the modeling package *nlme* (Pinheiro (2020)).

Model: Total Tax Reporting Over Time

For each category of audit, we separately estimate models (1a) and (1b) 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*. Models for Audit Category 2 are not controlled for whether the taxpayer itemized deductions since eligibility for this audit necessitates itemizing deductions. γ_{0i} denotes a random effect on individual *i*.

Model (1) is our two-group model that estimates the natural logarithm of total tax as a function of whether the taxpayer was audited or not audited. As such, $\beta_2\ audited_i$ is a time-invariant measure of whether the taxpayer was audited for the tax return filed at baseline year.

(1)
$$\begin{aligned}\ln(\text{total tax}+1)_{ij} &= \beta_0 + \gamma_{0i} \\ &+ \beta_1 priority_i + \beta_2 audited_i + \beta_{3-10} year\ after\ baseline_{ij} \\ &+ \beta_{11-18} audited_i * year\ after\ baseline_{ij} + \beta_{19} FS_{ij} + \beta_{20} TY_{ij} + \beta_{21} TPI_{ij} \\ &+ \beta_{22} any\ wages_{ij} + \beta_{23} total\ exemptions_{ij} + \beta_{24} any\ Child\ Tax\ Credit_{ij} \\ &+ \beta_{25} itemized\ deductions_{ij} + \beta_{26} mortgage\ interest_{ij} + \beta_{27} urban_{ij} \\ &+ \beta_{28} Total\ Tax_{i,-1} + \beta_{29} audit\ last\ 10\ TYs_i + \beta_{30} preparer_{ij} + \epsilon_{ij}\end{aligned}$$

Estimates of the Dollar Value of the Specific Indirect Effect

In order to use the estimates obtained as part of Model (1) in the context of resource allocation, we must first translate them into a form that quantifies the relative effect that a given category of audit has on taxpayers' future contributions to IRS revenue. Specifically, we convert estimates of the specific indirect effect obtained from the model (measured in a relative form of tax as percent changes) into an absolute form of tax, measured in 2018 USD. This allows us to compare each modeled enforcement activity on the same scale of predicted

revenue value. The specific methodology used to generate these dollar-valued estimates is detailed further in Appendix 1.

Results

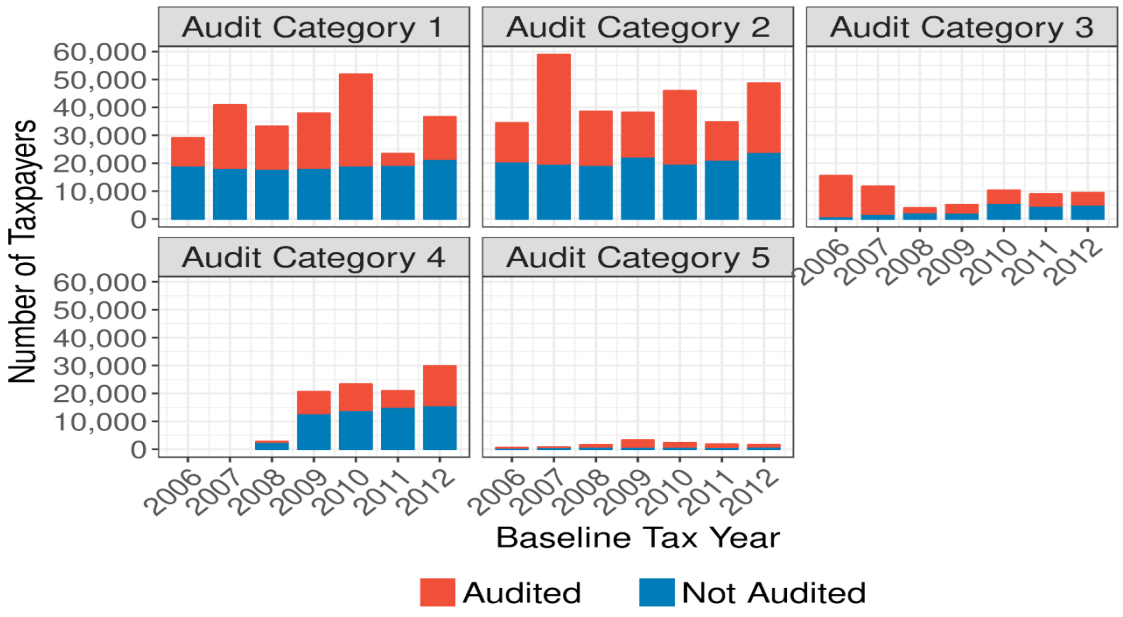
Sample Sizes

The sample sizes of each audit category and baseline year are shown in Figure 2. Audit Categories 1 and 2 have the largest samples with 252,890 and 299,348 taxpayers, respectively. For Audit Category 1, audits were most common in TY 2010 and least common in TY 2011. Similarly, for Audit Category 2, TY 2011 was a lighter year for audits, while TY 2007 has the highest audit frequency. Audit Category 3 has 64,807 taxpayers in total and had relatively few eligible, not-audited taxpayers in TYs 2006 and 2007. Audit Category 4 was not active until TY 2008 and has 97,504 taxpayers total. Lastly, Audit Category 5 is the smallest, with 11,891 taxpayers in total.

Timing of Audits

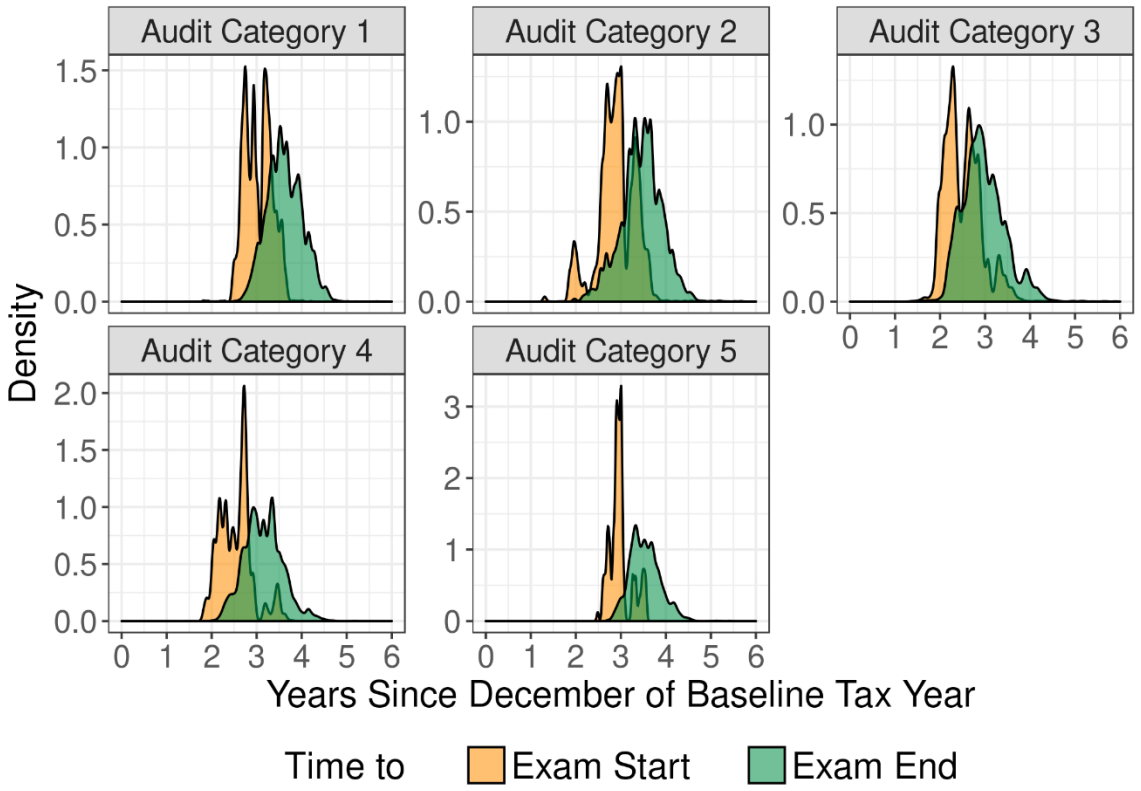
Figure 3 summarizes the time to exam start and end by audit category. We assume that the exam start date coincides with when the taxpayer is notified that their return is being examined, and thus marks when we might expect to observe a behavioral response to the audit. The distribution of time to exam start in Figure 3 indicates that for most taxpayers and all six audit categories, taxpayers are notified of their audit approximately 2 to 3 years after the December of the tax year for which they filed the audited return. Audit Category 5 has on average the longest notification time. Almost all audits open within 4 years after the tax year of the audited return and no taxpayers are notified within the first year. This suggests that if an indirect effect is present, it will most likely not manifest until 2 or 3 years after the tax year of the audited return. For example, if a taxpayer is audited for their TY 2008 return, which encompasses taxes paid through December 2008, they are likely to know about this audit by December 2011. They will file their TY 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 TY 2011 return (3 years after the baseline year).

FIGURE 2. Sample Sizes by Audit Category and Baseline TY



NOTE: Baseline tax year is defined as the tax year the return entered our sample due to audit or eligibility.

FIGURE 3. Densities of Time to Exam Start (orange) and End (green) by Audit Category



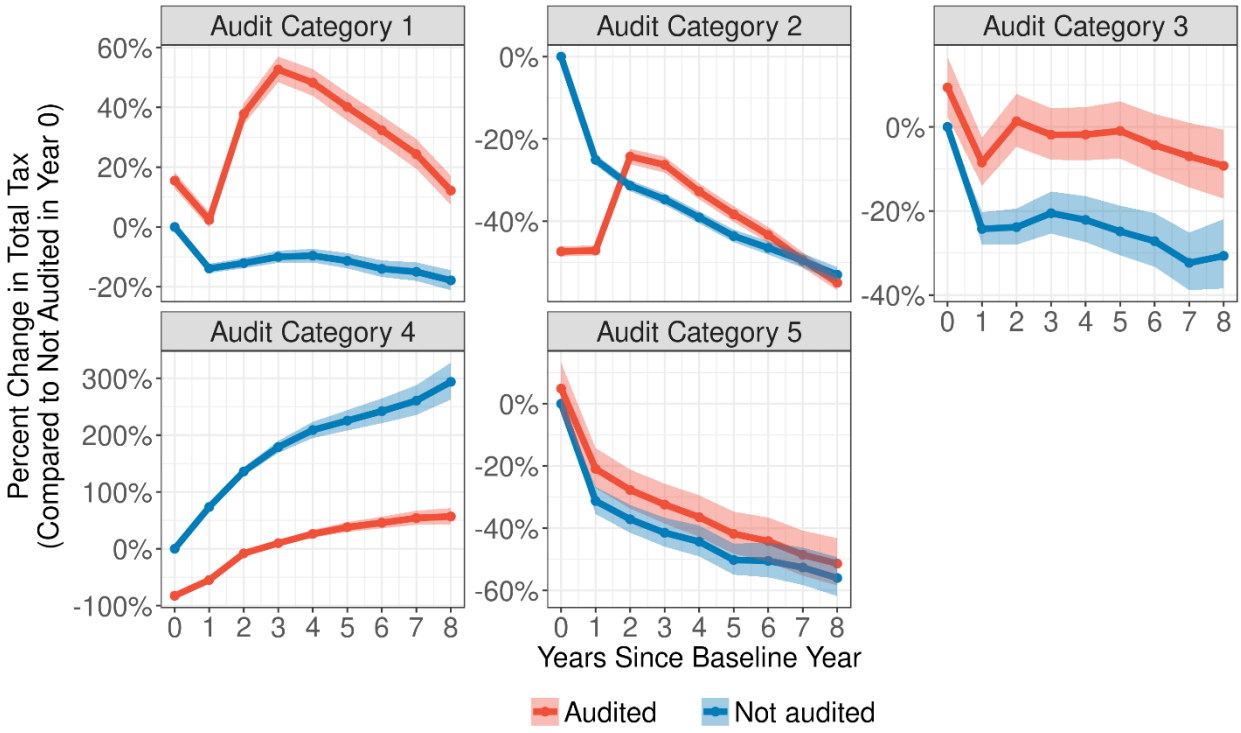
Modeling Results: Total Tax

The estimated coefficients from the linear mixed effects model for the log of total tax are displayed in Figure 4 and may be referenced in Table 1. The estimated coefficients of audit-time interaction are visualized alone in Figure 5.

Audit Category 1

Table 4 displays the estimates from the total tax model for Audit Category 1 for the audited versus not-audited groups, which deals with Schedule C line items. Figure 4 shows the estimated changes in total tax over time for the audited and not-audited groups based on the 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 15 percent more tax than that of the not-audited taxpayers (95-percent confidence interval (CI) 13–18), while holding the control variables constant. One year after baseline, the two groups decrease at roughly the same rate. However, in year 2, reporting among the audited group increases sharply, while that of the not-audited group remains relatively constant. Beyond 3 years after the baseline year, both groups show evidence of decreasing total tax over time when adjusting for control variables.

FIGURE 4. Estimated Percent Change in Total Tax Using Audit, Year, and Audit-Year Interaction Coefficients



NOTE: Shading represents 95-percent confidence intervals.

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 47 percent less than that of the not-audited taxpayer in the same year (CI 46–49). While there is evidence that not-audited taxpayers decrease 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 the baseline year. The slope of the audited taxpayers is estimated to decrease beginning 3 years after baseline, while the not-audited estimated total tax continues to decrease as well.

Audit Category 3

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 estimated values plotted in Figure 4. In year 0, for taxpayers with the same values of control variables, it is estimated that the audited taxpayer, on average, has a total tax 9 percent more than that of the not-audited taxpayer in the same year (CI 3–17). 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.

Audit Category 4

In Audit Category 4, the audited taxpayers pay significantly less total tax in the baseline year (17 percent of not audited). Both groups have positive trends in tax reporting over time, unlike any of the other audit categories considered. There is evidence to suggest their slopes are not parallel in years 2 and 3: the audited group's total tax reporting increases more starkly between years 2 and 3 than does that of the not-audited group. There is evidence of a weak indirect effect in year 2, after which point the audited taxpayers' total tax estimates plateau.

Audit Category 5

Being our smallest audit category, Audit Category 5 has the largest confidence intervals. There is insufficient evidence to suggest a significant difference in the groups' total tax reporting in the baseline year ($p=0.23$). Both groups have decreasing total tax over time and there is no evidence that a difference in slopes exists over time.

TABLE 1. Exponentiated Coefficients and 95-Percent Confidence Intervals for Effect on Total Tax for Audit, Year, and Interaction Variables

Variable	Audit Category 1		Audit Category 2		Audit Category 3		Audit Category 4		Audit Category 5	
	Estimate (CI)	p-value	Estimate (CI)	p-value	Estimate (CI)	p-value	Estimate (CI)	p-value	Estimate (CI)	p-value
Audited	1.15 (1.13, 1.18)	<0.0001	0.53 (0.51, 0.54)	<0.0001	1.09 (1.03, 1.17)	0.007	0.17 (0.17, 0.18)	<0.0001	1.05 (0.97, 1.13)	0.226
Year after baseline: 1	0.86 (0.85, 0.88)	<0.0001	0.75 (0.74, 0.76)	<0.0001	0.76 (0.72, 0.80)	<0.0001	1.73 (1.69, 1.78)	<0.0001	0.69 (0.64, 0.73)	<0.0001
Year after baseline: 2	0.88 (0.86, 0.90)	<0.0001	0.69 (0.67, 0.70)	<0.0001	0.76 (0.72, 0.81)	<0.0001	2.36 (2.29, 2.43)	<0.0001	0.63 (0.59, 0.67)	<0.0001
Year after baseline: 3	0.90 (0.88, 0.92)	<0.0001	0.65 (0.64, 0.67)	<0.0001	0.79 (0.75, 0.85)	<0.0001	2.79 (2.68, 2.9)	<0.0001	0.59 (0.54, 0.63)	<0.0001
Year after baseline: 4	0.90 (0.88, 0.93)	<0.0001	0.61 (0.59, 0.62)	<0.0001	0.78 (0.73, 0.84)	<0.0001	3.09 (2.95, 3.23)	<0.0001	0.56 (0.51, 0.61)	<0.0001
Year after baseline: 5	0.89 (0.86, 0.91)	<0.0001	0.56 (0.55, 0.58)	<0.0001	0.75 (0.69, 0.81)	<0.0001	3.25 (3.08, 3.44)	<0.0001	0.50 (0.45, 0.55)	<0.0001
Year after baseline: 6	0.86 (0.83, 0.89)	<0.0001	0.54 (0.52, 0.55)	<0.0001	0.73 (0.67, 0.79)	<0.0001	3.42 (3.21, 3.64)	<0.0001	0.49 (0.44, 0.55)	<0.0001
Year after baseline: 7	0.85 (0.82, 0.88)	<0.0001	0.50 (0.49, 0.52)	<0.0001	0.68 (0.61, 0.75)	<0.0001	3.61 (3.35, 3.88)	<0.0001	0.47 (0.42, 0.54)	<0.0001
Year after baseline: 8	0.82 (0.79, 0.86)	<0.0001	0.47 (0.45, 0.49)	<0.0001	0.69 (0.62, 0.78)	<0.0001	3.94 (3.63, 4.28)	<0.0001	0.44 (0.38, 0.51)	<0.0001
Audited*Year after baseline: 1	1.03 (1.0, 1.05)	0.024	1.34 (1.31, 1.37)	<0.0001	1.11 (1.04, 1.18)	0.002	1.5 (1.45, 1.56)	<0.0001	1.10 (1.01, 1.19)	0.031
Audited*Year after baseline: 2	1.36 (1.32, 1.39)	<0.0001	2.10 (2.05, 2.15)	<0.0001	1.22 (1.14, 1.3)	<0.0001	2.25 (2.16, 2.33)	<0.0001	1.10 (1.00, 1.20)	0.042
Audited*Year after baseline: 3	1.47 (1.43, 1.51)	<0.0001	2.14 (2.09, 2.20)	<0.0001	1.13 (1.05, 1.21)	<0.0001	2.28 (2.19, 2.36)	<0.0001	1.10 (1.01, 1.21)	0.038
Audited*Year after baseline: 4	1.42 (1.38, 1.46)	<0.0001	2.10 (2.05, 2.15)	<0.0001	1.15 (1.07, 1.24)	<0.0001	2.36 (2.27, 2.45)	<0.0001	1.09 (0.99, 1.19)	0.072
Audited*Year after baseline: 5	1.37 (1.33, 1.40)	<0.0001	2.08 (2.03, 2.13)	<0.0001	1.21 (1.12, 1.29)	<0.0001	2.45 (2.36, 2.55)	<0.0001	1.11 (1.02, 1.22)	0.023
Audited*Year after baseline: 6	1.33 (1.30, 1.37)	<0.0001	2.01 (1.96, 2.06)	<0.0001	1.2 (1.12, 1.29)	<0.0001	2.46 (2.37, 2.56)	<0.0001	1.08 (0.98, 1.18)	0.13
Audited*Year after baseline: 7	1.27 (1.23, 1.30)	<0.0001	1.90 (1.86, 1.96)	<0.0001	1.26 (1.16, 1.36)	<0.0001	2.47 (2.36, 2.58)	<0.0001	1.04 (0.94, 1.15)	0.496
Audited*Year after baseline: 8	1.18 (1.15, 1.22)	<0.0001	1.82 (1.77, 1.87)	<0.0001	1.2 (1.09, 1.32)	<0.0001	2.3 (2.19, 2.42)	<0.0001	1.05 (0.94, 1.18)	0.354

NOTE: Model is also adjusted for control variables listed in Data and Methods section.

Dollar Estimation

The direct effect for each audit category is calculated as the mean of exam enforcement tax among the audited taxpayers in that category. These data are drawn from the CDW Enforcement Revenue Information System (ERIS) database, and represent audits conducted for Audit Categories 1-5 for the baseline TYs 2006–2012. This is listed in Table 5, along with the estimated indirect effects. All direct effects are adjusted for inflation to 2018 USD. Audit Category 5 has the largest direct effect, with an average of $4,482.39. Audit Categories 1 and 2 have similar direct effects of $2,668.29 and $2,644.39, respectively. Audit Categories 3 and 4 have on average the smallest exam enforcement taxes.

After transforming our interaction estimates to the dollar scale by combining the estimated effect (as a percent) with the estimated total tax paid by the audited taxpayers at each time point, we find the indirect effect as a sum of those over years 1 through 5. The upper bound of 5 years was chosen due to prior research noting an attenuation in indirect effect after 5 years, as well as the fact that not all taxpayers have more than 6 years observed (i.e., those with baseline year TY 2012).

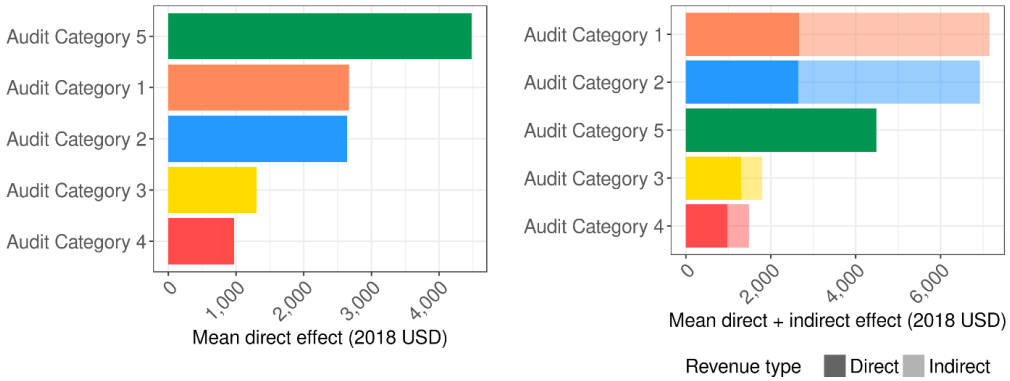
While having the largest direct effect, Audit Category 5 had insufficient evidence of any significant audit*years terms in the mixed effects model, and thus we estimate 0 indirect effects. As stated previously, this is the smallest audit category considered, with only 11,891 taxpayers total. We lacked sufficient statistical power to be able to detect an indirect effect. That brings Category 5's total revenue to $4,482.39, the same as its direct effect, and ranks it third in the right plot of Figure 5. Categories 1 and 2 again had similar indirect effects, which can be seen in Figure 4 around years 2-3. The estimated indirect effects were nearly twice that of their direct effects. Combining direct and indirect, we estimate Audit Category 1 to have the highest total revenue of all categories considered with $7,147.83. Category 2 is a close second at $6,926.97. Categories 3 and 4 had estimated indirect effects roughly half that of their direct effect, as seen in Table 2 and Figure 5. Audit Category 4, which deals with education credits, has the smallest total revenue of all categories considered.

TABLE 2. Direct and Indirect Effects by Audit Category (2018 USD)

Audit Category	Direct Effect	Indirect Effect	Total
1	2,668	4,480	7,148
2	2,644	4,283	6,927
3	1,302	502	1,803
4	972	509	1,481
5	4,482	0	4,482

NOTE: The by direct effect is the average exam enforcement tax after adjusting for inflation to 2018 USD.

FIGURE 5. Ranking of Audit Categories by Direct Effect Only (left) and by Sum of Direct and Indirect Effects (right)



Discussion

In this study, we investigated the indirect effect of experiencing an audit on subsequent total tax reporting and translated our results into absolute dollars back to the IRS for five categories of correspondence audit. We advance prior literature on the indirect effects of audits by explicitly considering how these estimates could be translated into data that support operational decision-making for resource allocation. More specifically, we push the existing literature forward in three ways: 1) by accounting for operational selection criteria; 2) by generating estimates of the indirect effect for multiple types of correspondence audits that decision-makers can compare when planning their workload; and 3) by comparing the estimated indirect effect with the direct effect of each audit category to better understand the overall "return on investment" for these audit categories. 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 *et al.* (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 the audit group based on criteria that are only partially known (Slemrod (2016)), even from within a narrowly defined candidate population.

For four of the five audit categories, we find evidence of an indirect effect among audited taxpayers; further, we see evidence that this effect varies in magnitude depending on the audit category, and therefore varies by taxpayer population. In other words, not all correspondence audits are created equal in terms of their specific 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 the baseline year, followed by an attenuation out to year 8. Considering that most Audit Category 1 exams will have started 3 years after the baseline year, we assume that most of the audited taxpayers have been notified by the peak in reporting observed in Figure 4 at 3 years after the baseline year. In this way, our results mirror previous findings from both research audit data on Schedule C filers (DeBacker *et al.* (2018a)) and findings using operational data on Schedule C filers (Beer *et al.* (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 (measured by the reporting of total tax) for Audit Category 3 (self-employment tax). To our knowledge, Schedule A and Schedule SE 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.

For Audit Category 4, we see a weak indirect effect around year 2. Interestingly, this is the only category considered in which both groups have increasing total tax over time. However, given that the context of this audit is to examine credits taken for education expenses, we posit that this is completely logical: taxpayers audited for Category 4 are likely to begin earning higher income each year after college, and therefore report higher tax each year. An indirect effect may be less applicable here, because by the time most taxpayers are notified of the audit of their education credits, their education is likely completed, and they are no longer claiming the same credits. Therefore, we hypothesize that an examination of credits one knows they will probably never claim again will have a relatively small deterrent effect on subsequent tax reporting.

In Audit Category 5, both the audited and not-audited taxpayers have decreasing total tax reporting after the baseline year. This category deals with passive activities, in which the taxpayer is not an active participant in the investment or business trade in question. A large loss claimed corresponds to less net income and therefore less tax.

Our efforts translating the multiplicative estimates from model 1 to absolute dollars highlight the utility of indirect effects for correspondence audit resource allocation. When considering only direct exam revenue, Audit Category 5 leads the pack. However, this audit category is also the smallest and has insufficient evidence to suggest any additional indirect effect, rendering it third in terms of total effect (direct and indirect) in the right panel of Figure 5. Categories 1 and 2 both have the largest volume of taxpayers over our 7-year period, and they also both have relatively large estimated indirect effects. For Category 1, the estimated indirect effect is nearly 1.7 times that of the direct effect. For Category 2, this number is 1.6. This implies that the subsequent deterrence from a single correspondence audit could be substantively larger than the average exam adjustment. The results of our dollar estimation also underscore the notion that not all correspondence audits are

created equal: Audit Categories 3 and 4 have indirect effects approximately half the size of their direct effects. In summary, allocating resources based purely on the direct effect could mislead calculations. Examining the total revenue, the sum of average direct and average indirect effects, paints a more representative picture of the value of an audit over a given time span.

Not all audits result in a direct effect at all: some audits end up being a "no-change" audit, in which the taxpayer is found to not owe any additional taxes. Prior work has shown that taxpayers whose audits resulted in a nonzero adjustment had more substantive increases in total tax and specific line-item reporting after audit (Beer *et al.* (2015)). Figure 6 in Appendix 2 shows total tax model estimations in which the audited group is broken out into two outcome groups: change (assessed revenue amount > $0) and no-change (assessed revenue amount = $0). For Audit Categories 1-4, it is evident the change group has a larger increase in total tax reporting than the no-change group. This means that the notification of a change on one's return likely has a larger deterrent effect than the notification of an examination that finds one in compliance.

Although exciting from a research perspective, the change vs. no-change dichotomy does not translate easily to operational use. That is, having an estimated indirect effect attributable to the change group, as well as an estimated effect attributable to the no-change group, and using that to allocate budget or other resources toward certain types of audits versus others, requires a priori knowledge whether an audit will have a change or no-change outcome. Deriving a model to classify taxpayers as change or no-change *before* they are audited is out of scope for the current work. Therefore, we choose to show these estimates as a supplementary analysis in Appendix 2 to highlight that the specific effect does indeed seem to be contingent on the audit outcome. However, we focus on our two-group models for resource allocation purposes.

While it is evident from Figure 4 that the audited and not-audited groups do not always have the same tax reporting at baseline, it is important to note that all models are controlling for the audit prioritization variable. This allows us to account for a degree of operational selection bias in the "treatment" condition of being audited. For two taxpayers with the same priority at baseline *and* with the same total tax reporting the year before baseline, we expect the relationships seen in Figure 4. Beyond the known operational filters that we already apply to define our control groups for each audit category, it is possible the IRS could have applied further exclusion criteria of which we are unaware. However, our research team has worked in direct collaboration with the IRS operations group responsible for these audits in order to measure and implement these operational criteria to the best of our abilities. We assert that 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. Therefore, we successfully account for a substantive level of selection bias and the modeling results apply to relatively homogenous taxpayers who are similarly likely to be audited.

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 analyzed here have been relatively stable over time, especially regarding 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 selection criteria (e.g., those used for TY 2018), but these filters may not necessarily apply to TYs 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 TYs 2006–2012.

The Tax Cuts and Jobs Act of 2017 raised the standard deduction amount, which will affect the available pool of taxpayers who itemize and file a Schedule A. This reduction in eligible taxpayers for Audit Category 2 may reduce the generalizability of our results for this model.

Further, there appears to be some overlap between the distributions of the priority variable for the audited and not-audited groups for all 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. The bias due to unobserved confounders poses an additional limitation to our work, and one that we hope to overcome in future research.

Additionally, audited taxpayers have varying notification times, even for audits of returns from the same tax year. Therefore, 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).

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. 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." *2015 Annual Report to Congress.* Washington, DC: Taxpayer Advocate Service, 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.

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, 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." National Bureau of Economic Research Working Paper 24465. https://www.nber.org/papers/w24465.pdf.

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.

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." University of Milan Bicocca Department of Economics, Management and Statistics Working Paper No. 359. 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://www.taxpayeradvocate.irs.gov/wp-content/uploads/2020/08/2014-ARC_VOL-2_2-Impact-of-Audits-508.pdf.

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

Pinheiro, José. 2020. Package "nlme: Linear and Nonlinear Mixed Effects Models." Version 3.1-151. https://cran.r-project.org/web/packages/nlme/nlme.pdf.

Plumley, Alan. 1996. *The Determinants of Individual Income Tax Compliance.* Publication 1916 (Rev. 11-96). Washington, DC: Internal Revenue Service. https://www.irs.gov/pub/irs-soi/pub1916b.pdf.

Pomeranz, Dina. 2015. "No Taxation Without Information: Deterrence and Self-Enforcement in the Value Added Tax." *The American Economic Review* 8: 2539.

Rettig, Charles P. 2016. "IRS Audit Selection and Classification Processes." March 20, 2016. *Journal of Tax Practice & Procedure,* February-March 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2771191.

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

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): 455–483. https://doi.org/10.1016/S0047-2727(99)00107-3.

United States Government Accountability Office (GAO). 2012. *Tax Gap: IRS Could Significantly Increase Revenues by Better Targeting Enforcement Resources.* GAO Report to Congressional Requesters (GAO-13-151). https://www.gao.gov/assets/660/650521.pdf.

Appendix 1. Estimating the Dollar Value of the Specific Indirect Effect

In this section, we go into greater detail on the method used to translate model predictions into dollar value estimates of the indirect effect of audit. Our objective is to produce values that can be used as a proxy for the indirect revenue of audits by estimating differences in changes to total tax between audited and not-audited taxpayers over time.

In Model (1), described in the text of this paper, the effect of an audit on subsequent yearly total tax liability is represented by the interaction terms $\beta_{11\text{-}18}$ *audited*$_i$**year after baseline*$_{ij}$ for an individual *i* at year *j* after the baseline. That is, we can expect that, on average, audited taxpayers will have increased their individual total tax reporting in year *j* over what they otherwise would have reported (if not audited) by a factor of $e^{\hat{\beta}_{10+j}}$ as a result of the audit.

Scaling Effects of Audit on Subsequent Yearly Total Tax to Audit Category

For a given audited taxpayer, we define the estimated difference in total tax reporting attributed by the model to lingering effects of the audit at years from baseline to be

$$\hat{\delta}_i = \text{total tax}_{audited,j} - \widehat{\text{total tax}}_{not\ audited,j} \quad (2)$$

Where $\widehat{\text{total tax}}_{not\ audited,j}$ is given such that

$$\ln(\text{total tax} + 1)_{j,audited} - \hat{\beta}_{10+j} = \ln(\widehat{\text{total tax}} + 1)_{not\ audited,j} \quad (3)$$

And $\hat{\beta}_{10+j}$ is a coefficient estimate corresponding to the audit-time interaction at year obtained in Model (1).

Under this definition, we denote the estimated dollar-valued effect of an audit on an audited taxpayer's total tax reporting at time *j* by

$$\hat{\delta}_j = \left(1 - \frac{1}{e^{\hat{\beta}_{10+j}}}\right) * (\text{total tax} + 1)_{audited,j} \quad (4)$$

Where individual *i* was audited at time point *j=0* . For the purposes of resource allocation among audit categories at the aggregate level, we use an estimated quantity of total tax observed under each category of audit at time *j*. To do this, we generate estimates of (total tax+1)$_{j,audited}$ for each audit category and time point *j* using data corresponding to individuals audited under that category. These estimates reflect the total tax reporting that might be expected of a hypothetically *average* taxpayer of the audited group, providing a basic way to compare the scales on which revenue from each category of audit exists.

Estimates for *(total tax+1)*$_{j,audited}$

To generate estimates of *(total tax+1)*$_{j,audited}$ for each audit category and time point, we use output from the model under inputs derived to reflect a hypothetically average audited taxpayer. That is, for each independent variable, we obtain the average of observed values among returns years after an audit under the relevant audit category. In cases of categorical independent variables, such as filing status, this means that each input given to the model reflects a proportion of observations known to match the relevant factor level, as opposed to a realistic binary value. Naturally, audit status and the number of years after baseline are preset to match an audited taxpayer observed at time *j*. The quantity of total tax having been produced by the model under such inputs is then used as our estimate for total tax at the relevant time point, denoted $\widehat{\text{total tax}}_{not\ audited,j}$.

In using this approach to estimate generation, we benefit from the interpretation of how the specific indirect effect may impact a theoretically average taxpayer of the audited group and category, as opposed to how it might impact taxpayers with an average reported total tax value among audit group and category. In addition, this method allows us to avoid many of the detriments associated with reliance on heavily skewed one-dimensional data, and it produces conservative estimates of total tax that maintain the relative differences in scale between each audit category.

Generating a Single Estimate of the Specific Indirect Effect per Audit Category

Once obtained, values of $\hat{\delta}_{ij}$ are summed over multiple years to obtain a more complete point estimate of the specific indirect effect on total tax reported over a period of interest. For the purpose of quantifying this amount for application in resource allocation, we consider such estimates only through year *j=5*. This aggregated estimate of $\hat{\Delta}$ specific indirect effect on tax reporting imposed by an audit category over 5 years is given by

$$\hat{\Delta} = \sum_{j=1}^{5} \hat{\delta}_j = \sum_{j=1}^{5} \left(1 - \frac{1}{e^{\hat{\beta}_{10+j}}}\right) * (\widehat{\text{total tax}} + 1)_{audited,j} \quad (5)$$

$\hat{\Delta}$ is then compared to an estimate of the average direct effect imposed by an audit category. Direct effect estimates are obtained using an average of direct audit revenue as recorded in the CDW ERIS database as enforcement tax, adjusted to 2018 USD.

Appendix 2. Estimates for Three-Group Models of Total Tax

In this section, we present a supplementary, three-group analysis of the indirect effect where we disaggregate the audited group into two groups based on their audit outcomes: the "change" group, and the "no-change" group. Some prior literature describing the indirect effect of audits on self-employed taxpayers suggests that the outcome of an audit is a key factor associated with taxpayer reporting trajectories over time after the audit (e.g., Beer *et al.* (2015)). As mentioned in the body of this paper, we do not use these three-group models to generate dollar estimates of the indirect effect for resource allocation purposes because current operational practices could not make use of dollar value estimates at the three-group level. Using these three-group estimates in order to allocate resources to different audit categories would require a way to know *a priori* whether a given audit could be expected to result in a change (adjustment) or a no-change (no adjustment). Obviously, if this were feasible operationally, the IRS would not select *any* no-change returns at all. This is currently outside the scope of IRS operations.

Audit ("Treatment") Group

To define the audited group, all primary taxpayer identification numbers associated with one of the five 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 TY 2018). For example, for baseline year 2006, we compiled return data up through TY 2014; for baseline year 2012, we compiled return data up to TY 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.

To define the audited/change and audited/no-change groups, we disaggregated the audited taxpayers into two groups based on whether their audit was recorded as resulting in tax revenue being assessed. The audited/change group includes taxpayers who were recorded as having assessed revenue greater than zero; the audited/no-change group includes taxpayers who were recorded as having assessed revenue equal to zero.

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 non-audited taxpayers available in CDW. We restricted the returned records to a random sample of up to 25,000 taxpayers from the eligible population in each of TYs 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 TY 2018). Note that this control group for the three-level models is identical to the control groups for the two-level models.

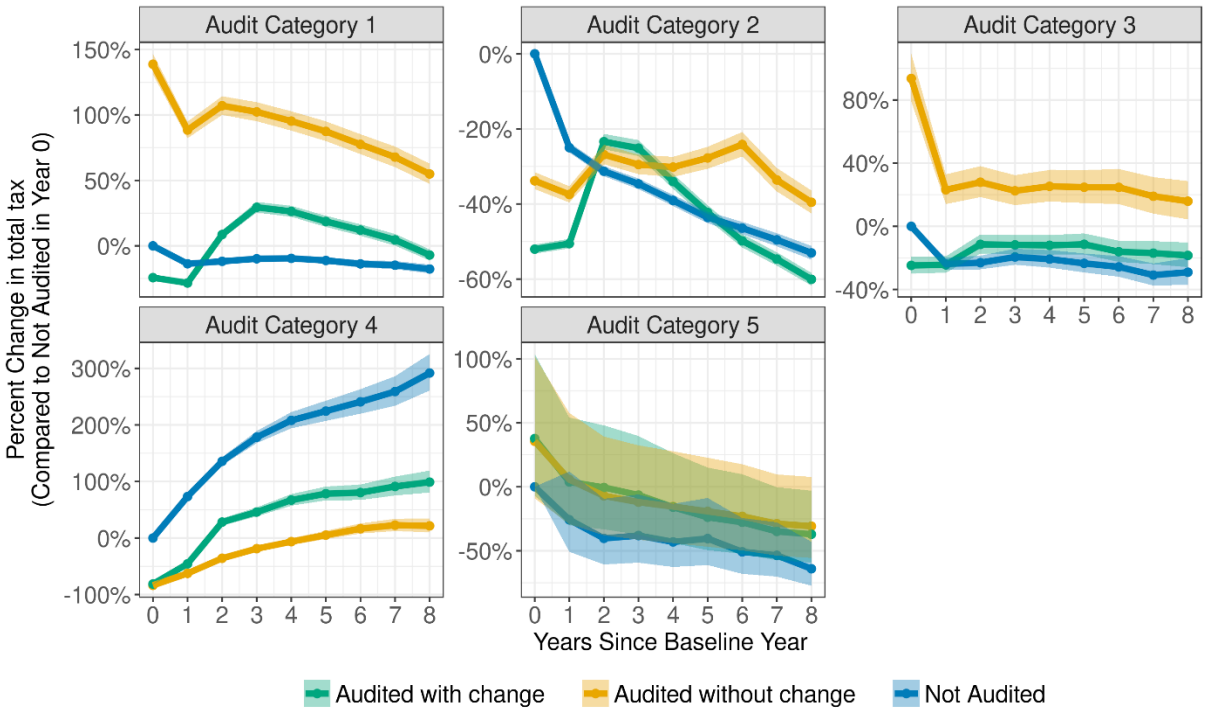
Statistical Analysis

Model (6) disaggregates the audited group into audited/change and audited/no-change groups to detect whether the effect of audit over time varies depending on the audit outcome. As such, it is a time-invariant measure of whether the taxpayer was audited with a change outcome (reference group is not audited), and it is a time-invariant measure of whether the taxpayer was audited with a no-change outcome (reference group is not audited).

$$
\begin{aligned}
(6)\quad & \ln(\text{total tax}+1)_{ij} \\
&= \beta_0 + \gamma_{0i} + \beta_1 FS_{ij} + \beta_2 TY_{ij} + \beta_3 TPI_{ij} + \beta_4 priority_i + \beta_5 any\ wages_{ij} \\
&+ \beta_6 total\ exemptions_{ij} + \beta_7 any\ Child\ Tax\ Credit_{ij} \\
&+ \beta_8 itemized\ deductions_{ij} + \beta_9 mortgage\ interest_{ij} + \beta_{10} urban_{ij} \\
&+ \beta_{11} audit\ last\ 10\ TYs_i \\
&+ \beta_{12} Total\ Tax_{i,-1} + \beta_{13} preparer_{ij} + \beta_{14} auditedChange_i \\
&+ \beta_{15} auditedNoChange_i + \beta_{16} year\ after\ baseline_{ij} + \beta_{17} auditedChange_i \\
&* year.after.baseline_{ij} + \beta_{18} auditedNoChange * year\ after\ baseline_{ij} \\
&+ \epsilon_{ij}
\end{aligned}
$$

Results

FIGURE 6. Estimated Changes in Total Tax by Audit Category and Audit Status



In Audit Category 1, the audited without change group has a significant increase in estimated total tax at year 2 before its slope decreases out to year 8. This group also has a significantly higher total tax at baseline, which could explain why their audit did not result in a change: these taxpayers were already remitted much higher tax than similar eligible taxpayers. The change group has a slightly different trajectory: their maximum tax reporting is around year 3. This group also has the lowest reported total tax at baseline. The magnitude of the increase in years 2–3 is larger for the change group than the no-change group, suggesting that an adjustment on one's return could have a larger deterrent effect than an exam without an adjustment.

A similar phenomenon is evident for Audit Category 2, however in this population the not-audited group has the highest total tax at baseline. The no-change group still has an uptick in reporting at year 2, but it is smaller in magnitude than that of the change group. Category 3 also has a discernable peak in tax reporting for the change group at year 2. In the two-group model, Audit Category 4 had weak evidence of an indirect effect. In this three-group model, we see that the change group does indeed have a significant uptick at year 2 while

the no-change group has no evidence of an indirect effect. This implies that for this population of education credit-claiming taxpayers, an audit without an ultimate adjustment on the return has little to no deterrence. Lastly, for Audit Category 5, there is insufficient evidence to suggest a difference in reporting in the three groups.

Considerations for Future Research

One limitation of our current modeling for the three-group models presented here is the suitability of the single control group in the three-group paradigm. That is, we assume homogeneity in potential audit outcomes among the entire not-audited group; the reality is that if those unaudited taxpayers *were* audited, some of those audits would result in changes (adjustments) and some would not. Theoretically, a large proportion of the not-audited taxpayers should be "no-change" audits, if they were indeed audited. One avenue for future research would be to predict whether a return would be change or no-change prior to estimating the indirect effects. Because the change and no-change returns are fundamentally different, this would lend itself better to a true estimation of the counterfactual reporting of audited taxpayers. While out of scope for our current study, such an approach would theoretically be useful for translating the change/no-change estimates to dollar values for resource allocation purposes. However, since the processes for allocating audit resources and selecting returns for audit cannot distinguish well between change- and no-change returns *before* auditing them, making this distinction in indirect effects estimation may have little operational application. Nonetheless, the results presented in this section serve to suggest an initial association between audit outcome and the specific indirect effect.

Audits, Audit Effectiveness, and Postaudit Tax Compliance[1]

James Alm (Tulane University) and Matthias Kasper (Tulane University and University of Vienna)

1. Introduction

Tax audits are an essential instrument in establishing and maintaining compliance, and increasing the number of audits has direct and indirect effects on taxpayer behavior. Audits have *direct* effects by raising revenue through the assessment of additional taxes, interest, and penalties on individuals who are audited. Additionally, tax audits have *indirect* effects by deterring future noncompliance among both audited taxpayers (specific deterrence) and unaudited taxpayers (general deterrence). A growing body of research analyzes these direct and indirect deterrent effects of tax audits and generally shows that more audits lead to more compliance (Alm (2019); Slemrod (2019)).

However, an important if largely unexamined feature of tax audits is that they do not always detect tax evasion when it is present, and they may even find evasion when it is not in fact present. For example, early work by Feinstein (1991) suggests that the average detection rates of senior tax examiners are around 50 percent. This affects the revenue collected from audits. In Fiscal Year 2018, U.S. taxpayers challenged over $10 billion in additional taxes recommended by the Internal Revenue Service (IRS), while almost $4 billion of tax and penalties were under appeal in U.S. tax courts (Internal Revenue Service (2019)). In addition to these direct effects, audit "effectiveness," or the tax administration's capacity to detect noncompliance in an audit, might affect a taxpayer's behavioral response to enforcement. For example, recent work suggests that the specific deterrent effect of audits depends on the audit outcome. These studies find that tax audits increase subsequent, or postaudit, compliance among taxpayers who were found to be noncompliant, while they decrease compliance among those who were determined to be compliant (Gemmell and Ratto (2012); Beer *et al.* (2020)). This raises questions about the effect of audit effectiveness on postaudit tax compliance. More specifically, it remains unclear whether inefficient audits undermine the specific deterrence effect of enforcement. A related question is whether truly compliant and truly noncompliant taxpayers differ in their behavioral response to enforcement.

This study addresses these questions and investigates the specific deterrent effect of audits on postaudit tax compliance. We run a preregistered laboratory experiment with 333 participants in which we test how variation in the risk of detection affects subsequent tax compliance. An important feature of our experimental design is the addition of audit "effectiveness" to our audit mechanism, where effectiveness is defined as the share of undeclared income that the tax agency detects in an audit (Rablen (2014)). This addition allows us to examine the effects of audit effectiveness on postaudit compliance. We also study whether enforcement has differential effects on different types of taxpayers, as distinguished by their prior reporting behavior. Addressing these questions with field data is difficult, even problematic, because tax agencies typically do not know a taxpayer's true tax liability. In particular, the audit outcome is not a perfect measure of a taxpayer's true compliance, so that the identification of audit effectiveness and its effects on truly compliant and noncompliant taxpayers is challenging. In contrast to the use of field data, data generated from a laboratory experiment allows us to introduce changes in both audit probability and audit effectiveness, and thereby allows clean identification of the effects of these changes on postaudit compliance of truly compliant and noncompliant individuals.

[1] This study was preregistered under: https://osf.io/uhpmw/. It has been approved by the Institutional Review Board of Tulane University (2019–1077) and the Institutional Review Board of the Institute of Work, Economy, and Social Psychology at the University of Vienna (2019/W/001). We thank the Vienna Centre of Experimental Economics (VCEE), University of Vienna, for allowing us to run our experiments in their laboratory. We also thank Steven Sheffrin and the Murphy Institute for the generous support that made this study possible. We appreciate valuable comments from Sebastian Beer, Linda Dezsoe, Brian Erard, Janet Holzblatt, Christoph Kogler, Luigi Mittone, Stephan Muehlbacher, Jerome Olsen, Alan Plumley, Ziga Puklavec, Alexander Siebert, and from participants at the Tulane University/Murphy Institute Conference on "Economic and Behavioral Dimensions of Tax Compliance" held in New Orleans, Louisiana, in March 2019 and the 9th Annual IRS/TPC Joint Research Conference on Tax Administration.

Our study differs from the previous literature by making contributions in two important dimensions. First, unlike most existing work, we account for the possibility that tax audits might not detect all undeclared income. This allows us to investigate whether ineffective audits reduce taxpayers' propensity to comply in the future, and it also allows us investigate whether presenting the compliance decision as a two-stage compound lottery (where an audit does not result in certain detection) changes a taxpayer's willingness to comply compared to a single-stage lottery (where an audit results in certain detection) (Kahneman and Tversky (1979); Bernasconi and Bernhofer (2020)). Second, we investigate whether behavioral responses to enforcement depend on taxpayers' prior reporting behavior. More specifically, we distinguish between "compliant," "partly compliant," and "noncompliant" individuals, where compliant taxpayers are defined as those who report all income *in the round that is audited* and noncompliant taxpayers report zero income in this round. Similarly, we distinguish between "honest" and "dishonest" individuals, where honest taxpayers report all income *in all rounds prior to their first audit* and dishonest taxpayers report zero income in these rounds. This latter distinction allows us to identify the effect of enforcement on taxpayers who do not respond to changes in the incentives to evade prior to experiencing their first audit. In sum, our design allows us to disentangle the possible mechanisms that drive postaudit tax compliance, and we are also able to investigate in detail the effect of audits on different types of taxpayers (Torgler (2003)).

Our results indicate that increasing the probability of detection (the product of the audit probability and the audit effectiveness) results in higher compliance levels, but we find no evidence for a misperception of compound detection lotteries. Moreover, we find that audit effectiveness is an important determinant of the specific deterrent effect of audits. Taxpayers declare a larger share of their income after experiencing an audit that detects all undeclared income while ineffective audits decrease postaudit compliance. Moreover, we find that prior reporting compliance affects these behavioral responses to audits. While individuals who have been found to underreport their entire income (noncompliant taxpayers) declare substantially more income in subsequent rounds, postaudit compliance declines considerably among those who have been found to report all income correctly (compliant taxpayers). We also find that audits increase compliance among dishonest individuals who never declared any income before experiencing their first audit. However, we find no evidence that audits "crowd out" compliance among honest taxpayers who reported all income correctly in all rounds prior to their first audit.

Our study adds to the literature on behavioral responses to enforcement. Moreover, we provide a new perspective on the tradeoff between audit frequency and audit effectiveness (Rablen (2014)) and the analysis of optimal tax administration (Keen and Slemrod (2017)). Our results suggest that a complete analysis of a revenue-maximizing audit strategy requires the consideration of postaudit behavior and in particular behavioral responses to audit effectiveness as well as differential responses of compliant and noncompliant taxpayers.

2. Related Literature

Prior work on the specific deterrent effect of tax audits typically has used administrative data to analyze the aggregate response of those taxpayers who have been audited. Overall, these studies find that enforcement has a positive effect on subsequent reporting compliance.[2] For example, Kleven *et al.* (2011) show that tax audits increase self-reported income among Danish taxpayers in the subsequent tax year. Similarly, Advani *et al.* (2017) find that reported income of self-employed UK taxpayers increases for at least 5 years after an audit, while DeBacker *et al.* (2018) show that compliance of U.S. taxpayers improves for 3 years after an audit before ultimately reverting to previous (and lower) levels. A more recent study of U.S. taxpayers by Beer *et al.* (2020) investigates whether the effect of audits on postaudit reporting behavior depends on the audit outcome, and they find that the specific deterrent effect of tax audits is positive in the aggregate but that subsequent compliance depends on the outcome of the examination. In particular, taxpayers who receive an additional tax assessment as a result of their audit report more income in subsequent years, while those who do not receive an additional assessment report less. This result is in line with a study by Gemmell and Ratto (2012) for the UK that finds that audited taxpayers who were found to be noncompliant report more income in their subsequent

[2] An exception is Erard (1992), who analyzes microlevel data from the U.S. Taxpayer Compliance Measurement Program (TCMP) of the IRS and who finds no significant effect of a prior tax audit on subsequent compliance.

tax return than those who were not audited, while taxpayers who were found to be compliant show the opposite response. A study on the effects of audits on VAT compliance in Argentina and Chile also finds that audits have a differential effect on postaudit compliance, however this study finds that audits decrease compliance among those who were found to be cheating (Bergman and Nevarez (2006)).

Taken together, these studies suggest that tax audits increase subsequent reporting compliance in the aggregate. However, they raise the question why enforcement appears sometimes to encourage rather than deter future noncompliance.[3]

Several behavioral explanations have been suggested for these results (Kirchler (2007); Alm (2019); Beer *et al.* (2020)), but the underlying mechanisms remain unclear. One possible explanation is that ineffective audits might stimulate a taxpayer's willingness to take risks; that is, if an audit fails to detect undeclared income, the taxpayer might infer that the agency is unable to discover cheating and thus underreport his income in subsequent years (Andreoni *et al.* (1998)). Indeed, prior work finds that unsanctioned criminal offenses reduce perceived risk of detection and punishment (Matsueda *et al.* (2006)). However, almost all prior work that estimates behavioral responses to tax enforcement assumes that tax audits always detect all undeclared income. The few exceptions employ laboratory experiments to investigate how variation in audit effectiveness affects the general population of taxpayers, rather than those taxpayers who experienced the audit. For example, Alm and McKee (2006) vary the fraction of undeclared income that the tax agency detects in an audit, and, surprisingly, they find higher compliance levels when audit effectiveness is low. Similarly, Bernasconi and Bernhofer (2020) find some support for the hypothesis that ineffective tax audits increase compliance in the aggregate. However, while these two studies suggest that the general deterrent effect of ineffective tax audits might be positive, and potentially even greater than the effect of effective tax audits, the effects of ineffective audits on postaudit tax compliance remain unknown.

A second explanation for the unintended consequences of tax audits is the "bomb crater effect" (Guala and Mittone (2005); Mittone (2006)). Contrary to the standard model of tax evasion (Allingham and Sandmo (1972)), it is common in laboratory experiments to find that participants declare a smaller share of their income after being audited. Such a response might result from an underestimation of the risk of future audits (Mittone *et al.* (2017)) or from loss-repair motivations (Maciejovsky *et al.* (2007)). However, it remains unclear whether the perceived risk of future examinations is affected by the audit outcome or whether the tendency to make up for past losses pertains to individuals who have been found to be noncompliant. For example, some studies find that a decline in reported income after an audit cannot be explained by loss-repair motivations alone because individuals who were found to be compliant also report less income after experiencing an audit (Kastlunger *et al.* (2009); McKee *et al.* (2018); Bernasconi and Bernhofer (2020)).

A third potential explanation is that audits have differential effects on different types of taxpayers. Some scholars have suggested that taxpayers comply for different reasons (e.g., Erard and Feinstein (1994); Torgler (2003); Braithwaite (2009)). While some taxpayers are motivated entirely by the expected value of the evasion gamble, others comply regardless of any incentive to cheat (Braithwaite (2003)). However, such honest taxpayers may find being audited unfair, perceive the audit as a breach of trust, or experience negative emotions (Olsen *et al.* (2018); Enachescu *et al.* (2019)). This experience might crowd out their intrinsic motivation to comply and reduce their propensity to comply in the future (Frey (1997); Mendoza *et al.* (2017); Lederman (2018); Hu and Ben-Ner (2020)). Therefore, a decline in postaudit compliance might also result from honest individuals who are less likely to comply after experiencing an audit.

Taken together, prior studies suggest different behavioral explanations of responses to tax audits, but without resolving the actual mechanisms that drive these responses. Our study allows us to discern the potential explanations that have been proposed in the literature. To our knowledge, our study is also the first to investigate the effect of audit effectiveness on postaudit compliance.

[3] A body of research in criminology investigates the effect of punishment on an individual's future proclivity for crime. Reviews of this literature suggest mixed evidence for specific deterrence effects, but there is some indication that the experience of punishment might increase, rather than decrease future offending (Cullen *et al.* (2011), Nagin *et al.* (2009), Nagin (2013a); Nagin (2013b)).

3. Theoretical Foundations

Theories of deterrence distinguish between threat of punishment and experience of punishment (Chalfin and McCrary (2007)). The literature in economics focuses almost exclusively on the prior. A taxpayer's compliance decision is typically analyzed within an expected utility framework that follows Becker's (1968) economics-of-crime approach. The standard model of tax evasion (Allingham and Sandmo (1972); Srinivasan (1973); Yitzhaki (1974)) describes a taxpayer's reporting decision as a decision under risk, where all relevant parameters are fixed and known with certainty. The model assumes that a taxpayer receives income I and must decide how much to report to the tax agency. Reported income R is taxed at the rate t, and unreported income is not taxed. The taxpayer faces the risk of being audited with a probability p. In case of an audit, the agency is assumed to detect all undeclared income and to impose a fine f on the undeclared amount;[4] in case of no audit, the taxpayer simply pays taxes on reported income. The taxpayer chooses R to maximize the expected utility of the evasion gamble, or:

$$\text{EU(I)} = (1 - p)\ \text{U}(I - tR) + g\,p\ \text{U}(I - tR - tf(I - R)), \tag{1}$$

where utility U( ) depends only upon income and E is the expectation operator. The model predicts that an increase in the audit probability p or the penalty rate f translates into higher compliance levels.[5]

One major problem with the standard expected utility approach to tax compliance is that the observed levels of tax evasion are not as high as the theory predicts. Taxpayers typically face a low risk of being audited and modest fines for noncompliance. Assuming reasonable risk preferences, a taxpayer that is motivated by financial incentives alone should evade more than the evidence suggests (Skinner and Slemrod (1985)). One explanation for this "tax compliance puzzle" is that taxpayers overestimate the risk of an audit (Alm *et al.* (1992)). An alternative explanation is that a taxpayer's compliance decision is not determined by financial considerations alone.[6] For example, Erard and Feinstein (1994) point out that some taxpayers are inherently honest and report all income correctly even when they face financial incentives to underreport their income.

In light of these findings, several authors have suggested to apply rank dependent expected utility theories to tax compliance (Bernasconi (1998); Yaniv (1999); Bernasconi and Zanardi (2004); Alm and McKee (2006); Dhami and al-Nowaihi (2007); Hashimzade *et al.* (2013)). These models allow individuals to overweigh the probability of an audit and to exhibit more extreme forms of risk aversion. As a result, they generate predicted levels of compliance that better approximate observed levels. With rank dependent expected utility, the basic maximization problem of equation (1) now becomes

$$\text{EU(I)} = (1 - g\,p)\ \text{U}(I - tR) + g\,p\ \text{U}(I - tR - tf(I - R)), \tag{2}$$

where g serves to overweight the probability of detection and punishment.[7]

These models typically assume that an audit detects all undeclared income, but they can be easily adjusted to allow for ineffective audits. In this case both the audit and the outcome of the audit are uncertain, rendering the evasion gamble a two-stage (compound) decision. In a variation of the expected utility model (1), a taxpayer now faces a compliance choice given by:

$$\text{EU(I)} = (1 - p)\ \text{U}(I - tR) + p(e\ \text{U}(I - tR - tf(I - R)) + (1 - e)\ \text{U}(I - tR)), \tag{3}$$

where e is the probability that the audit is effective and detects all undeclared income.[8]

If taxpayers compute compound lotteries correctly, the compliance effect of a change in the audit probability is the same as the effect of an equivalent change in audit effectiveness.[9] However, presenting a decision as a two-stage compound lottery, rather than a single-stage lottery with identical expected outcomes, might

[4] In Yitzhaki, (1974) the fine is imposed on unpaid taxes.

[5] There is ample evidence that increasing p and f increases compliance. Alm (2019) and Slemrod (2019) provide comprehensive surveys of the literature.

[6] Kirchler (2007) provides an overview of nonfinancial determinants of tax compliance.

[7] This alternative approach also helps illuminate the roles of information dissemination by the tax authority. Any information provided by the tax authority that describes audits and their ability to detect undeclared income should increase the weighted probability of an audit, while information that suggests the ineffectiveness of audits should lower the weighted probability of an audit. It is straightforward to derive comparative statics results from this approach.

[8] It can directly be seen that (3) collapses to (1) if e =1.

[9] Specifically, simplifying (3) yields that an x percentage point increase in p is offset by an $1/x$ percentage point increase in e and vice-versa.

induce a shift in preferences. Assuming a nonlinear probability weighing function, where small probabilities are overestimated and large probabilities are underestimated, decision-makers who evaluate the two stages in isolation exhibit different risk preferences than those who consider the compound lottery (Kahneman and Tversky (1979)). Whether a taxpayer misperceives the risk of detection when the audit probability is distinct from the audit effectiveness depends on the magnitude of these parameters, the shape of the taxpayer's probability weighing function, and the taxpayer's cognitive capacity (Dillenberger (2010); Harrison *et al.* (2015); Prokosheva (2016)). For example, Bernasconi and Bernhofer (2020) find that taxpayers' probability weighing functions adjust over time due to learning effects.

It is important to note that all these models predict that audits do not affect a taxpayer's subsequent reporting decision, because they assume that the audit does not provide the taxpayer with new information. As audit and penalty rates are fixed and known, experiencing an audit is merely a case of losing the evasion gamble, which should not affect postaudit compliance. However, in most cases a taxpayer does in fact not know how likely his noncompliance will be detected and a tax audit might provide new information to the taxpayer that affects his postaudit compliance (Snow and Warren (2007); Kleven *et al.* (2011)). For example, if the audit detects more noncompliance than expected, the taxpayer may increase his prior on the probability of detection and increase his postaudit compliance. Conversely, a taxpayer may decrease his prior on the probability of detection, and thus decrease his postaudit compliance, if the audit detects less noncompliance than expected (Slemrod (2019)). The tax audit would have a specific deterrent effect in the prior case and a specific counterdeterrent effect in the latter case. But even after experiencing an audit, a taxpayer does not know exactly the risk of detection he faces in the future (Alm (1988); Scotchmer and Slemrod (1989); Polinsky and Shavell (2000)). This implies that postaudit compliance depends on perceived rather than actual changes in the probability of detection. In fact, prior studies find that the experience of enforcement may change behavior, even absent any change in the underlying probability of detection (Haselhuhn *et al.* (2012); Earnhart and Friesen (2013); Simonsohn *et al.* (2008)). This effect is particularly well documented in laboratory experiments on tax compliance, where the relevant tax system parameters are typically unaffected by the audit outcome (Alm (2019); Alm and Kasper (2020)).

This raises questions about the behavioral determinants of postaudit tax compliance. For example, Mittone (2006) suggests that taxpayers falsely assume dependency of statistically independent events, such as experiencing a random tax audit. Such a bias is related to the "gambler's fallacy" and implies the misconception that a recent audit experience reduces the risk of a future audit ("bomb-crater effect"). Conversely, Spicer and Hero (1985) suggest that audited taxpayers overestimate the risk of future audits because they apply the "availability heuristic" (Tversky and Kahneman (1973)) and assess the probability of a future audit by the ease of recalling their previous audit. In fact, the availability heuristic provides a behavioral rationale for the finding that the audit experience informs a taxpayer's decision to revise upwards or to revise downwards his prior on the probability of a future audit even when the relevant parameters do not change. More specifically, Tversky and Kahneman (1974, p. 1128) argue that individuals evaluate the risk of a decision by imagining the negative outcome. If the negative outcome is "*vividly portrayed*," this event may "*appear exceedingly dangerous, although the ease with which disasters are imagined need not reflect their actual likelihood. Conversely, the risk […] may be grossly underestimated if some possible dangers are either difficult to conceive of, or simply do not come to mind*." The effectiveness (the share of undeclared income that the tax agency detected) as well as the outcome (whether or not the taxpayer was found to be noncompliant) of a past audit should thus affect a taxpayer's assessment of the risk of a future audit. More specifically, the specific deterrent effect of an effective audit should be stronger than the specific deterrent effect of an ineffective audit. Likewise, an audit that found the taxpayer to be noncompliant should have a stronger deterrent effect than an audit that found him to be compliant.

Another theory assumes that the audit experience changes a taxpayer's motivation to comply, rather than the perceived risk of future detection. As taxpayers comply for different reasons, the audit experience might have differential effects on postaudit tax compliance. For example, an honest taxpayer may find being audited unfair, or perceive the audit as a breach of trust. Similarly, the audit experience might induce negative emotions in honest individuals (Olsen *et al.* (2018); Enachescu *et al.* (2019)). Tax audits might thus have the potential to crowd out the intrinsic motivation to comply and to reduce postaudit compliance among honest taxpayers (Frey (1997); Mendoza *et al.* (2017); Lederman (2018)). Dishonest taxpayers, on the other hand,

might respond to an audit by increasing their postaudit compliance, because the experience of being punished motivates them to comply more in the future (Kirchler *et al.* (2008); Braithwaite (2003)).

In sum, theoretical studies of tax compliance suggest that financial incentives determine a taxpayer's reporting decision and that increasing the audit probability and the fines for noncompliance deter tax evasion. Behavioral studies suggest that other factors, such as a taxpayer's intrinsic motivation, determine his compliance decision. However, the effect of the audit experience on postaudit compliance is not well understood and the existing literature does not resolve crucial aspects. First, the effect of audit effectiveness on postaudit tax compliance remains unknown. Second, the mechanism by which tax audits affect truly compliant and truly noncompliant taxpayers is also unknown, even though there are various explanations for this behavior. The next section discusses our experimental design for examining these issues.

4. Experimental Setup: Design, Procedure, and Sample

Our experiment implements the fundamental elements of voluntary income tax reporting and follows the standard procedure of tax compliance experiments (Alm and Jacobson (2007)). In each round of the experiment, participants receive a random amount of income that varies between 2,000 and 3,500 Experimental Currency Units (ECU).[10] They must decide how much income to report to the tax agency, and they may report any amount between 0 ECU and the amount they received. Reported income is taxed at $t = 0.25$. Participants face the risk of being randomly selected for audit. Audit probabilities p range from 0.18 to 0.70, and tax audits differ in their effectiveness. While audits detect all undeclared income in some rounds, they detect only some fraction of undeclared income in others. Specifically, the audit effectiveness e ranges from 0.30 and 1. Consequently, the detection probability (the product of p and e) ranges from 0.18 to 0.49. The fine for noncompliance is twice the evaded amount that has been detected. Once participants have reported their income, they learn whether they have been audited or not and the outcome of the audit. This process is repeated over 28 rounds in random order. Participants do not know the number of rounds.

Table 1 shows our experimental parameters. We calibrate these parameters such that a reasonably risk-averse taxpayer should not report any income to maximize his expected profit.[11] By distinguishing between and introducing variation in the audit probability p and the audit effectiveness e, our design enables us to test whether effective and ineffective audits differ in their capacity to deter future noncompliance of audited taxpayers. Moreover, it also allows us to investigate whether taxpayers misperceive compound lotteries (where p and $e < 1$) relative to one stage lotteries (where $e = 1$) with identical detection risk (p multiplied with e); see column "*Audit Type.*" We also systematically vary the display of information on the audit probability p and the audit effectiveness e to rule out the possibility that order effects drive our results; see column "*Parameter Order.*"

All parameters are known to the participants in each round. Also, to facilitate the compliance decision, we program a calculator that shows how declared income translates into after-tax income conditional on audit effectiveness. We provide a screenshot of the experimental task in Appendix A.[12]

[10] 1,000 ECU equals € 3.50.

[11] An individual with realistic levels of constant relative risk aversion ($e \leq 1.5$) would optimally declare zero income for p = 0.26 (the average detection probability), t = 0.25, and f = 2 (see Alm (2019) for details).

[12] The experiment was programmed in z-Tree (Fischbacher (2007)).

TABLE 1. Experimental Parameters

Task	Audit Type	Parameter Order	Audit Probability	Audit Effectiveness	Detection Risk
1	Effective audit (e = 1)	p first	0.18	1.00	0.18
2			0.21	1.00	0.21
3			0.24	1.00	0.24
4			0.28	1.00	0.28
5		e first	0.18	1.00	0.18
6			0.21	1.00	0.21
7			0.24	1.00	0.24
8			0.28	1.00	0.28
9	Low audit probability (p)	p first	0.30	0.60	0.18
10			0.33	0.63	0.21
11			0.37	0.67	0.24
12			0.40	0.70	0.28
13		e first	0.30	0.60	0.18
14			0.33	0.63	0.21
15			0.37	0.67	0.24
16			0.40	0.70	0.28
17	Low audit effectiveness (e)	p first	0.60	0.30	0.18
18			0.63	0.33	0.21
19			0.67	0.37	0.24
20			0.70	0.40	0.28
21		e first	0.60	0.30	0.18
22			0.63	0.33	0.21
23			0.67	0.37	0.24
24			0.70	0.40	0.28
25	High audit probability (p) and effectiveness (e)	p first	0.60	0.60	0.36
26			0.63	0.63	0.40
27			0.67	0.67	0.44
28			0.70	0.70	0.49

NOTE: Participants face all 28 tasks in random order. Parameters are presented to participants at the beginning of each round. Parameter Order indicates how the audit probability p and the audit effectiveness e are presented to participants (p before e or vice versa).

The experiment was conducted at the Vienna Center of Experimental Economics (VCEE) in December 2019 and January 2020. Participants were recruited via ORSEE (Greiner (2015)). We used a power analysis to determine the sample size and pre-registered our study at https://osf.io/uhpmw/.[13] The final sample (n = 333) comprises data from 13 experimental sessions, and is slightly larger than the aspired sample (n = 327).

At the beginning of the experiment participants learn that their information is private and that it is impossible to identify individual participants. The study starts with a few demographic questions. Subsequently, participants learn about the compensation mechanism. Each participant receives a show-up fee of € 5.00 and an additional compensation that is based on the after-tax income of a randomly selected round. Participants

[13] Our target sample size estimate is based on a power analysis, which indicated that a sample size of N = 327 is required to detect a difference between two means (mean compliance rate after an effective vs. ineffective audit) (continued) with the following parameters: power = 0.95, alpha = 0.05, Cohen's d = 0.2, t-test for two dependent means, two-tailed).

are encouraged to earn as much money as they can. After reading a detailed introduction to the experimental task and an example of the tax compliance decision, participants must answer two questions on the definition of "audit probability" and "audit effectiveness" correctly before they can proceed. Next, they play three practice rounds. One practice round is not audited, while the two other rounds result in one effective and one ineffective audit, respectively. Participants then proceed to the experiment. After completing the 28th round, they answer a few final questions. The experiment lasts approximately 45 minutes, and the mean payoff is € 12.66.

The participant pool has a slightly larger percentage of female subjects (57 percent) than male subjects, and the pool includes students and nonstudents. The mean age is 26 years (SD = 6.06) with a range from 18 to 59 years. Most participants hold at least a high-school degree (49 percent) and study business (19 percent). While 95 percent indicate that they participated in a laboratory experiment in the past, only 16 percent state that they participated in a study on tax compliance before. Moreover, 29 percent indicate that they self-prepared a tax return in the past.

5. Results

Table 2 presents descriptive statistics. We observe 9,324 compliance decisions from 333 individuals. The actual audit probability was 0.44, and the average audit effectiveness was 0.66. Our main dependent variable is the *Compliance rate*, defined as the share of received income that was reported to the tax agency. The mean compliance rate was 0.54 (SD = 0.41), which indicates substantial underreporting in the aggregate.

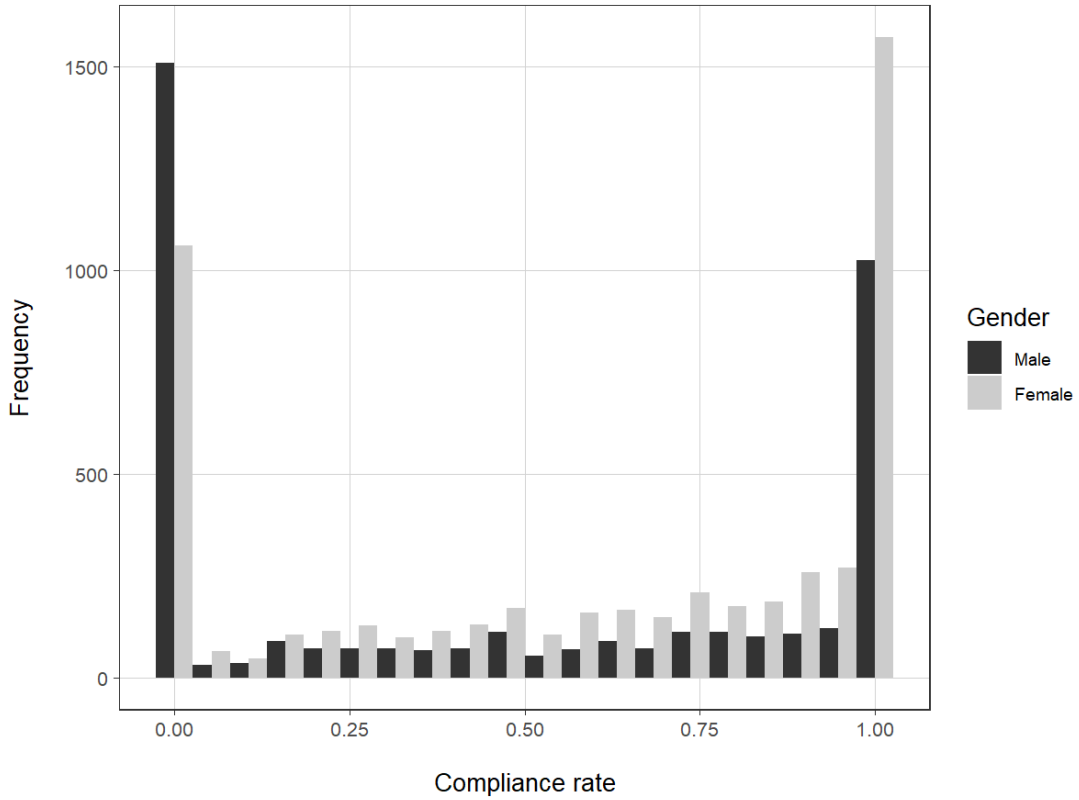
TABLE 2. Data Description

Variable	Description	Mean	SD
Dependent Variables			
Compliance rate	Reported income divided by received income	0.54	0.41
Evaded income	Income not reported on tax return (in ECU)	1248.97	1156.05
Experimental Treatment Variables			
Received income	Income received (in ECU)	2700.16	430.04
Detection risk	Audit probability multiplied with audit effectiveness	0.26	0.08
Audit probability	Probability of being audited	0.44	0.19
Audit effectiveness	Share of evaded income that the audit detects	0.66	0.25
Audit probability first	= 1 if audit probability presented before audit effectiveness	0.57	0.50
Round after audit	= 1 if round succeeds an audit and 0 if round is audited		
Noncompliant	= 1 if reported income equals 0	0.25	0.44
Compliant	= 1 if reported income equals received income	0.26	0.44
Dishonest	= 1 if reported income equals 0 for each round prior to first audit	0.11	0.31
Honest	= 1 if reported income equals received income for each round prior to first audit	0.14	0.35
Demographic Variables			
Female	= 1 if participant is female	0.57	0.50
Age	Participant's age in years	25.94	6.06
Higher education	= 1 if completed Bachelor Studies or higher	0.51	0.49
Economics major	= 1 if Major in Economics	0.08	0.27
German speaking	= 1 if Austrian or German	0.48	0.50
Prior experiments	= 1 if prior participation in laboratory experiments	0.95	0.23
Prior tax experiments	= 1 if prior participation in tax experiments	0.16	0.37
Self-preparation	= 1 if self-prepared tax return in the past	0.29	0.46
Risk seeking#	Do you like to gamble? (0 to 9)	4.36	2.36
Income maximization#	To what extent did you try to maximize your income? (0 to 9)	6.27	2.34
Tax morale#	Do you think cheating on tax if you have a chance can be justified? (0 to 9)	6.05	2.68

NOTES: # denotes a scale from 0 to 9, where higher values indicate more risk-seeking, more income maximization, and higher tax morale.

Figure 1 shows a bimodal distribution of the *Compliance rate*. Participants report zero income in 0.25 of all rounds and all income in 0.26 of all rounds. This indicates that participants differ fundamentally in their propensity to comply. While some appear to be motivated entirely by the expected value of the evasion gamble and never report any income, others report their income correctly irrespective of any incentive to cheat. Moreover, we find that female participants are substantially more compliant (mean compliance = 0.60) than male participants (mean compliance = 0.43).

FIGURE 1. Histogram of the Compliance Rate for Male and Female Participants



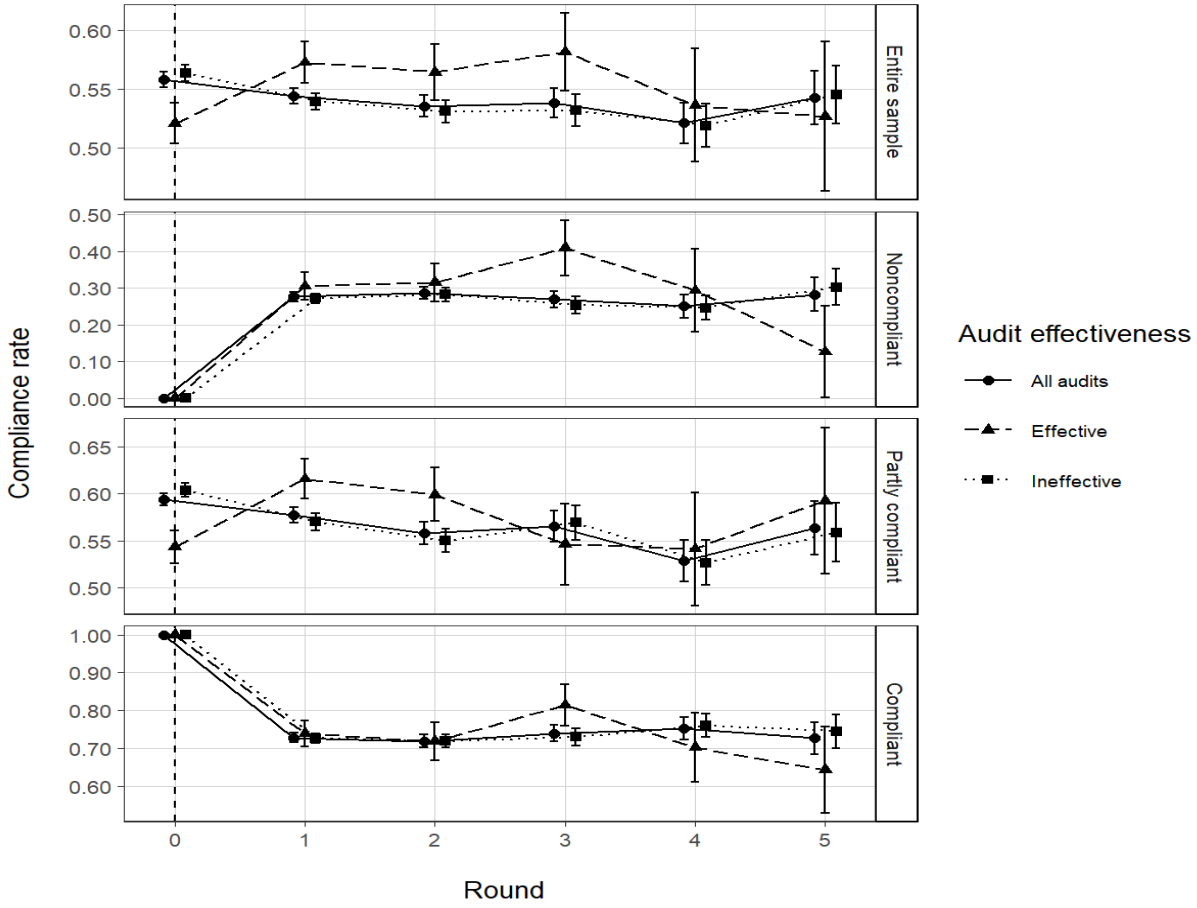
To investigate the effect of tax audits on postaudit compliance, we compare mean compliance levels in the rounds that were audited to compliance levels in subsequent rounds. Figure 2 depicts the compliance implications of effective and ineffective tax audits. We refer to the round that was audited as Round 0, where this round represents data from all rounds that were audited (4,131 rounds). Round 1 then comprises data from all rounds that follow a tax audit (4,016 rounds) and Rounds 2 to 5 summarize information from subsequent rounds.[14]

Panel 1 indicates that the aggregate effect of tax audits on subsequent compliance tends to be slightly negative. However, behavioral responses seem to depend strongly on audit effectiveness. Participants who experience an effective audit declare a larger share of their income in subsequent rounds, while postaudit compliance declines among taxpayers who experience an ineffective audit. Panel 1 also suggests that behavioral responses to ineffective audits seem to be slightly more persistent than behavioral responses to effective audits.

Panels 2 to 4 depict the effect of audits on taxpayers who were noncompliant, partly compliant, or compliant in the round that was audited. Overall, the effect of audits on postaudit compliance seems to depend strongly on prior reporting levels. While audits increase postaudit compliance considerably among noncompliant taxpayers who did not report any income in the round that was audited (Panel 2), the behavioral response of partly compliant individuals, who reported some but not all of their income seems to depend strongly on audit effectiveness (Panel 3). Finally, audits seem to decrease postaudit compliance substantially among compliant individuals who reported all income in the round that was audited (Panel 4).

[14] In case of a subsequent audit (e.g., in Round 3), information from that round is reflected both in Round 3 (reported income three rounds after experiencing an audit) and in Round 0 (reported income in a round that is audited). The reporting decision in the subsequent round is then reflected in Round 1 so that Graph 2 depicts the average effect of audits on postaudit compliance.

FIGURE 2. Effect of Audits on Postaudit Compliance



NOTES: Taxpayers were audited after declaring their income in Round 0 (dashed vertical line) and not audited again through Round 5. Panel 1 (*Entire Sample*) comprises data from all individuals (4,131 observations in Round 0). Taxpayers who were found to be *Noncompliant* (Panel 2) did not report any income in Round 0 (1,049 observations in Round 0). Taxpayers who were found to be *Partly Compliant* (Panel 3) reported some but not all of their income in Round 0 (1,916 observations in Round 0). Taxpayers who were found to be *Compliant* (Panel 4) reported all income to the tax agency in Round 0 (1,166 observations in Round 0). Effective audits detect all undeclared income. Ineffective audits detect between 30 percent and 70 percent of undeclared income. Error bars represent standard errors.

Taken together, our descriptive analyses indicate that the audit effectiveness has a strong effect on postaudit tax compliance. While effective audits increase postaudit compliance, ineffective audits seem to have the opposite effect. Moreover, audits appear to have differential effects on compliant and noncompliant (including partly compliant) taxpayers. While taxpayers who reported all their income to the tax agency in the round that was audited declare less in subsequent rounds, postaudit compliance increases among individuals who were found to be noncompliant. The next section employs regression analyses to analyze the effect of audit effectiveness on postaudit tax compliance of compliant and noncompliant taxpayers.

5.1. Regression Results

We report our main results in Tables 3 and 4. Table 3 presents regression results on the effect of audits on tax reporting in the round that follows the audit, while Table 4 shows behavioral responses in subsequent rounds (two to five rounds after the audit).[15] Our regression results provide strong evidence that tax audits have differential effects on postaudit compliance, effects that vary by audit effectiveness and also by taxpayer type. In particular, Table 3 reveals three important results. First, we find that audits have the potential to increase

[15] To identify the effect of audits on postaudit compliance, we compare compliance rates in rounds that were audited (Round 0, $n_0 = 4{,}131$) to compliance rates in the five subsequent rounds ($n_1 = 4{,}016$, $n_2 = 2{,}113$, $n_3 = 1{,}112$, $n_4 = 592$, $n_5 = 312$). Taxpayers were audited only once (in Round 0) through Round 5. Our main analysis thus identifies within-subject variation in reporting compliance that results from experiencing an audit. To test the robustness of our results, we also compare reporting compliance across audited and unaudited individuals. These results are presented in Appendix Tables B3 and B4. Our results are unaffected.

or to decrease postaudit tax compliance. Second, we find that effective audits have a more positive effect on postaudit tax compliance than ineffective audits. Third, we find that audits have differential effects on compliant and noncompliant taxpayers. More specifically, audits increase the postaudit compliance of noncompliant individuals, who did not report any income in the round that was audited, while they reduce the postaudit compliance of compliant taxpayers, who have been found to report all income correctly. Finally, Table 4 reveals that audits have sustainable effects on postaudit compliance. While the audit effectiveness has a positive effect on the postaudit compliance of taxpayers who did not report some fraction of their income for three rounds after the audit, the differential responses of compliant and noncompliant individuals persists for five rounds after the audit.

Our baseline specifications (Models 1 and 2) estimate the effect of basic economic factors (*Received income*), deterrence factors (*Detection risk, Audit probability, Audit effectiveness),* the audit experience (*Round after audit),* and several *Demographic Variables* (listed in Table 2) on the *Compliance rate.*[16] *The interaction Round after audit x Experienced effectiveness* measures the effect of the experienced audit effectiveness on postaudit compliance. While the *Detection Risk* (the product of the audit probability and the audit effectiveness) has a strong effect on compliance, the insignificant coefficients of the *Audit probability* and the *Audit effectiveness* provide no indication for a systematic misperception of either of these factors. This suggests that the risk of detection drives compliance decisions; in contrast, the presentation of the compliance decision as a one-stage or a two-stage compound lottery with identical expected outcomes does not drive compliance decisions. Similarly, the insignificant coefficient of *Audit probability* first shows that whether the audit probability is shown before the audit effectiveness (or vice versa) has no effect on compliance. Importantly, Models 1 and 2 indicate that postaudit compliance depends strongly on the audit effectiveness. While the coefficient of *Round after audit* indicates that ineffective audits reduce the postaudit compliance rate by 3 percentage points in the aggregate, the interaction term *Round after audit x Experienced effectiveness* is significant and positive. All else equal, we estimate that experiencing an effective audit increases postaudit compliance by 3 percentage points.

Models 3 to 6 complement these findings and show that prior compliance has a strong effect on postaudit compliance. Specifically, Models 3 and 4 add the indicator variable *Noncompliant* that equals 1 if a taxpayer did not report any income in a round that was audited. The negative coefficient of *Round after audit* shows that audits reduce postaudit compliance among taxpayers who reported at least some fraction of their income by approximately 8 percentage points, while the insignificant interaction term *Round after audit x Experienced effectiveness* suggests that the experienced audit effectiveness has no effect on the postaudit compliance of these taxpayers. As discussed below, these results are driven by compliant taxpayers, whose substantial decline in postaudit compliance is unaffected by the audit effectiveness. Moreover, Models 3 and 4 show that audits increase postaudit tax compliance of *Noncompliant* taxpayers substantially. On average *Noncompliant* individuals report over fifty percentage points less income than other taxpayers. However, noncompliant individuals increase their reported income by approximately 20 percentage points one round after experiencing an audit (*Round after audit x Noncompliant*).

Finally, Models 5 and 6 replace the indicator variable *Noncompliant* with the indicator variable *Compliant* that equals 1 if a taxpayer reported all income in a round that was audited. Our estimates indicate that ineffective audits increase the postaudit compliance of individuals who were not found to be compliant by approximately 5 percentage points (*Round after audit*) and that an effective audit increases postaudit compliance of those taxpayers by 6 percentage points compared to an ineffective audit (*Round after audit x Experienced Effectiveness*). Moreover, we estimate that *Compliant* taxpayers, who report over 40 percentage points more income than other taxpayers, reduce their postaudit tax compliance by approximately 24 percentage points in the round after an audit (*Round after audit x Compliant).*

With regard to the demographic variables, we find that age and being female has a positive effect on compliance, that participants from German-speaking countries are less compliant than participants from other countries, and that individuals who indicated in the post-experimental survey that they tried to maximize their income reported smaller shares of their income.

[16] To test the robustness of our results, we also use *Evaded income* as the dependent variable. These results are presented in Appendix Tables B1 and B2. Our results are unaffected.

TABLE 3. Effect of Audits One Round After the Audit

Dependent variable: Compliance rate

Independent variable	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	0.2984*** (0.0433)	0.3718*** (0.0945)	0.4659*** (0.0379)	0.4897*** (0.0679)	0.2256*** (0.0412)	0.2821*** (0.0804)
Received income	-0.0145*** (0.0030)	-0.0146*** (0.0030)	-0.0140*** (0.0028)	-0.0142*** (0.0028)	-0.0109*** (0.0030)	-0.0111*** (0.0030)
Detection risk	0.0082*** (0.0009)	0.0082*** (0.0009)	0.0056*** (0.0008)	0.0056*** (0.0008)	0.0071*** (0.0009)	0.0071*** (0.0009)
Audit probability	0.0003 (0.0006)	0.0003 (0.0006)	0.0012** (0.0006)	0.0012** (0.0006)	0.0001 (0.0006)	0.0001 (0.0006)
Audit effectiveness	0.0002 (0.0004)	0.0002 (0.0004)	0.0004 (0.0004)	0.0004 (0.0004)	0.0001 (0.0004)	0.0001 (0.0004)
Audit probability first	-0.0035 (0.0064)	-0.0034 (0.0064)	-0.0036 (0.0059)	-0.0035 (0.0059)	-0.0032 (0.0063)	-0.0031 (0.0063)
Round after audit	-0.0274** (0.0123)	-0.0273** (0.0123)	-0.0829*** (0.0120)	-0.0835*** (0.0121)	0.0463*** (0.0127)	0.0465*** (0.0127)
Round after audit x Experienced effectiveness	0.0006*** (0.0002)	0.0006*** (0.0002)	0.0002 (0.0002)	0.0002 (0.0002)	0.0006*** (0.0002)	0.0006*** (0.0002)
Noncompliant			-0.5369*** (0.0105)	-0.5340*** (0.0106)		
Round after audit x Noncompliant			0.2829*** (0.0124)	0.2836*** (0.0124)		
Compliant					0.4285*** (0.0108)	0.4272*** (0.0108)
Round after audit x Compliant					-0.2861*** (0.0128)	-0.2864*** (0.0128)
Demographic variables		included		included		included
Observations	8,147	8,147	8,147	8,147	8,147	8,147
n	333	333	333	333	333	333
R^2	0.681	0.656	0.633	0.644	0.638	0.637

NOTES: ** and *** indicate significance at the 5%, and 1% levels, respectively; none were significant at the 10% level. Robust standard errors (in parentheses) are clustered at the individual level. Continuous predictors are scaled.

Table 4 presents regression results for subsequent rounds (two to five rounds after the audit). The effect of the Received income on compliance remains negative across all specifications, while the effect of the Detection risk on compliance remains large and positive. Again, the coefficients of the Audit probability and the Audit effectiveness provide no indication for a systematic misperception of either of these factors. This suggests that the experience of an audit (whether effective or not) does not induce a bias in the evaluation of these factors in subsequent compliance decisions.

The interaction terms Round after audit x Experienced effectiveness suggest that the effect of Audit Effectiveness is strongest for taxpayers who did not report some fraction of their income. Among those taxpayers, effective audits increase postaudit compliance by 6 percentage points two rounds after the audit (Model 9), and 8 percentage points three rounds after the audit (Model 12) compared to ineffective audits. Surprisingly, our estimates indicate that this effect reverts over time: while the interaction effect is insignificant four rounds after the audit (Model 15), experienced audit effectiveness has a negative effect on audited taxpayers who were not found to be compliant five rounds after the audit (Model 18), where an effective audit reduces postaudit compliance by 17 percentage points.

The differential responses of Noncompliant and Compliant taxpayers are even more persistent. We estimate that audits increase postaudit compliance of Noncompliant taxpayers for five rounds after the audit (Round after audit x Noncompliant). While the increase in postaudit compliance attenuates from approximately 19 percentage points two rounds after the audit (Model 8) to approximately 14 percentage points increase four rounds after the audit (Model 14), we estimate that postaudit compliance levels of taxpayers who have been found to be noncompliant are 19 percentage points higher five rounds after the audit than they were before the audit (Model 17). Similarly, our estimates indicate that the audit experience reduces postaudit compliance of Compliant taxpayers for five rounds (Round after audit x Compliant). Those taxpayers report approximately 24 percentage points less income two rounds after an audit (Model 9), and five rounds after the audit (Model 18) postaudit compliance is still approximately 10 percentage points below preaudit levels.

TABLE 4. Effect of Audits on Postaudit Tax Compliance

Dependent variable: Compliance rate

Independent variable	Two rounds after the audit			Three rounds after the audit			Four rounds after the audit			Five rounds after the audit		
	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
Intercept	0.5014*** (0.0987)	0.6003*** (0.0644)	0.4069*** (0.0790)	0.4382*** (0.1013)	0.6118*** (0.0593)	0.3720*** (0.0785)	0.4483*** (0.1052)	0.6109*** (0.0559)	0.3405*** (0.0783)	0.4328*** (0.1049)	0.6401*** (0.0542)	0.3489*** (0.0777)
Received income	-0.0166*** (0.0033)	-0.0149*** (0.0029)	-0.0134*** (0.0033)	-0.0148*** (0.0037)	-0.0128*** (0.0028)	-0.0076** (0.0034)	-0.0143*** (0.0038)	-0.0114*** (0.0026)	-0.0059* (0.0034)	-0.0131*** (0.0040)	-0.0103*** (0.0026)	-0.0044 (0.0034)
Detection risk	0.0084*** (0.0010)	0.0041*** (0.0008)	0.0069*** (0.0010)	0.0085*** (0.0011)	0.0032*** (0.0008)	0.0064*** (0.0010)	0.0084*** (0.0011)	0.0026*** (0.0008)	0.0058*** (0.0010)	0.0087*** (0.0012)	0.0028*** (0.0008)	0.0062*** (0.0011)
Audit probability	-0.0008 (0.0007)	0.0008 (0.0006)	-0.0013* (0.0007)	-0.0003 (0.0008)	0.0013** (0.0006)	-0.0008 (0.0008)	-0.0001 (0.0008)	0.0015** (0.0006)	-0.0006 (0.0008)	-0.0001 (0.0009)	0.0013** (0.0006)	-0.0009 (0.0008)
Audit effectiveness	-0.0005 (0.0005)	0.0001 (0.0004)	-0.0006 (0.0005)	-0.0002 (0.0006)	0.0003 (0.0004)	-0.0004 (0.0005)	-0.0002 (0.0006)	0.0004 (0.0004)	-0.0001 (0.0005)	-0.0003 (0.0006)	0.0001 (0.0004)	-0.0004 (0.0006)
Audit probability first	-0.0034 (0.0072)	-0.0002 (0.0061)	-0.0002 (0.0070)	-0.0033 (0.0080)	-0.0063 (0.0060)	-0.0054 (0.0074)	-0.0070 (0.0081)	-0.0056 (0.0057)	-0.0084 (0.0073)	-0.0091 (0.0086)	-0.0060 (0.0056)	-0.0080 (0.0074)
Round after audit	-0.0262 (0.0165)	-0.0889*** (0.0148)	0.0374** (0.0166)	-0.0394* (0.0225)	-0.0833*** (0.0176)	0.0158 (0.0213)	-0.0450 (0.0294)	0.0660*** (0.0217)	0.0060 (0.0269)	0.0631 (0.0408)	0.0298 (0.0281)	0.1537*** (0.0359)
Round after audit x Experienced effectiveness	0.0004 (0.0003)	0.0002 (0.0002)	0.0006** (0.0003)	0.0006 (0.0004)	0.0003 (0.0003)	0.0008** (0.0003)	0.0004 (0.0005)	-0.0002 (0.0003)	0.0003 (0.0004)	-0.0015** (0.0007)	-0.0017*** (0.0004)	-0.0017*** (0.0006)
Noncompliant		-0.6178*** (0.0100)			-0.6781*** (0.0091)			-0.7079*** (0.0084)			-0.7268*** (0.0081)	
Round after audit x Noncompliant		0.2839*** (0.0137)			0.2326*** (0.0152)			0.2091*** (0.0182)			0.2281*** (0.0241)	
Compliant			0.4736*** (0.0109)			0.5220*** (0.0108)			0.5534*** (0.0104)			0.5693*** (0.0103)
Round after audit x Compliant			-0.2768*** 0.0150)			-0.2471*** (0.0182)			-0.1959*** (0.0225)			-0.2596*** (0.0301)
Demographic variables	included	included	included	included	included	included	included	included	included	included	included	included
Observations	6,244	6,244	6,244	5,255	5,255	5,255	4,723	4,723	4,723	4,443	4,443	4,443
n	333	333	333	333	333	333	333	333	333	333	333	333
R^2	0.668	0.704	0.654	0.662	0.770	0.687	0.694	0.813	0.731	0.678	0.832	0.738

NOTES: *, **, and *** indicate significance at the 10%, 5%, and 1% level respectively. Robust standard errors (in parentheses) are clustered at the individual level. Continuous predictors are scaled.

To further investigate the effect of tax audits on individuals who differ in their propensity to comply, Models 19 to 22, presented in Table 5, estimate the effect of the first audit that taxpayers experience on the *Compliance rate* in the subsequent round. This reduces the number of observations to 666 (n = 333). Due to the small sample size, we do not distinguish between different levels of audit effectiveness. Our experimental parameters are calibrated such that the profit-maximizing strategy is to report zero income in every round (see Footnote 11 for details). To identify the effect of an audit on individuals who are motivated entirely by the expected value of the evasion gamble, we introduce the indicator variable *Dishonest* (n_D = 37), which equals 1 if a taxpayer reported zero income in all rounds prior to his first audit. The interaction *Round after audit x Dishonest* thus identifies the effect of the first audit on *Dishonest* taxpayers. Our estimates indicate that the experience of the first audit increases the postaudit *Compliance rate* of *Dishonest* taxpayers by 19 percentage points (Models 19 and 20).

Conversely, Models 21 and 22 investigate the hypothesis that audits "crowd out" the intrinsic motivation of honest taxpayers to comply regardless of any incentive to cheat. We therefore replace the variable *Dishonest* with the indicator variable *Honest* that equals 1 if a taxpayer reported all income in all rounds prior to his first audit (n_H = 46). The interaction *Round after audit x Honest* thus identifies the effect of the first audit on *Honest* taxpayers. Our estimates indicate that the first audit does not reduce the postaudit compliance of *Honest* taxpayers (p = .105 in Model 21 and p = .115 in Model 22). Therefore, we find no support for the hypothesis that audits crowd out the intrinsic motivation to comply among honest individuals.

TABLE 5. Effect of First Audits on Dishonest and Honest Taxpayers
Dependent variable: Compliance rate

Independent variable	(19)	(20)	(21)	(22)
Intercept	0.3996**	0.3839*	0.2803	0.2688
	(0.1730)	(0.1959)	(0.1812)	(0.2030)
Received income	0.0001	0.0014	0.0002	0.0016
	(0.0127)	(0.0130)	(0.0133)	(0.0133)
Detection risk	0.0100***	0.0110***	0.0116***	0.0121***
	(0.0038)	(0.0039)	(0.0040)	(0.0040)
Audit probability	-0.0004	-0.0008	-0.0013	-0.0014
	(0.0028)	(0.0029)	(0.0030)	(0.0030)
Audit effectiveness	0.0002	-0.0002	-0.0005	-0.0007
	(0.0020)	(0.0020)	(0.0020)	(0.0020)
Audit probability first	-0.0024	-0.0021	0.0152	0.0118
	(0.0278)	(0.0284)	(0.0291)	(0.0291)
Round after audit	-0.0088	-0.0082	0.0325	0.0326
	(0.0223)	(0.0232)	(0.0230)	(0.0232)
Dishonest	-0.6459***	-0.5676***		
	(0.0648)	(0.0657)		
Round after audit x Dishonest	0.1893***	0.1903***		
	(0.0658)	(0.0685)		
Honest			0.4725***	0.3972***
			(0.0637)	(0.0656)
Round after audit x Honest			-0.1008	-0.0988
			(0.0622)	(0.0628)
Demographic variables		included		included
Observations	666	666	666	666
N	333	333	333	333
R^2	0.605	0.577	0.612	0.604

NOTES: *, **, and *** indicate significance at the 10%, 5%, and 1% level. Robust standard errors (in parentheses) are clustered at the individual level. Continuous predictors are scaled.

5.2. Supplemental Analysis

We also estimate additional regression models to examine the robustness of our findings; these results are reported in Appendix B. First, we investigate whether the experienced fine for noncompliance, rather than the audit effectiveness, determines postaudit compliance. Therefore, we add the *Experienced fine* to our explanatory variables (Table B1). The interaction *Round after audit x Experienced fine* captures changes in postaudit compliance that result from the experienced fine, while the term *Round after audit x Experienced effectiveness x Experienced fine* identifies whether behavioral responses to differences in audit effectiveness depend on the experienced fine. Our estimates indicate that audit effectiveness, but not the experienced fine determine postaudit tax compliance. However, while the interaction *Round after audit x Experienced fine* is insignificant, the significant 3-way interaction indicates that effective audits increase postaudit tax compliance when experienced fines are high, but not when experienced fines are low. The dynamic between the experienced audit effectiveness and the experienced fine is depicted in Figure B1.

Finally, we test whether our results are robust to changes in the dependent variable (Table B2), with Models V to VIII testing whether using *Evaded income* (i.e., received income minus reported income) as the dependent variable affects the results. As expected, changing the dependent variable does not affect our results. Likewise, the effect of the first audit on the evaded income of *Honest* and *Dishonest* individuals is in line with the results reported above (Models IX and X). However, Model X indicates a marginally significant increase in evaded income among *Honest* taxpayers who were audited in the last round ($p = 0.082$).

6. Conclusions

How do tax audits affect postaudit tax compliance? In this paper we study the specific deterrent effect of tax audits by analyzing two aspects of behavioral responses to enforcement. First, we investigate how ineffective audits that do not detect all undeclared income affect subsequent reporting behavior. This also allows us to test whether presenting the compliance decision as a two-stage compound lottery with uncertain detection affects compliance decisions relative to a single-stage lottery with certain detection. Second, we analyze how tax audits affect truly compliant and truly noncompliant taxpayers, by examining the behavioral mechanisms that drive these responses. We investigate these issues in a preregistered laboratory experiment in which taxpayers receive income and decide how much they declare to the tax agency. They face the risk of being audited and a fine for undeclared income that is detected on audit. We introduce variation in the audit probability and audit effectiveness in order to assess behavioral responses to changes in these factors.

Our results suggest that tax audits have different effects on postaudit compliance and that behavioral responses to enforcement are not always in line with the assumptions of the standard model of tax evasion (Allingham and Sandmo (1972)). Specifically, we do not find that tax audits have a positive effect on subsequent reporting compliance in the aggregate. However, our estimates indicate that the specific deterrent effect of tax audits depends strongly on audit effectiveness. While taxpayers who experienced an effective audit that detected all undeclared income comply more in subsequent periods, those who experienced an ineffective audit show the opposite response. This suggests that ineffective tax audits stimulate risk-taking, and that taxpayers whose underreporting was not detected during an audit contribute to the decline in postaudit compliance found in prior studies (Gemmell and Ratto (2012); Beer *et al.* (2020)). As compound compliance lotteries (with ineffective audits) do not affect compliance compared to single-stage lotteries (with certain detection), we can rule out that a misperception of either of these factors drives behavioral responses to effective and ineffective audits. Indeed, it is important to recognize that participants knew the exact consequences of their reporting decisions, which reduces the margin for such bias. We also show that compliance choices are unaffected by the way in which the relevant factors are presented (e.g., showing the audit probability before the audit effectiveness and vice versa).

We also find consistent and robust evidence that postaudit compliance depends on taxpayers' prior reporting behavior. While taxpayers who were caught cheating report substantially more income for five rounds after the audit, individuals who reported all income in the round that was audited reduce their postaudit tax payments for five rounds. This result provides a more nuanced perspective on the finding that audited taxpayers generally tend to underestimate the risk of future examinations (Guala and Mittone (2005); Mittone

(2006); Mittone *et al.* (2017)), and indicates that loss-repair motivations alone do not explain behavioral responses to enforcement because taxpayers who were found to be compliant seem to infer that the risk of a future examination is low (Maciejovsky *et al.* (2007); McKee *et al.* (2018)).

An alternative explanation for differential responses to audits is that audits affect different types of taxpayers differently. In particular, some studies suggest that compliant taxpayers might reduce their postaudit compliance because these individuals perceive the audit as a sign of distrust of the tax agency, which reduces their intrinsic motivation to comply in the future (Frey (1997); Mendoza *et al.* (2017); Lederman (2018); Hu and Ben-Ner (2020)). To investigate this hypothesis, we analyze how audits affect honest and dishonest taxpayers who always report all or zero income prior to their first audit. While postaudit compliance increases among dishonest individuals, the effect of audits on the reporting compliance of honest taxpayers is insignificant. Thus, we do not find evidence that experiencing an audit crowds out the intrinsic motivation to comply of honest taxpayers.

Taken together, our findings challenge the standard result—and common assumption—that more audits always lead to more compliance. This has important implications for tax administrations. Our study suggests that increasing the capacity of tax audits to detect noncompliance as well as improving the targeting of noncompliant taxpayers are crucial in establishing and maintaining compliance.

Future work should investigate the effect of the audit selection mechanism on subsequent compliance. While in practice most audits target taxpayers with a relatively high likelihood of noncompliance, our study employs a random audit selection mechanism, common to many if not all laboratory experiments. A taxpayer, and particularly a compliant taxpayer, who has been randomly selected for audit might fall for the "bomb crater" fallacy, underestimate the risk of a future examination, and thus decide to report less income after the audit. Conversely, taxpayers who have been targeted based on their prior reporting behavior might be less likely to exhibit such bias. Finally, future studies might investigate how uncertainty about the audit probability and the audit effectiveness affects subsequent compliance.

References

Advani, A., Elming, W., & Shaw, J. (2017). "The dynamic effects of tax audits." No. W17/24. London, UK: The Institute for Fiscal Studies.

Allingham, M.G., & Sandmo, A. (1972). "Income tax evasion: A theoretical analysis." *Journal of Public Economics* 1 (3–4): 323–338.

Alm, J. (1988). "Uncertain tax policies, individual behavior, and welfare." *The American Economic Review,* 78 (1): 237–245.

Alm, J. (2019). "What motivates tax compliance?" *Journal of Economic Surveys* 33 (2): 353–388.

Alm, J., & Jacobson, S. (2007). "Using laboratory experiments in public economics." *National Tax Journal* 60 (1): 129–152.

Alm, J., & Kasper, M. (2020). "Laboratory Experiments," in B. Van Rooij & D. Sokol, ed., Cambridge Handbook of Compliance. Cambridge, UK: Cambridge University Press (in press).

Alm, J., McClelland, G.H., & Schulze, W.D. (1992). "Why do people pay taxes?" *Journal of Public Economics* 48: 21–38.

Alm, J., & McKee, M. (2006). "Audit certainty, audit productivity, and taxpayer compliance." *National Tax Journal* 59 (4): 801–816.

Andreoni, J., Erard, B., & Feinstein, J. (1998). "Tax compliance." *The Journal of Economic Literature* 36 (2): 818–860.

Becker, G.S. (1968). "Crime and punishment—An economic approach." *The Journal of Political Economy* 76 (2): 169–217.

Beer, S., Kasper, M., Kirchler, E., & Erard, B. (2020). "Do audits deter or provoke future tax noncompliance? Evidence on self-employed taxpayers." *CESifo Economic Studies* 66 (3): 248–264.

Bergman, M., & Nevarez, A. (2006). "Do audits enhance compliance? An empirical assessment of VAT enforcement." *National Tax Journal,* 817–832.

Bernasconi, M. (1998). "Tax evasion and orders of risk aversion." *Journal of Public Economics* 67 (1): 123–134.

Bernasconi, M., & Bernhofer, J. (2020). "Catch me if you can: Testing the reduction of compound lotteries axiom in a tax compliance experiment." *Journal of Behavioral and Experimental Economics* 84. Article 101479.

Bernasconi, M., and Zanardi, A. (2004). "Tax evasion, tax rates and reference dependence." FinanzArchiv 60: 422–445.

Braithwaite, V. (2003). "Taxing democracy: understanding tax avoidance and tax evasion," in V. Braithwaite, ed., Dancing with Tax Authorities: Motivational Postures and Noncompliant Actions. Aldershot, UK: Ashgate,15–39.

Braithwaite, V. (2009). Defiance in Taxation and Governance—Resisting and Dismissing Authority in a Democracy. Cheltenham, UK, and Northhampton, MA: Edward Elgar Publishing.

Chalfin, A., & McCrary, J. (2017). "Criminal deterrence: A review of the literature." *Journal of Economic Literature,* 55(1), 5–48.

DeBacker, J., Heim, B.T., Tran, A., & Yuskavage, A. (2018). "Once bitten, twice shy? The lasting impact of IRS audits on individual tax reporting." *The Journal of Law and Economics* 61 (1): 1–35.

Cullen, F. T., Jonson, C. L., & Nagin, D. S. (2011). "Prisons do not reduce recidivism: The high cost of ignoring science." *The Prison Journal*, 91(3_suppl), 48S–65S.

Dhami, S., and al-Nowaihi, A. (2007). "Why do people pay taxes? Prospect theory versus expected utility theory." *Journal of Economic Behavior & Organization* 64(1): 171–192.

Dillenberger, D. (2010). "Preferences for one-shot resolution of uncertainty and Allais-type behavior." *Econometrica* 78 (6): 1973–2004.

Earnhart, D., & Friesen, L. (2013). "Can punishment generate specific deterrence without updating? Analysis of a stated choice scenario." *Environmental and Resource Economics,* 56(3), 379–397.

Enachescu, J., Olsen, J., Kogler, C., Zeelenberg, M., Breugelmans, S. M., & Kirchler, E. (2019). "The role of emotions in tax compliance behavior: A mixed-methods approach." *Journal of Economic Psychology* 74.

Erard, B. (1992). "The influence of tax audits on reporting behavior," in J. Slemrod, ed., *Why People Pay Taxes: Tax Compliance and Enforcement.* Ann Arbor, MI: The University of Michigan Press, 95–114.

Erard, B., & Feinstein, J. S. (1994). "Honesty and evasion in the tax compliance game." *The RAND Journal of Economics,* 1–19.

Feinstein, J. S. (1991). "An econometric analysis of income tax evasion and its detection." *The RAND Journal of Economics,* 14–35.

Fischbacher, U. (2007). "z-Tree: Zurich toolbox for ready-made economic experiments." *Experimental Economics* 10 (2): 171–178.

Frey, B. (1997). *Not Just for the Money—An Economic Theory of Personal Motivation.* Cheltenham, United Kingdom: Edward Elgar Publishing Limited.

Gemmell, N., & Ratto, M. (2012). "Behavioral responses to taxpayer audits: Evidence from random taxpayer inquiries." *National Tax Journal* 65 (1): 33–58.

Greiner, B. (2015). "Subject pool recruitment procedures: Organizing experiments with ORSEE." *Journal of the Economic Science Association* 1 (1): 114–125.

Guala F., & Mittone, L. (2005), "Experiments in economics: External validity and the robustness of phenomena." *Journal of Economic Methodology* 12: 495–515.

Hashimzade, N., Myles, G.D., & Tran-Nam, B. (2013). "Applications of behavioural economics to tax evasion." *Journal of Economic Surveys* 27(5): 941–977.

Harrison, G.W., Martinez-Correa, J., & Swarthout, J.T. (2015). "Reduction of compound lotteries with objective probabilities: Theory and evidence." *Journal of Economic Behavior and Organization* 119: 32–55.

Haselhuhn, M. P., Pope, D. G., Schweitzer, M. E., & Fishman, P. (2012). "The impact of personal experience on behavior: Evidence from video-rental fines." *Management Science,* 58(1), 52–61.

Hu, F., & Ben-Ner, L. (2020). "The effects of feedback on lying behavior: Experimental evidence." *Journal of Economic Behavior and Organization* 171: 24–34.

Internal Revenue Service (2019). Data Book, 2018, Publication 55B. Washington, D.C.: Internal Revenue Service.

Kahneman, D., & Tversky, A. (1979). "Prospect theory: An analysis of decision under risk." *Econometrica* 47 (2): 263–292.

Kastlunger, B., Kirchler, E., Mittone, L., & Pitters, J. (2009). "Sequences of audits, tax compliance, and taxpaying strategies." *Journal of Economic Psychology* 30: 405–418.

Keen, M., & Slemrod, J. (2017). "Optimal tax administration." *Journal of Public Economics.* 152: 133–142.

Kirchler, E. (2007). *The Economic Psychology of Tax Behavior.* Cambridge, UK: Cambridge University Press.

Kirchler, E., Hoelzl, E., & Wahl, I. (2008). "Enforced versus voluntary tax compliance: the 'slippery slope' framework." *Journal of Economic Psychology* 29: 210–225.

Kleven, H. J., Knudsen, M. B., Kreiner, C.T., Pedersen, S., & Saez, E. (2011). "Unwilling or unable to cheat? Evidence from a randomized tax audit experiment in Denmark." *Econometrica* 79 (3): 651–692.

Lederman, L. (2018). "Does enforcement reduce voluntary tax compliance?" *Brigham Young University Law Review* 3: 623–694.

Maciejovsky, B., Kirchler, E., & Schwarzenberger, H. (2007). "Misperceptions of chance and loss repair: On the dynamics of tax compliance." *Journal of Economic Psychology* 28 (6): 678–691.

Matsueda, R. L., Kreager, D. A., & Huizinga, D. (2006). "Deterring delinquents: A rational choice model of theft and violence." *American Sociological Review,* 71(1): 95–122.

McKee, M., Siladke, C. A., & Vossler, C. A. (2018). "Behavioral dynamics of tax compliance when taxpayer assistance services are available." *International Tax and Public Finance* 25 (3): 722–756.

Mendoza, J. P., Wielhouwer, J. L., & Kirchler, E. (2017). "The backfiring effect of auditing on tax compliance." *Journal of Economic Psychology* 62: 284–294.

Mittone, L. (2006). "Dynamic behaviour in tax evasion: An experimental approach." *Journal of Socio-Economics* 35 (5): 813–835.

Mittone, L., F. Panebianco, F., & Santoro, A. (2017). "The bomb-crater effect of tax audits: Beyond the misperception of chance." *Journal of Economic Psychology* 61: 225–243.

Nagin, D. S. (2013a). "Deterrence in the twenty-first century." *Crime and Justice* 42(1): 199–263.

Nagin, D. S. (2013b). "Deterrence: A review of the evidence by a criminologist for economists." *Annual Review of Economics,* 5(1): 83–105.

Nagin, D. S., Cullen, F. T., & Jonson, C. L. (2009). "Imprisonment and reoffending." *Crime and Justice* 38(1): 115–200.

Olsen, J., Kasper, M., Enachescu, J., Benk, S., Budak, T., & Kirchler, E. (2018). Emotions and tax compliance among small business owners: An experimental survey. *International Review of Law and Economics,* 56: 42–52.

Polinsky, A. M., & Shavell, S. (2000). "The economic theory of public enforcement of law." *Journal of Economic Literature* 38 (1): 45–76.

Prokosheva, S. (2016). "Comparing decisions under compound risk and ambiguity: The importance of cognitive skills." *Journal of Behavioral and Experimental Economics* 64: 94–105.

Rablen, M.D. (2014). "Audit probability versus effectiveness: The Beckerian approach revisited." *Journal of Public Economic Theory* 16 (2): 322–342.

Scotchmer, S., & Slemrod, J. (1989). "Randomness in tax enforcement." *Journal of Public Economics,* 38 (1): 17–32.

Simonsohn, U., Karlsson, N., Loewenstein, G., & Ariely, D. (2008). "The tree of experience in the forest of information: Overweighing experienced relative to observed information." *Games and Economic Behavior* 62 (1): 263–286.

Snow, A., & Warren, R. S. (2007). "Audit uncertainty, Bayesian updating and tax evasion." *Public Finance Review* 35 (5): 555–571.

Skinner, J., & Slemrod, J. (1985). "An economic perspective on tax evasion." *National Tax Journal* 38 (3): 345–353.

Slemrod, J. (2019). "Tax Compliance and Enforcement." The *Journal of Economic Literature* 57 (4): 904–954.

Spicer, M. W., & Hero, R. E. (1985). "Tax evasion and heuristics: A research note." *Journal of Public Economic,* 26: 263–267.

Srinivasan, T. N. (1973). "Tax evasion: A model." *Journal of Public Economics* 2(4): 339–346.

Torgler, B. (2003). "Tax morale, rule-governed behaviour, and trust." *Constitutional Political Economy* 14 (2): 119–140.

Tversky, A., & Kahneman, D. (1973). "Availability: A heuristic for judging frequency and probability." *Cognitive Psychology* 5 (2): 207–232.

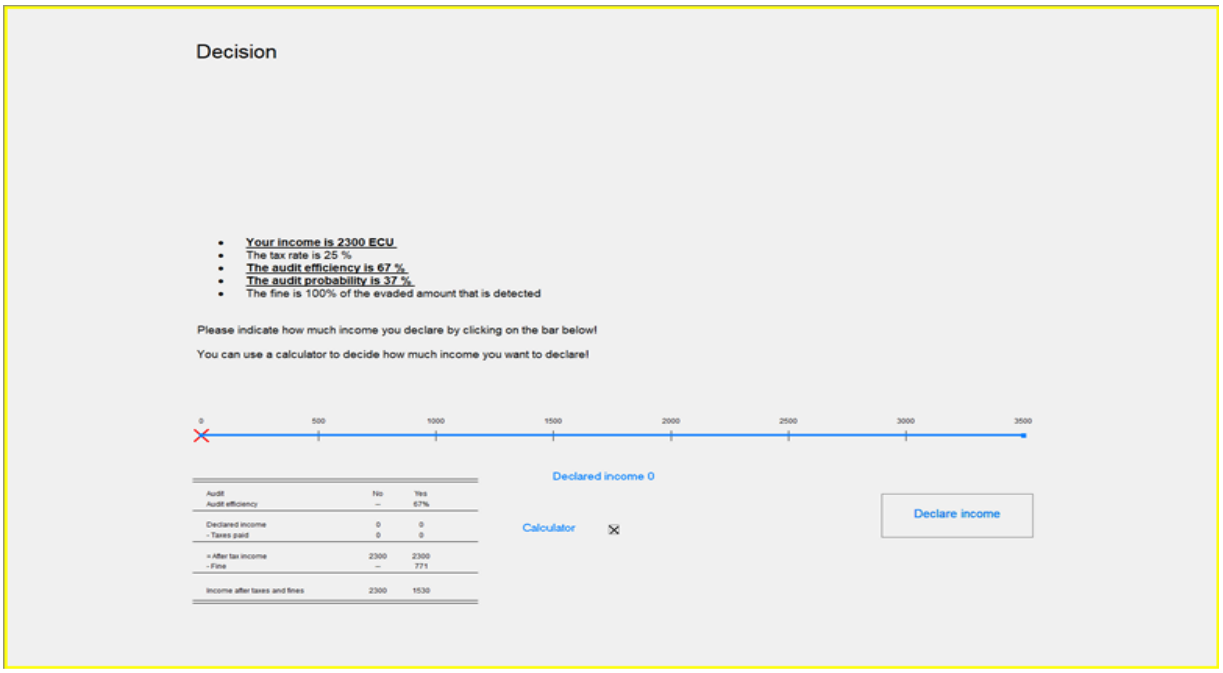
Tversky, A., & Kahneman, D. (1974). "Judgment under uncertainty: Heuristics and biases." *Science* 185 (4157): 1124–1131.

Yaniv, G. (1999) "Tax compliance and advance tax payments: A prospect theory analysis." *National Tax Journal* 52 (4): 753–764.

Yitzhaki, S. (1974). "A note on income tax evasion: A theoretical analysis." *Journal of Public Economics* 3: 201–202.

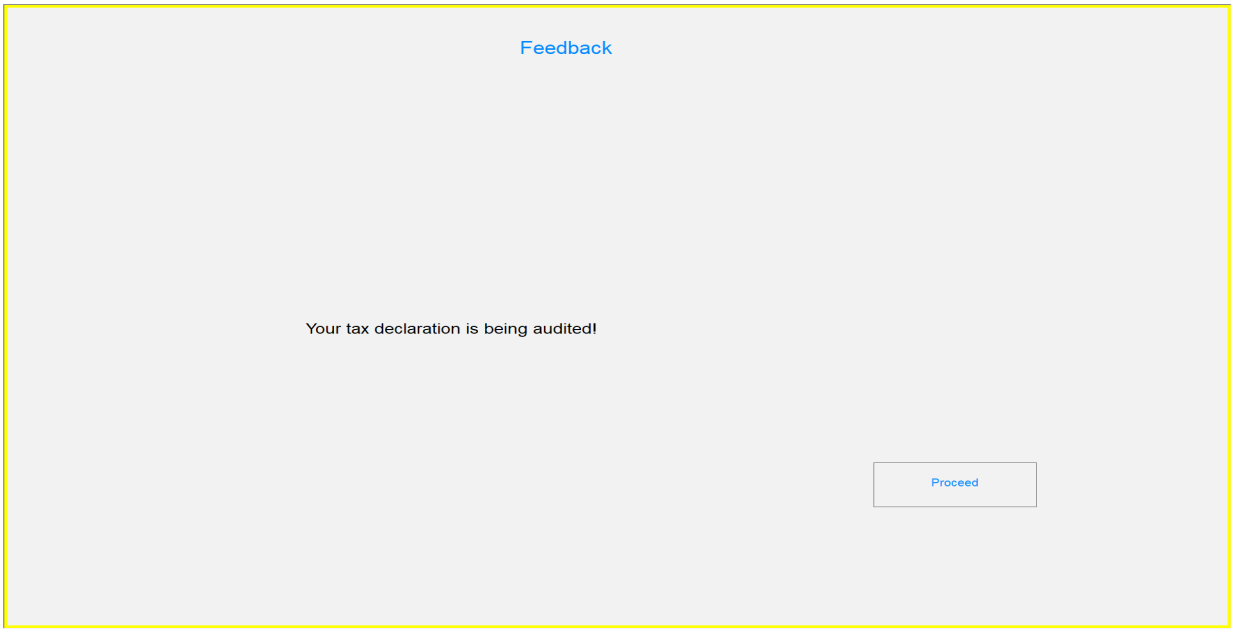
Appendix A

EXPERIMENTAL Task

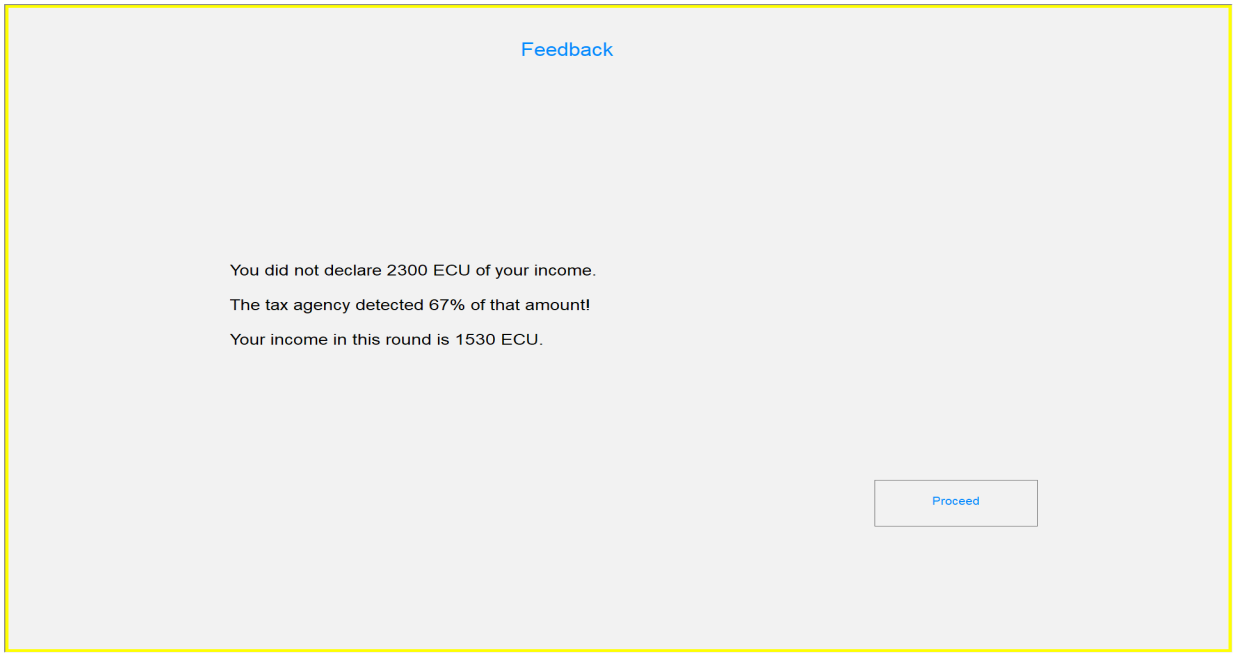


NOTES: Compliance choice for Task 23: “low audit effectiveness,” “e first,” e = .67, p = .37, detection risk = .24.

FEEDBACK: Tax declaration is being audited



FEEDBACK: Audit result



Appendix B

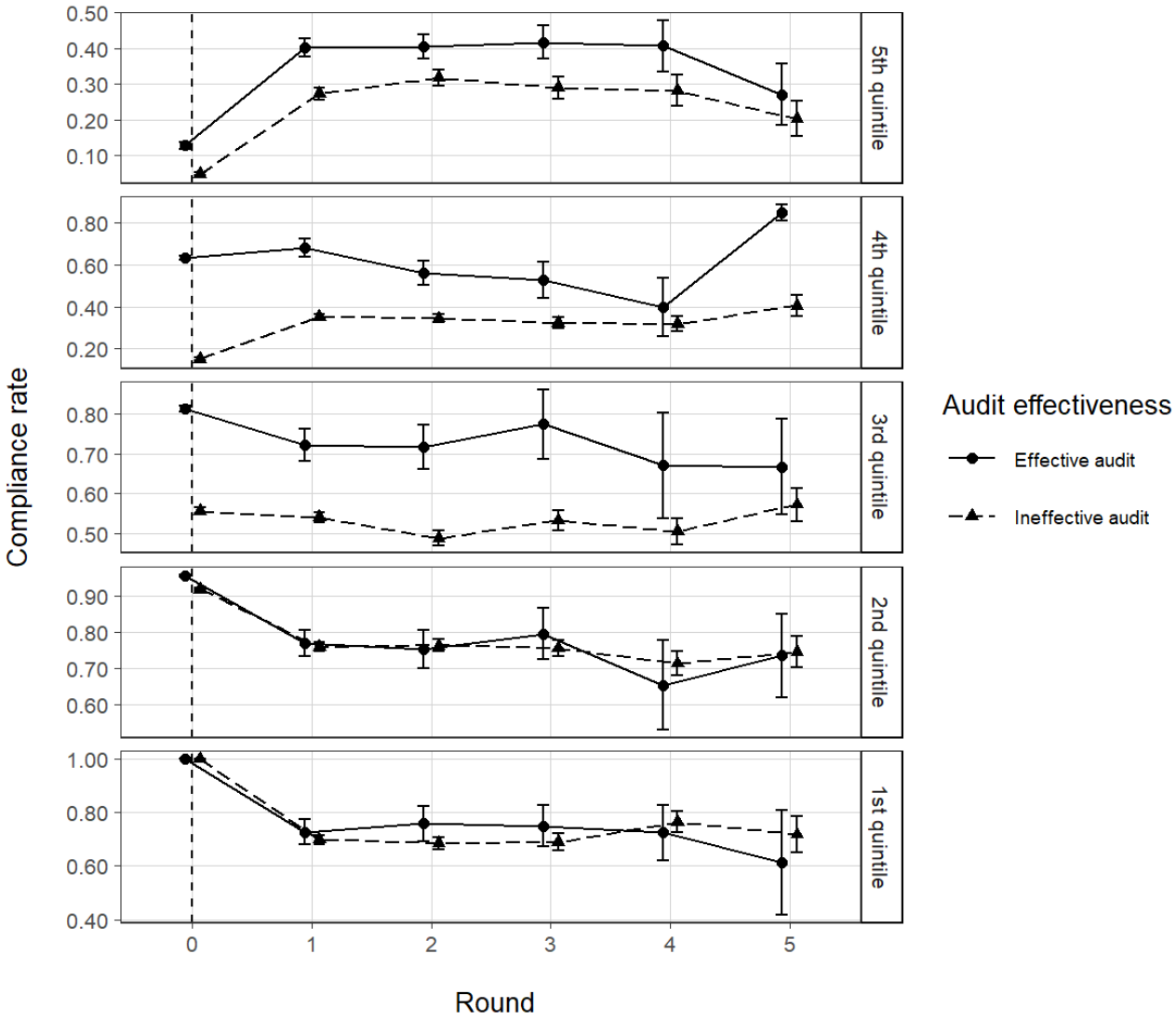
TABLE B1. Effect of Experienced Fines on Postaudit Compliance

Dependent variable: Compliance rate

Independent variable	One round after audit	Two rounds after audit	Three rounds after audit	Four rounds after audit	Five rounds after audit
	(I)	(II)	(III)	(IV)	(V)
Intercept	0.3480*** (0.0607)	0.4219*** (0.0557)	0.3302*** (0.0505)	0.3059*** (0.0477)	0.2835*** (0.0461)
Received income	0.0100*** (0.0027)	0.0190*** (0.0028)	0.0342*** (0.0028)	0.0416*** (0.0027)	0.0471*** (0.0027)
Detection risk	0.0076*** (0.0008)	0.0068*** (0.0008)	0.0047*** (0.0008)	0.0043*** (0.0008)	0.0036*** (0.0008)
Audit probability	-0.0010 (0.0006)	-0.0023*** (0.0006)	-0.0015* (0.0006)	-0.0015* (0.0006)	-0.0014* (0.0006)
Audit effectiveness	0.0015*** (0.0004)	0.0020*** (0.0004)	0.0035*** (0.0004)	0.0043*** (0.0004)	0.0047*** (0.0004)
Audit probability first	-0.0028 (0.0057)	0.0034 (0.0060)	0.0011 (0.0058)	-0.0022 (0.0057)	0.0010 (0.0056)
Round after audit	-0.1598*** (0.0120)	-0.1908*** (0.0146)	-0.2648*** (0.0174)	-0.2640*** (0.0218)	-0.1856*** (0.0282)
Experienced fine	-0.2657*** (0.0044)	-0.3054*** (0.0042)	-0.3393*** (0.0039)	-0.3577*** (0.0037)	-0.3681*** (0.0036)
Round after audit x Experienced effectiveness	0.0023*** (0.0002)	0.0025*** (0.0002)	0.0038*** (0.0003)	0.0030*** (0.0004)	0.0022*** (0.0005)
Round after audit x Experienced fine	0.0056 (0.0140)	-0.0016 (0.0171)	-0.0535** (0.0207)	-0.0528* (0.0260)	-0.0493 (0.0336)
Round after audit x Experienced effectiveness x Experienced fine	0.0019*** (0.0002)	0.0022*** (0.0002)	0.0027*** (0.0003)	0.0029*** (0.0003)	0.0028*** (0.0004)
Demographic variables	included	included	included	included	included
Observations	8,147	6,244	5,255	4,723	4,443
n	333	333	333	333	333
R^2	0.671	0.710	0.778	0.816	0.835

NOTES: *, **, and *** indicate significance at the 10%, 5%, and 1% level respectively. Robust standard errors (in parentheses) are clustered at the individual level. Continuous predictors are scaled.

FIGURE B1. Effect of Effective and Ineffective Audits Conditional on Experienced Fine



NOTES: Taxpayers were audited after declaring their income in Round 0 (dashed vertical line) and not audited again through Round 5. Panel 1 (5th quintile) comprises data from audited rounds that result in high fines (top 20 percent) as well as subsequent rounds. The mean Experienced fine is 338.30 ECU (SD = 375.27). Effective audits detect all undeclared income. Ineffective audits detect between 30 percent and 70 percent of undeclared income. Error bars represent standard errors.

TABLE B2. Effect of Audits on Evaded Income One Round After the Audit
Dependent variable: Evaded income

Independent variable	Aggregate effect			Effect of first audit	
	(VI)	(VII)	(VIII)	(IX)	(X)
Intercept	1751.0874*** (254.4807)	1976.6103*** (218.2346)	1413.2132*** (184.9973)	1621.9170*** (529.0160)	1930.4631*** (553.5839)
Received income	233.8380*** (8.5125)	223.9386*** (8.3077)	225.1960*** (7.8109)	154.4487*** (35.2222)	159.2713*** (36.3148)
Detection risk	-21.7878*** (2.4644)	-18.2884*** (2.4053)	-14.3268*** (2.2668)	-29.6249*** (10.6482)	-32.5972*** (11.0103)
Audit probability	-1.6236 (1.8072)	-1.0998 (1.7628)	-3.9564** (1.6595)	2.8631 (7.8824)	4.4694 (8.1252)
Audit effectiveness	-0.9267 (1.2483)	-0.8464 (1.2176)	-1.4162 (1.1462)	0.9793 (5.4049)	2.2718 (5.5881)
Audit probability first	9.6645 (18.1273)	10.7328 (17.6788)	11.6946 (16.6409)	19.5717 (76.8972)	-16.4343 (79.5490)
Round after audit	85.0889** (35.0975)	-117.2726*** (35.6436)	237.1545*** (34.0904)	33.7191 (63.1871)	-80.7758 (63.9287)
Round after audit x Experienced effectiveness	-1.7158*** (0.5348)	-1.7950*** (0.5216)	-0.6975 (0.4941)		
Noncompliant			1433.7787*** (29.8289)		
Round after audit x Noncompliant			-777.1469*** (35.1549)		
Compliant		-1135.5209*** (30.3969)			
Round after audit x Compliant		780.0467*** (36.0062)			
Dishonest				1545.7704*** (176.1982)	
Round after audit x Dishonest				-514.9945*** (186.8062)	
Honest					-1061.5841*** (177.3145)
Round after audit x Honest					297.3044* (172.8129)
Demographic variables	included	included	included	included	included
Observations	8,147	8,147	8,147	666	666
n	333	333	333	333	333
R^2	0.638	0.629	0.635	0.575	0.595

NOTES: *, **, and *** indicate significance at the 10%, 5%, and 1% level respectively. Robust standard errors (in parentheses) are clustered at the individual level. Continuous predictors are scaled.

TABLE B3. Effects of Audits on Compliance (Between-Subject Comparison)

Dependent variable: Compliance rate

Independent variable	(VII)	(VIII)	(IX)	(X)
Income	-0.04***	-0.05***	-0.05***	-0.05***
	(0.01)	(0.01)	(0.01)	(0.01)
Detection risk	2.14***	2.14***	2.14***	2.14***
	(0.10)	(0.10)	(0.10)	(0.10)
Audit probability first	0.01	0.01	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.02)
Audited last round	0.01	-0.06	-0.08	-0.11**
	(0.01)	(0.04)	(0.05)	(0.05)
Efficiency		-0.01	-0.03	-0.07
		(0.04)	(0.05)	(0.05)
Audited last round x Efficiency		0.14**	0.19**	0.21***
		(0.06)	(0.07)	(0.08)
Compliant			-0.01	
			(0.08)	
Audited last round x Compliant			0.05	
			(0.10)	
Efficiency x Compliant			0.09	
			(0.10)	
Audited last round x Efficiency x Compliant			-0.18	
			(0.15)	
Noncompliant				-0.22***
				(0.07)
Audited last round x Noncompliant				0.21**
				(0.10)
Efficiency x Noncompliant				0.24**
				(0.09)
Audited last round x Efficiency x Noncompliant				-0.31**
				(0.15)
Sex	0.24***	0.24***	0.24***	0.23***
	(0.08)	(0.08)	(0.08)	(0.08)
Age	0.14***	0.14***	0.14***	0.14***
	(0.05)	(0.05)	(0.05)	(0.05)
Education	-0.10	-0.10	-0.10	-0.10
	(0.07)	(0.07)	(0.07)	(0.07)
Study	-0.00	-0.00	-0.00	-0.00
	(0.04)	(0.04)	(0.04)	(0.04)
Nationality	0.15***	0.16***	0.15***	0.15***
	(0.04)	(0.04)	(0.04)	(0.04)
No prior experiments	0.08	0.10	0.10	0.09
	(0.17)	(0.17)	(0.17)	(0.17)
No prior tax experiments	-0.06	-0.06	-0.06	-0.06
	(0.08)	(0.08)	(0.08)	(0.08)
No self-preparation	0.03	0.03	0.03	0.03
	(0.08)	(0.08)	(0.08)	(0.08)
Risk-seeking	-0.02	-0.01	-0.01	-0.02
	(0.04)	(0.04)	(0.04)	(0.04)
Income maximization	-0.29***	-0.29***	-0.29***	-0.28***
	(0.04)	(0.04)	(0.04)	(0.04)
Low tax morale	-0.07*	-0.07*	-0.07*	-0.07*
	(0.04)	(0.04)	(0.04)	(0.04)
Intercept	-1.14***	-1.15***	-1.14***	-1.07***
	(0.35)	(0.35)	(0.35)	(0.35)
Observations	9324	8991	8991	8991
n	333	333	333	333
R^2	0.582	0.588	0.586	0.582

NOTES: *, **, and *** indicate significance at the 10%, 5%, and 1% level. Robust standard errors (in parentheses) are clustered at the individual level. The dependent variable and all continuous predictors are standardized.

TABLE B4. Effects of First Audits on Compliance (Between-Subject Comparison)

Dependent variable: Compliance rate

Independent variable	(XI)	(XII)
Income	-0.01 (0.02)	-0.01 (0.02)
Detection risk	2.36*** (0.26)	2.38*** (0.26)
Audit probability first	0.02 (0.04)	0.01 (0.04)
First audit last round	0.04 (0.04)	-0.02 (0.04)
Honest	0.90*** (0.15)	
First audit last round * Honest	-0.18 (0.12)	
Dishonest		-1.30*** (0.15)
First audit last round * Dishonest		0.37*** (0.13)
Sex	0.23** (0.09)	0.08 (0.09)
Age	0.05 (0.06)	0.07 (0.05)
Education	-0.02 (0.08)	-0.02 (0.07)
Study	0.01 (0.05)	0.03 (0.04)
Nationality	0.07 (0.05)	0.05 (0.05)
No prior experiments	-0.04 (0.20)	-0.08 (0.18)
No prior tax experiments	-0.10 (0.09)	-0.06 (0.09)
No self-preparation	0.00 (0.09)	-0.02 (0.09)
Risk-seeking	0.00 (0.05)	-0.02 (0.04)
Income maximization	-0.16*** (0.05)	-0.18*** (0.04)
Low tax morale	-0.06 (0.04)	-0.08* (0.04)
Intercept	-0.97** (0.40)	-0.51 (0.38)
Observations	1181	1181
n	333	333
R^2	0.689	0.671

NOTES: *, **, and *** indicate significance at the 10%, 5%, and 1% level respectively. Robust standard errors (in parentheses) are clustered at the individual level. The dependent variable and all continuous predictors are standardized.

2

New Insights on Taxpayer Behavior

Bazzoli ◆ Di Caro ◆ Figari ◆ Fiorio ◆ Manzo

An Analysis of Self-Employed Income Tax Evasion in Italy With a Consumption-Based Methodology

Martina Bazzoli, Paolo Di Caro, and Marco Manzo (Italian Ministry of Economy and Finance), Francesco Figari (University of Insubria), and Carlo Fiorio (University of Milan)

1. Introduction

The study of personal income tax evasion and individual underreporting is important, among other factors, for knowing the true income distribution in a given country and for providing more accurate evaluations of the redistributive effects of tax policies (Matsaganis e*t al.* (2010)). This is particularly relevant in countries like Italy, where tax evasion is high in comparison to other developed countries and it shows persistence across time (Schneider e*t al.* (2015)): in 2018, the Italian personal income tax (PIT) gap was equal to about 31.5 million euro, one-third of PIT revenues for the same year (Ministry of Economy and Finance (2020)). Measuring personal income tax evasion, however, is not a trouble-free task given the invisible nature of evasion activities and the need of having detailed information on individuals (Slemrod and Weber (2012)).

In some countries, such as the United Kingdom, the United States, and Denmark, the availability of administrative micro data based on random tax audits provides good information that can be used for estimating personal income tax evasion with a bottom-up approach. Alternative bottom-up techniques have been used in those countries where survey and tax data can be merged, either statistically and/or exactly through personal identification codes: discrepancies methods (Paulus (2015)), and expenditure-based analyses (Hurst e*t al.* (2014); Cabral e*t al.* (2019)). In Italy, due to the lack of random tax audits and the unavailability of tax microdata until now, income tax evasion has been mainly estimated by using the top-down approach that combines aggregate information on national accounts and tax data. In this country, bottom-up applications have been applied for research purposes (Bernasconi and Marenzi (1997); Fiorio and D'Amuri (2005)), with renewed interest in recent years (Albarea e*t al.* (2019); Lalla e*t al.* (2019)). In the next section, which contains the literature review, we discuss the added value of using bottom-up approaches for analyzing personal income tax evasion in countries such as Italy where the top-down methodology is the only one available.

In this work, for the first time for Italy, we study self-employed personal income tax evasion by applying the bottom-up approach that relies on the consumption-based methodology (Pissarides and Weber (1989)). Specifically, we build a novel dataset based on the exact matching of tax administrative microdata from individual tax declarations over the period 2010–2016 with information from the Italian Household Budget Survey (HBS) for the year 2013 that does not contain income variables. The exact matching of income and consumption data, which has been conducted by the IT Department of the Ministry of Economy and Finance (MEF) to preserve anonymity, allows us to rule out the issues that are present when adopting statistical matching techniques (Atkinson and Brandolini (2001)). Moreover, the availability of panel data regarding income covering 7 years gives us the possibility of overcoming problems related to the usage of current income in the estimation of the consumption-income curves (Engström and Hagen (2017)).

The second contribution of our study is to provide evidence on the heterogeneity of the estimates of self-employed income tax evasion in Italy. Specifically, we start by investigating the different evasion rates of the self-employed across the Italian macroareas (North, Centre, South), which is justified by the relevant territorial economic and social differences that are present in Italy, which can have consequences on the tax evasion behaviour (D'Attoma (2019)). One of the policy implications of such results is that we support possible region-specific tax compliance actions. In addition, we depart from the aggregate definition of self-employed, and we make a distinction between small entrepreneurs and liberal professionals (e.g., lawyers, doctors, accountants,

etc.). This separation can be made thanks to our administrative data that allows us to identify the particular category of self-employed under analysis. From an economic point of view, recent evidence suggests that entrepreneurs can show different characteristics (i.e., risk profile, education, etc.) than the rest of self-employed workers (Levine and Rubinstein (2017)). From a policy perspective, the knowledge of differences in tax evasion rates within the category of self-employed is important to better tailoring policies aimed at reducing tax evasion.

Our results, which are robust to alternative consumption and income variables, and remain valid after comparing Ordinary Least Squares (OLS) and Instrumental Variable (IV) estimates, suggest that the underreporting gap of self-employed households ranges from 27 percent to 35 percent. These findings are not significantly dissimilar to the results obtained by applying the same methodology to other institutional contexts such as the United States (Hurst *et al.* (2014)), and the United Kingdom (Cabral *et al.* (2019)). Interestingly, this result supports, in contrast to the popular wisdom, the recent experimental evidence suggesting that the extent of tax evasion in Italy is not so different from that registered in other countries (D'Attoma *et al.* (2017)). In addition, we find that self-employed households located in the North of the country evade more income, relative to dependent-worker households living in the same macroarea, than in the rest of the country. Also, we document that liberal professionals underreport a share of income that is about twice that underreported by small entrepreneurs.

The rest of the work is organized as follows: The next section overviews the related literature. Then, we present the data and the methodology. The fourth section contains the results. The final section concludes with some policy implications.

2. Literature Review

There are two approaches commonly used for quantifying personal income tax evasion: top-down and bottom-up. The top-down approach is used by tax administrations where good microdata are not available and/or not accessible, and relies upon aggregate comparisons between national account data, which generally include evasion, and information collected by tax authorities, based on reported income only. There are some advantages in using the top-down approach. First, it provides time-series estimates of tax evasion. Second, it allows for the separation of gross and net tax gap, the latter taking into account the effects of tax compliance policies. Third, this approach does not request the availability of and the access to microdata. Yet, the top-down method presents the following shortcomings: It is not possible to disaggregate tax gap for different categories of taxpayers; and, it does not permit the study of the distributional effects of tax evasion in detail. For a more detailed discussion and an application to Italy, see Braiotta *et al.* (2020).

The bottom-up approach uses different sources of microdata and includes three different methods. The first method uses information derived from individual tax audits for approximating true income and calculating tax evasion. This method is typically applied in countries where random audits are available (United Kingdom, United States, and Denmark), and it requests the adoption of statistical corrections (e.g., uplift factor) for extending the results obtained for the used sample to the whole population (Clotfelter (1983); Feinstein (1991); Kleven *et al.* (2011)). This bottom-up method is able to provide time-series data on tax evasion; the main shortcoming is the cost of setting up random enquiry programs where they are not available.

The second method is based on the comparison of income data deriving from individual surveys and aggregate administrative data, on the general idea that surveys provide larger aggregate taxable income than administrative data, and assuming that taxpayers declare a closer-to-true income in an anonymous interview than in tax forms (Fiorio and D'Amuri (2005); Paulus (2015); Albarea *et al.* (2019)). This method, called the discrepancy approach, relies on the assumption that survey data are without errors and/or survey errors can be managed by the researcher in order to use income declared in surveys as true income (Koijen *et al.* (2014)). Moreover, given that surveys are usually available as repeated cross-sections, this method does not allow one to provide time-series estimates of tax evasion.

The third method is based on the comparison of income and consumption data for particular categories of taxpayers. Specifically, the so-called consumption-based method (Pissarides and Weber (1989)) relies upon

the estimation of expenditure curves for different groups of taxpayers with different underreporting possibilities, such as self-employed versus dependent workers, to approximate income tax evasion by the former relative to that of the latter. This method requires using as a consumption variable a set of items that—after controlling for observable characteristics—are assumed to be independent of selected groups, such as food. This methodology was first applied in the UK (Pissarides and Weber (1989)), and later applied in several other countries (Kukk *et al.* (2020)), including the United States (Hurst *et al.* (2014)), Canada (Tedds (2010)), and Sweden (Engström and Hagen (2017)). The consumption-based method requires the availability of detailed microdata, and the solution of some empirical issues such as: i) the choice of a good measure of permanent income; ii) the selection of consumption variables that does not conditionally depend on taxpayer occupations; iii) the matching between survey and administrative data, with statistical matching producing additional noise in the estimates. Moreover, this method does not allow for the production of time-series data of tax evasion given that it is usually based on cross-section survey collection. In the next sections, we discuss the application of the consumption-based method to the Italian case, and how we dealt with the practical issues in our case.

Despite the presence of some data and methodological problems, bottom-up estimates of tax evasion have recently regained importance among researchers and policymakers given the progressive accessibility to administrative microdata (Card *et al.* (2010)). In particular, bottom-up methods allow for integrating top-down estimates in several ways, particularly in those countries like Italy where bottom-up estimates are not generally used for policymaking. First, bottom-up results are able to integrate top-down findings, by providing robustness checks to the calculations obtained by using aggregate data. Second, the adoption of bottom-up methods allows for the identification of heterogeneous profiles of tax evasion based on individual and/or household characteristics. This can be particularly helpful for profiling tax evaders and supporting the design of more tailored tax audit policies. Third, microestimates of tax evasion used in combination with tax-benefit microsimulation models are important for throwing new light on the distributional implications of underreporting activities. For a discussion on the value-added of bottom-up results applied to Italy, see MEF (2020).

3. Data and Methodology

3.1 Data description

We use a novel consumption-income dataset for a representative sample of Italian households by linking the 2013 Italian Household Budget Survey (HBS), which is provided by the Italian National Institute of Statistics (ISTAT) on a yearly basis, with data on individual tax returns, available at the MEF, for the years 2010-2016. The HBS provides detailed information on consumption expenditures, with data on about 300 consumption items, and household characteristics (number of children, education of parents, age profiles, etc.) for about 20,700 households corresponding to about 50,000 individuals (Rondinelli (2014)). Unfortunately, and differently from other countries, the Italian HBS does not contain information on household income. The main expenditure variable that we use as a dependent variable in the empirical analysis is the monetary value (in euros) of total food consumption expenditures that are recorded in the HBS on a daily basis from a diary kept by a member of the household for 2 weeks.[1]

In this study, we use administrative information deriving from individual tax returns for measuring household income. Moreover, we employ individual and household characteristics present in tax returns for having a large set of observables. Administrative data allow for the measurement of the stock of property wealth at cadastral values that we use as an additional control variable. The panel structure of fiscal data allows us to construct a measure of declared individual income from year t-3 to year t+3, where t=2013, which is the year of the HBS, providing a good proxy of permanent income over a 7-year period. This implies that our results with the adoption of the permanent income proxy rule out the issues related to asymmetric income fluctuations among taxpayer categories that are present when using a measure of current income only (Engström and Hagen (2017)).

[1] The use of food consumption as dependent variable is motivated by the fact that food expenditures are usually uncorrelated with the self-employment status of a household, holding constant all other observable characteristics (Pissarides and Weber (1989)). Other contributions used different consumption items, available in the surveys, such as home utilities and health expenditures (Albarea *et al.* (2019)). In our data, we have also information on these additional consumption expenditures. Results with different dependent variables, available upon request, confirm the main findings of our work.

Although the merge between income (administrative) and consumption (survey) data is performed at the individual level, given that in Italy tax declarations are made individually, we perform our analysis at the household level. In this study, in line with the international literature following the initial contribution of Pissarides and Weber (1989), we define self-employed households as those households whose total income from self-employment is at most equal to 25 percent of total household income. In a companion work (Bazzoli *et al.* (2020)), we defined a household as self-employed if 50 percent of its income comes from self-employment. This choice is not without implications in terms of the aggregate consequences of self-employment underreporting that are sensitive to the particular definition of self-employed (Hurst *et al.* (2014)). It is worth noticing that our classification of self-employed households allows for the detection of about 12 percent of the total sample as self-employed, a share that is close to the total share of self-employed workers in the tax records. Our results are robust to the alternative classification of self-employed households including the self-declared status in a survey.

Table 1 shows some descriptive statistics for the whole sample, the share of self-employed households defined as those earning at least 25 percent of income from self-employment and the remaining ones, defined as dependent workers, which also include pensioners. Food expenditures (in logs) are higher for self-employed than for the rest of the population, while differences in declared household income are less marked notwithstanding the definition of income that is adopted (e.g., pre- and post-tax income, current vs 7-year average). These preliminaries, which are in line with the evidence for the U.S. (Hurst *et al.* (2014)), suggest that the income-consumption relations among categories of taxpayers shall be further investigated, as we will do in the next pages. Observe that, moreover, self-employed households are younger, mostly concentrated in the North of Italy, and headed by males, in comparison to dependent workers.

3.2 Methodology

To investigate the underreporting (tax evasion) rate of self-employed households, in comparison to the income reported by dependent worker households, we use the following consumption-income relationship:

Our dependent variable is the (log of) household food consumption lnC_i, where i denotes a given household; our main income variable is the (log of) household income declared in tax returns over the years 2010–2016. We call this measure a proxy of permanent income (Engström and Hagen (2017)). We also use an alternative income variable the (log of) household income declared in tax returns in 2013, the same year of the HBS survey used in this study, in order to provide a measure of current income.

$$lnC_i = \beta lnY_i + \mathbf{X}'\alpha + \gamma SE_i + \varepsilon_i \tag{1}$$

The set of baseline controls $\mathbf{X}'$ include household head age and gender, in-couple dummy interacted with education (primary, secondary, or tertiary) of the spouse, household size, a dummy for presence of kids, family consumption of sin goods, a full set of the macroarea of residence dummies. Additional controls include also household head education and building property wealth (cadastral values). The controls, which are common in this literature (Cabral e*t al.* (2019)), are introduced to estimate the Engle curve conditional to the same individual and household characteristics for different categories of taxpayers, namely self-employed versus dependent workers. Specifically, the controls are used to rule out the influence of possible observable differences between categories of taxpayers in the investigated relationship.

The covariate SE_i is a dummy variable that takes value 1 for a given self-employed household, which we define as those households whose total income from self-employment is at least equal to 50 percent of total household income. The term ε_i is the error term of relation (1). The share of underreported income of self-employed households can be calculated as follows:

$$1 - \hat{k}_{SE} = 1 - exp\left[-\frac{\hat{\gamma}_{SE}}{\hat{\beta}}\right]. \tag{2}$$

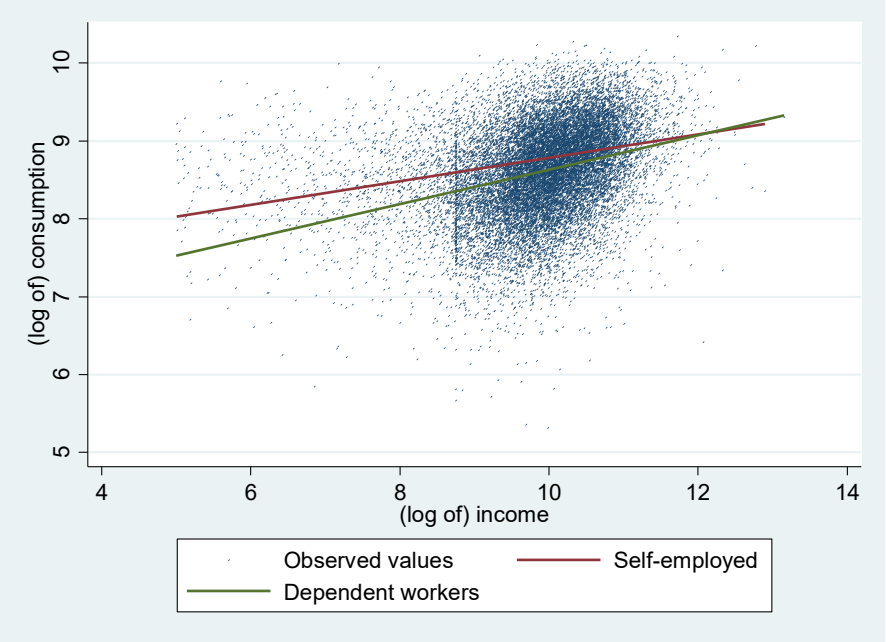
TABLE 1. Descriptive Statistics

	Whole sample	Self-employed	Dependent workers
log food expenditures	8.622	8.768	8.606
	(0.004)	(0.014)	(0.005)
Pre-tax Total Household Income:			
current (in logs)	10.088	9.974	10.101
	(0.007)	(0.030)	(0.007)
average (7-year, in logs)	10.098	10.037	10.105
	(0.006)	(0.026)	(0.006)
Post-tax Total Household Income:			
current (in logs)	9.914	9.798	9.927
	(0.006)	(0.028)	(0.006)
average (7-year, in logs)	9.920	9.848	9.929
	(0.006)	(0.024)	(0.006)
% of female-headed households	0.320	0.191	0.335
	(0.003)	(0.009)	(0.004)
% families with kids	0.264	0.307	0.259
	(0.003)	(0.011)	(0.003)
Average household size	2.377	2.852	2.323
	(0.009)	(0.032)	(0.010)
Household head: 35 and below	0.075	0.096	0.073
	(0.002)	(0.007)	(0.002)
Household head: 36-50	0.290	0.499	0.266
	(0.003)	(0.012)	(0.003)
Household head: 51-65	0.281	0.333	0.276
	(0.003)	(0.011)	(0.003)
Household head: 66 and over	0.353	0.073	0.385
	(0.004)	(0.006)	(0.004)
North	0.498	0.542	0.493
	(0.004)	(0.012)	(0.004)
Center	0.205	0.187	0.207
	(0.003)	(0.009)	(0.003)
South	0.297	0.272	0.300
	(0.003)	(0.011)	(0.004)
Sample size	18,198	1,767	16,431

NOTES: Our calculation is based on the selected sample; standard errors in parentheses. Self-employed households are identified as those with self-employment income equal to or larger than to 50 percent of total household income.

Relation (2) describes the proportion of unreported income of self-employed households ($1 - \hat{k}_{SE}$). It derives from the underlying assumption that self-employed households misreport their income, which is not third-party reported as in the case of dependent worker households, by a factor k, namely $Y_i^T = K_i Y_i^R$, with $K_i \leq 1$ where Y_i^T and Y_i^R denote true and reported income, respectively. For dependent workers, by assumption, $Y_i^T = Y_i^R$ and $k_i = 1$. Note that, in this approach, the factor k_i is assumed to be different among categories of households (i.e., self-employed vs dependent workers), but constant within the same category. In a different contribution, we relax this assumption by allowing for the possibility of having heterogeneous values for the factor k_i (Bazzoli *et al.* (2020)).

FIGURE 1. Income-Consumption Relation, Preliminary Evidence



NOTE: The graph reports the estimated values of the relation in (1) by applying the OLS estimator, when self-employed households are defined as having at most 25 percent of their total income from self-employment. The red line shows the predictions for self-employed (when the dummy SEi =1), while the green line shows the predictions for dependent workers (when the dummy SEi =0).

The graph in Figure 1 provides an illustration of the methodology that we use in this paper. It reports the values of the income-consumption relationship (dots), as estimated from the relation in (1). The red and green lines show the predicted values for self-employed and dependent workers, respectively. Two aspects are worth commenting upon. The predicted values for self-employed households are above those observed for dependent worker households, by suggesting that, for the same level of declared income, self-employed households have higher food expenditures than dependent workers. This difference, which is conditional to the same individual and household characteristics, can imply that self-employed households underreport the extent of their declared income, by denoting the presence of tax evasion. We are interested in quantifying the share of such underreporting that can be approximated by relation (2). Lastly, it is important to remember that we assume that dependent workers do not underreport their income, which can be restrictive particularly for private dependent workers (Paulus (2015)). If dependent workers can also misreport their income, our estimates of the tax evasion by self-employed households can be interpreted as a lower bound of the true level of tax evasion for such a category.

4. Results

4.1 Self-employed income tax evasion in Italy

In Tables 2 and 3, we report the estimates of the relation (1), and the estimated values of relation (2) reported in the tables as evasion rates, with the adoption of pre- and post-tax income, respectively. Using after-tax income, although subject to its own measurement issues, allows us to check to what extent fewer taxes paid by self-employed are allocated to consumption (Hurst *et al.* (2014)). We use both current and permanent income definitions in order to see how results change when smoothing income fluctuations with the adoption of the proxy of permanent income. For expositional convenience, we show the estimated coefficients of the self-employed dummy and income variables only. Estimates are obtained by clustering the errors at a provincial level for the 109 Italian provinces that describe the residence of the family.

In specifications (A-B), we use no controls, namely log consumption is regressed on a constant, the self-employment dummy, and the log of income. The specifications (C-D) include the set of controls, that is, gender and age of the household head, in-couple dummy interacted with education (primary, secondary, or college) of the partner, household size, a dummy for presence of kids, and family consumption of sin goods. The specifications (A-D) are obtained by applying OLS techniques. In the last two specifications (E-F), we apply the Instrumental Variable (IV) strategy, according to the existing literature since Pissarides and Weber (1989), where we use as an instrument the building property wealth measured using cadastral values. The IV strategy is useful for dealing with the endogeneity of current income in relation (1) and, moreover, for limiting measurement errors in the 7-year average income measure of permanent income (Engström and Hagen (2017)). The model diagnostics confirm the robustness of our findings.

Our results suggest that self-employed households consume on average more than 5 percent of what dependent worker households consume. The elasticity of consumption estimates suggest that changes in current income affect less than changes in the 7-year average income, consistently, with an interpretation of the latter as a better measure of permanent income. As for tax evasion, and when considering average income, we find that the underreporting gap of self-employed households ranges from 26 percent (specification F) to 35 percent (specification (D) when using the definition of after-tax family income (Table 3). The results are similar when using the definition of pre-tax family income, as in Table 2. Interestingly, such results are not significantly different from the findings obtained by applying the same methodology to other countries such as the United States (Hurst *et al.* (2014)), and the United Kingdom (Cabral *et al.* (2019)). In a different work (Bazzoli *et al.* (2020)), we showed that the average self-employment income tax evasion rate that we find here derives from heterogeneous underreporting shares that depend on specific individual and family characteristics (e.g., singles vs couples, age and educational levels, etc.).

TABLE 2. Self-Employment Income Tax Evasion, Pre-Tax Total Family Income

	(A)	(B)	(C)	(D)	(E)	(F)
	OLS	OLS	OLS	OLS	IV	IV
Self-employed	0.187***	0.177***	0.053***	0.055***	0.091***	0.083***
	(0.017)	(0.017)	(0.016)	(0.017)	(0.017)	(0.016)
Current income	0.197***		0.076***		0.201***	
	(0.009)		(0.008)		(0.022)	
Average income (7-yr)		0.233***		0.094***		0.201***
		(0.009)		(0.009)		(0.022)
Evasion rate	0.612***	0.534***	0.501***	0.441***	0.363***	0.340***
	(0.038)	(0.037)	(0.107)	(0.098)	(0.057)	(0.059)
Controls	No	No	Yes	Yes	Yes	Yes
R-squared	0.098	0.116	0.261	0.263	0.235	0.248
N. observations	18,198	18,198	18,198	18,198	18,198	18,198
N. obs self-employed	1,767	1,767	1,767	1,767	1,767	1,767
Share self-employed	0.775	0.775	0.775	0.775	0.775	0.775
F-stat					982.16	951.13

***Significant to the 1% level.
NOTE: Controls include household head age and gender, in-couple dummy interacted with education (primary, secondary or tertiary) of the spouse, household size, a dummy for presence of kids, family consumption of sin goods, a full set of macro area of residence dummies, household head education and building property wealth (cadastral values). Standard errors are adjusted for 109 clusters at the province of family residence.

TABLE 3. Self-Employment Income Tax Evasion, Post-Tax Total Family Income

	(A)	(B)	(C)	(D)	(E)	(F)
	OLS	OLS	OLS	OLS	IV	IV
Self-employed	0.189***	0.183***	0.052***	0.056***	0.095***	0.089***
	(0.017)	(0.017)	(0.016)	(0.017)	(0.017)	(0.017)
Current income	0.216***		0.077***		0.222***	
	(0.011)		(0.009)		(0.024)	
Average income (7-yr)		0.259***		0.099***		0.223***
		(0.010)		(0.011)		(0.024)
Evasion rate	0.584***	0.506***	0.492***	0.432***	0.348***	0.329***
	(0.038)	(0.035)	(0.107)	(0.096)	(0.053)	(0.055)
Controls	No	No	Yes	Yes	Yes	Yes
R-squared	0.098	0.119	0.26	0.262	0.234	0.247
N. observations	18,198	18,198	18,198	18,198	18,198	18,198
N. obs self-employed	1,767	1,767	1,767	1,767	1,767	1,767
Share self-employed	0.775	0.775	0.775	0.775	0.775	0.775
F-stat					950.121	921.893

***Significant to the 1% level.
NOTE: Controls include household head age and gender, in-couple dummy interacted with education (primary, secondary or tertiary) of the spouse, household size, a dummy for presence of kids, family consumption of sin goods, a full set of macro area of residence dummies, household head education and building property wealth (cadastral values). Standard errors are adjusted for 109 clusters at the province of family residence.

4.2 Regional distribution of self-employed tax evasion

In Italy, one of the most relevant dimensions of inquiry for analysing economic issues is represented by geography, given the long-lasting economic and social differences between the North and the South of the country. Such territorial differences produce several effects, including implications on inequality (Fiorio (2011); Di Caro (2017)), the distribution of evasion (Carfora e*t al.* (2018)), and the concentration of informal occupations (Di Caro and Sacchi (2020)). Understanding the region-specific patterns of self-employed tax evasion, a novelty of our contribution, is relevant because it provides further information on the concentration of evasion activities across the space (Wiseman (2013)), and, most importantly, it throws light into the regional distribution of tax revenues within the same country (González-Fernández and González-Velasco (2014)).

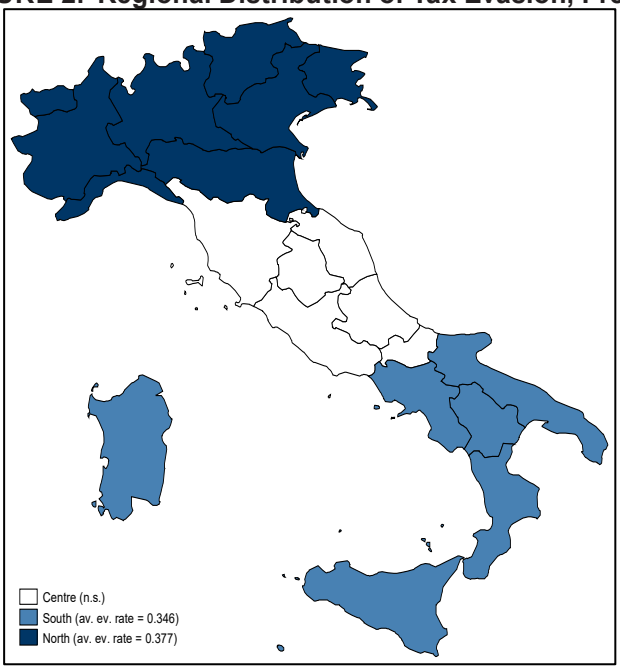
Our administrative data matched with the HBS consumption data allow for the analysis of the regional aspects of self-employed tax evasion, by providing a good sample size from a regional perspective. To keep a significant number of observations, however, we have preferred to produce estimates based on the three Italian macro-areas (North, Centre, South), which are obtained by aggregating the twenty Italian regions. In particular, we have estimated the relation (1) for each macro-area sub-sample separately. The results that we have obtained can be interpreted as the tax evasion rate of self-employed households compared to dependent workers households living in the same macro-area. In Figures 2 and 3, we report the shares of underreported income, as defined in the relation (2), for each macro-area when the income variable is pre- and post-tax household income, respectively. We have used the results obtained from the estimates of specification (F), with the IV strategy and all the control set, as discussed in the previous section. High self-employment evasion rates are marked in dark blue.

Some comments are worth discussing. We find that self-employed households underreport income relatively to dependent workers households located in the same area more in the regions located in the North (37 percent of their income) than in the rest of country. Indeed, in the South we detect a share of income underreported by self-employed equals to about 34 percent, while for the sub-sample of taxpayers located in the

[2] The lack of statistical significance for the analysis restricted to the sample of households located in the Centre can be due, among other factors, to the relatively lower number of observations in this sample (less than 18 percent of total observations). Moreover, in this macro-area, the income-consumption differences between self-employed and dependent workers households, conditional to other covariates, are very limited, possibly because there is the Lazio region where the Italian capital Rome is located and the category of dependent workers is the majority.

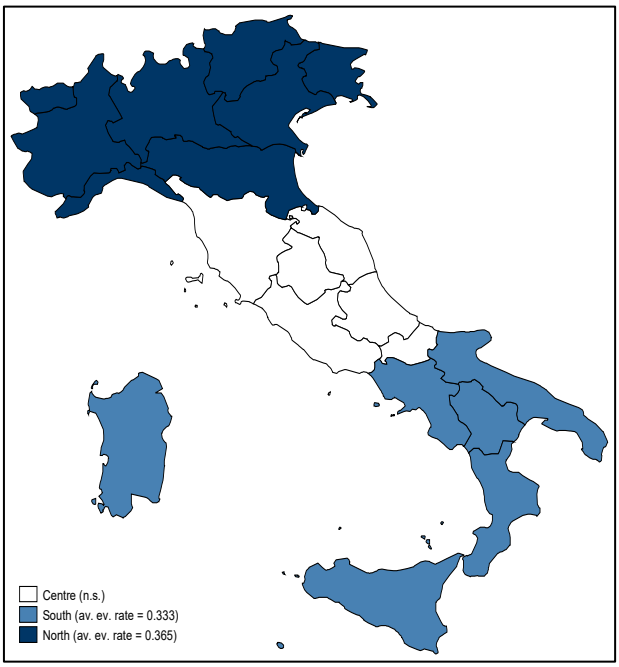
Centre we do not find significant results.[2] There are different reasons that can explain the higher evasion rates of self-employed registered in the Northern regions, which we left for future research. Note that, for instance, in this study we are not able to cover informal self-employed occupations that do not fill tax returns, which are possibly higher in the South than in the rest of the country (Di Caro and Sacchi (2020)).

FIGURE 2. Regional Distribution of Tax Evasion, Pre-Tax Income



NOTE: the graph shows the regional distribution of estimates in (1) by applying the specification (F), as in Table 2, for the sub-samples covering the three Italian macro-areas (North, Centre, South) separately; self-employed households are defined as having at least 50 percent of their total income from self-employment

FIGURE 3. Regional Distribution of Tax Evasion, Post-Tax Income



NOTE: the graph shows the regional distribution of estimates in (1) by applying the specification (F) as in Table 3 for the sub-samples covering the three Italian macro-areas (North, Centre, South), separately; self-employed households are defined as having at least 50 percent of their total income from self-employment.

4.3 Estimates for small entrepreneurs

There has been recent empirical evidence, particularly for the United States, on the fact that defining different categories of self-employed workers in a single way can produce misleading results (Levine and Rubinstein (2017)). Specifically, small and medium entrepreneurs, which are often classified as self-employed for the lack of detailed data, show significant differences in risk attitudes, organizational abilities, financial constraints and other economic and social traits, in comparison to the rest of self-employed workers (Levine and Rubinstein (2018)). Due to the lack of adequate data, to our knowledge, the consumption-income method has been applied in different countries by treating self-employed as a single category of workers. This has important policy implications since different types of self-employed can show different attitudes towards tax evasion and, most importantly, they need different tax compliance strategies. For instance, the introduction of compulsory electronic invoicing can be a good strategy for increasing tax compliance of small entrepreneurs in business-to-business (B2B) transactions, but not a sufficient tool for liberal professionals that are mostly involved in business-to-consumer (B2C) transactions. Therefore, throwing light into the different evasion profiles within the category of self-employed is necessary for guiding policymakers and, in particular, to clarify the distinction between the contrast to tax evasion in B2B transactions, which is due to omission to declare, and that in B2C transactions, which is more related to omission to invoice.

The tax return data that we use in this study gives us the possibility of making a distinction within the category of self-employed, by identifying small entrepreneurs (e.g., shop vendors, individual service firms). In this section, we have estimated the relation (1) for this category of self-employed households. The results that we have obtained can be interpreted as the tax evasion rate of small entrepreneurs households compared to dependent workers households. In Tables 4 and 5, we show the findings obtained for small entrepreneurs. Interestingly, our results, which are robust to alternative specifications and definition of the income variable, suggest that the share of income underreported by small entrepreneurs' households, relatively to dependent workers households, is lower than that registered for the entire category of self-employed households, namely 27 percent vs 34 percent. This difference, which needs further investigation on the reasons behind it, suggests the adoption of different compliance strategies with different costs for the tax administration, when trying to improve the compliance of self-employed.

TABLE 4. Income Tax Evasion, Small Entrepreneurs, Pre-Tax Total Family Income

	(A)	(B)	(C)	(D)	(E)	(F)
	OLS	OLS	OLS	OLS	IV	IV
Self-employed	0.177***	0.174***	0.034**	0.038**	0.068***	0.065***
	(0.017)	(0.017)	(0.016)	(0.016)	(0.017)	(0.017)
Current income	0.197***		(0.016)		0.203***	
	(0.010)		(0.016)		(0.021)	
Average income (7-yr)		0.233***		0.093***		0.209***
		(0.009)		(0.009)		(0.024)
Evasion rate	0.594***	0.526***	0.369***	0.336***	0.285***	0.276***
	(0.038)	(0.037)	(0.132)	(0.114)	(0.057)	(0.059)
Controls	No	No	Yes	Yes	Yes	Yes
R-squared	0.094	0.113	0.261	0.262	0.236	0.248
N. observations	18,198	18,198	18,198	18,198	18,198	18,198
N. obs self-employed	1,305	1,305	1,305	1,305	1,305	1,305
Share self-employed	0.769	0.769	0.769	0.769	0.769	0.769
F-stat					961.237	944.079

***Significant to the 1% level.

NOTE: Controls include household head age and gender, in-couple dummy interacted with education (primary, secondary or tertiary) of the spouse, household size, a dummy for presence of kids, family consumption of sin goods, a full set of macro area of residence dummies, household head education and building property wealth (cadastral values). Standard errors are adjusted for 109 clusters at the province of family residence.

TABLE 5. Income Tax Evasion, Small Entrepreneurs, Post-Tax Total Family Income

	(A)	(B)	(C)	(D)	(E)	(F)
	OLS	OLS	OLS	OLS	IV	IV
Self-employed	0.176***	0.176***	0.033**	0.039**	0.070***	0.070***
	(0.017)	(0.017)	(0.016)	(0.016)	(0.017)	(0.017)
Current income	0.215***		0.076***		0.224***	
	(0.011)		(0.009)		(0.024)	
Average income (7-yr)		0.259***		0.097***		0.224***
		(0.010)		(0.011)		(0.024)
Evasion rate	0.560***	0.494***	0.356***	0.329***	0.267***	0.269***
	(0.038)	(0.036)	(0.133)	(0.111)	(0.053)	(0.054)
Controls	No	No	Yes	Yes	Yes	Yes
R-squared	0.095	0.116	0.26	0.261	0.232	0.246
N. observations	18,198	18,198	18,198	18,198	18,198	18,198
N. obs self-employed	1,305	1,305	1,305	1,305	1,305	1,305
Share self-employed	0.769	0.769	0.769	0.769	0.769	0.769
F-stat					932.401	915.656

***Significant to the 1% level.
NOTE: Controls include household head age and gender, in-couple dummy interacted with education (primary, secondary or tertiary) of the spouse, household size, a dummy for presence of kids, family consumption of sin goods, a full set of macro area of residence dummies, household head education and building property wealth (cadastral values). Standard errors are adjusted for 109 clusters at the province of family residence.

5. Concluding remarks

This study, which is part of joint a research project between the Department of Finance of the Italian Ministry of Economy and Finance, the Universities of Milan and Insubria, and the research institution FBK-IRVAPP started two years ago, provided novel evidence on the self-employed income tax evasion in Italy. We have applied a consolidated methodology based on consumption-income comparisons between categories of taxpayers to new microdata that combines information on tax returns and consumption survey. The main results of the work can be listed as follows. First, we document that the share of self-employed income tax evasion in Italy, ranging from 30 to 40 percent of total income, is not dissimilar to that observed in different countries (United States, United Kingdom) where the same methodology has been applied. This confirm the recent view that Italy is not so exceptional internationally regarding tax evasion (D'Attoma e*t al.* (2017)). Second, we find that self-employed households located in the North of the country evade more income, about 3 percent higher, than in the rest of the country. Contrary to the popular wisdom that indicates Southern taxpayers as more evaders, we have discussed some of the possible explanations behind this result. Third, our findings point out that there are different attitudes towards tax evasion within the category of self-employed, with small entrepreneurs underreporting a lower share of income than the rest of self-employed households.

There are some policy implications that can be derived from our results. Bottom-up approaches for estimating tax evasion can be very useful instruments for complementing tax gap estimates obtained with top-down methodologies. Since two years, in Italy, in the official report on tax evasion both top-down and bottom-results regarding self-employment income tax evasion are published (MEF (2020)). In the presence of territorial differences in tax evasion behavior, as we have documented in this work, it is useful to adopt place-specific tax compliance actions in order to make the action of the tax administration more effective. Lastly, the fact that specific types of self-employed (small entrepreneurs) evade less than others highlights the importance of designing tax compliance policies, which have different costs for the administration, for particular categories of taxpayers.

References

Albarea, A., Bernasconi, M., Marenzi, A. & Rizzi, D. (2019). Income underreporting and tax evasion in Italy: Estimates and distributional effects. *Review of Income and Wealth.*

Atkinson, A. B. & Brandolini, A. (2001). Promise and pitfalls in the use of "secondary" datasets: Income inequality in OECD countries as a case study. *Journal of Economic Literature,* 39, 771–799.

Bazzoli, M., Di Caro, P., Figari,F., Fiorio, C.V. & Manzo, M. (2020). Size, heterogeneity and distributional effects of self-employment income tax evasion in Italy. Dipartimento Finanze Working Paper 8/2020, available at Dipartimento Finanze—Working Papers.

Bernasconi, M. & Marenzi, A. (1997). Gli effetti redistributivi dell'evasione fiscale in Italia. Universita degli studi di Pavia.

Braiotta, A., Carfora, A., Vega Pansini, R. & Pisani, S. (2020). Tax gap and redistributive aspects across Italy. *Scienze Regionali,* 19(2), 205–226.

Cabral, A. C. G., Kotsogiannis, C. & Myles, G. (2019). Self-employment income gap in Great Britain: How much and who? *CESifo Economic Studies,* 65(1), 84–107.

Card D., Chetty R., Feldstein, M. S. & Saez, E. (2010). Expanding access to administrative data for research in the United States. Mimeo.

Carfora, A., Pansini, R. V. & Pisani, S. (2018). Regional tax evasion and audit enforcement. *Regional Studies,* 52(3), 362–373.

Clotfelter, C. T. (1983). Tax evasion and tax rates: An analysis of individual returns. *The Review of Economics and Statistics,* 363–373.

D'Attoma, J. (2019). What explains the North–South divide in Italian tax compliance? An experimental analysis. *Acta Politica,* 54(1), 104–123.

D'Attoma, J., Volintiru, C. & Steinmo, S. (2017). Willing to share? Tax compliance and gender in Europe and America. *Research & Politics,* 4(2), 2053168017707151.

Di Caro, P. (2017). The contribution of tax statistics for analysing regional income disparities in Italy. *Journal of Income Distribution,* 25(1), 1–27.

Di Caro, P. & Sacchi, A. (2020). The heterogeneous effects of labor informality on VAT revenues: Evidence on a developed country. *Journal of Macroeconomics*, 63, 103190.

Engström P. & Hagen, J. (2017). Income underreporting among the self-employed: A permanent income approach. *European Economic Review,* 92, 92–109.

Feinstein, J. S. (1991). An econometric analysis of income tax evasion and its detection. *The RAND Journal of Economics,* 14–35.

Fiorio, C. V. (2011). Understanding Italian inequality trends. *Oxford Bulletin of Economics and Statistics,* 73(2):255–275.

Fiorio, C. V. & D'Amuri, F (2005). "Workers' tax evasion in Italy." *Giornale Degli Economisti e Annali di Economia:* 247–270.

González-Fernández, M. & González-Velasco, C. (2014). Shadow economy, corruption and public debt in Spain. *Journal of Policy Modeling,* 36(6), 1101–1117.

Hurst, E., Li, G. & Pugsley, B. (2014). Are household surveys like tax forms? Evidence from income underreporting of the self-employed. *Review of Economics and Statistics,* 96(1), 19–33.

Kleven, H. J., Knudsen, M. B., Kreiner, C. T., Pedersen, S. & Saez, E. (2011). Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica,* 79(3), 651–692.

Koijen, R., Van Nieuwerburgh, S. & Vestman R. (2014). Judging the quality of survey data by comparison with "truth" as measured by administrative records: Evidence from Sweden. In: I*mproving the Measurement of Consumer Expenditures.* 308–346. University of Chicago Press.

Kukk, M., Paulus, A. & Staehr, K. (2020). Cheating in Europe: underreporting of self-employment income in comparative perspective. *International Tax and Public Finance,* 27(2), 363–390.

Lalla, M., Mantovani, D. & Frederic, P. (2019). Measurement errors and tax evasion in annual incomes: Evidence from survey data matched with fiscal data.

Levine, R. & Rubinstein, Y. (2017). Smart and illicit: Who becomes an entrepreneur, and do they earn more? *The Quarterly Journal of Economics*, 132(2), 963–1018.

Levine, R. & Rubinstein, Y. (2018). Selection into entrepreneurship and self-employment (No. w25350). National Bureau of Economic Research.

Matsaganis, M., Benedek, D., Flevotomou, M., Lelkes, O., Mantovani, D. & Nienadowska, S. (2010). Distributional implications of income tax evasion in Greece, Hungary and Italy. Mimeo.

Ministry of Economy and Finance (2020). Relazione sull'economia non osservata e sull'evasione fiscale e contributiva. Rome.

Paulus, A. (2015). Tax evasion and measurement error: an econometric analysis of income survey data linked with tax records. ISER Working Paper 2015-10, University of Essex.

Pissarides, C.A. & Weber, G. (1989). An expenditure-based estimate of Britain's black economy. *Journal of Public Economics,* 39, 17(32).

Rondinelli, C. (2014). On the structure of Italian households: Consumption patterns during the recent crises. *Politica Economica,* 30(2–3), 235–260.

Schneider, F., Raczkowski, K. & Mróz, B. (2015). Shadow economy and tax evasion in the EU. *Journal of Money Laundering Control,* 18(1), 34–51.

Slemrod, J. & Weber, C. (2012). Evidence of the invisible: Toward a credibility revolution in the empirical analysis of tax evasion and the informal economy. *International Tax and Public Finance,* 19(1), 25–53.

Tedds, L. M. (2010). Keeping it off the books: An empirical investigation of firms that engage in tax evasion. *Applied Economics,* 42(19), 2459–2473.

Wiseman, T. (2013). U.S. shadow economies: a state-level study. *Constitutional Political Economy,* 24(4),310–335.

3

Advances in Taxpayer Service

Goldin ♦ Homonoff ♦ Javaid ♦ Schafer

Herlache ♦ Orlett ♦ Roy ♦ Turk

Scollan ♦ ten Brink

Filing Season 2019 Outreach Experiments on Paper Filers and Nonfilers

Jacob Goldin (Stanford University), Tatiana Homonoff (New York University), and Rizwan Javaid and Brenda Schafer (IRS, Research, Applied Analytics, and Statistics)

With approximately 4 percent of taxpayers filing their returns on paper and almost 12 percent not filing a return at all (nonfilers), there is an opportunity to reach millions of people with information about the benefits of preparing their return using a free assisted tax preparation method.[1] According to IRS tax return data, 15 percent of returns prepared by taxpayers using paper forms (paper filers) contained math errors, almost 30 times more than returns prepared by taxpayers using tax preparation software.[2] Software-prepared returns can be filed electronically, which helps taxpayers receive their refunds faster. Electronic filing also helps the IRS. The IRS spends approximately $0.20 processing an electronic return as opposed to almost $5.50 for a paper return.[3]

During Filing Season 2017, the IRS conducted a postcard outreach experiment on prior paper filers to encourage the use of free assisted tax preparation methods. Information about in-person Voluntary Income Tax Assistance (VITA) sites and free online assistance (through Free File or MyFreeTaxes) was provided in the outreach. The outreach resulted in a significant increase in the use of assisted tax preparation methods, particularly the use of VITA sites, which were more frequently used when addresses to the nearest sites were provided to the taxpayer. A followup study was designed for Filing Season 2019 using these results.

For the Filing Season 2019 outreach, the IRS sent letters with information about VITA, Free File, or both programs to a statistically random sample of prior-year paper filers and prior-year nonfilers. In this experiment, all treatments with VITA information included site addresses. For communications that included information on Free File, half of the treatments provided a link for the general information for Free File, while the other half provided a link to the Free File Wizard. The Free File Wizard asks taxpayers about their tax situations and then provides a list of free online software that best suits their needs. The results from this experiment yielded significant increases in overall VITA usage and Free File usage, while significantly decreasing nonfiling.

This paper discusses the results from the Filing Season 2019 outreach and how it was designed, including methodological changes based on lessons learned from the previous Filing Season 2017 experiment. We examine how filing rates and preparation methods of both paper filers and nonfilers were impacted for their TY2018 return. We also discuss how the use of VITA and Free File was impacted by single messaging and paired messaging about both options. Additionally, we explore the treatment effects when age and income groups are broken out.

Background

As technology continues to shape the world, more taxpayers are filing their returns online, whether it is one that they have completed independently, or one completed through a tax preparer. Less than 30 years ago, only about 8 percent of taxpayers used software to prepare their returns; but, at that time, there was no option to file online (Guyton *et al.* (2005)). By 2007, as the Internet became more accessible for Americans, taxpayers were

[1] Internal Revenue Service. Compliance Data Warehouse. Individual Return Transaction File. TY2017 Returns. Data Extracted March 2019.

[2] Internal Revenue Service. Compliance Data Warehouse. Individual Return Transaction File. TY2015 Returns. Data Extracted January 2017.

[3] Internal Revenue Service. 2020. Document 6746: *Cost Estimate Reference* FY2019. http://publish.no.irs.gov/catlg.html.

filing Federal tax returns electronically at a rate of 61 percent (Gunter (2016)). Jumping a decade to 2017, some 89 percent of taxpayers filed their returns online. Despite this shift to software preparation and electronic filing, 4 percent of taxpayers, approximately 6 million taxpayers nationwide, filed their returns on paper without using any type of tax preparation assistance.

During Filing Season 2017, the IRS Office of Research, Applied Analytics, and Statistics (RAAS) collaborated with the IRS Refundable Credits Administration (RCA) and academic partners at Stanford University and New York University to conduct an outreach experiment on taxpayers who had previously filed paper returns. The objective was to inform taxpayers about free assisted tax preparation methods that were available to them. The outreach was developed as a postcard communication and contained information about free in-person assistance and/or free online assistance. Despite key issues created by the printer, the outreach yielded significant increases in the use of assisted tax preparation. Because of these encouraging results, a new outreach was developed for Filing Season 2019.

The Filing Season 2019 outreach approach follows its predecessor closely with two key differences. In addition to including prior-year paper filers, this outreach also includes taxpayers who did not file a return the previous year. The other difference was to mail the taxpayer a letter in a sealed envelope, rather than send a postcard. Previous studies have found that letters from the IRS are more effective in increasing the number of responses, compared to postcards (Orlett *et al.* (2017)). The free in-person and online assistance methods that were used in the communication are outlined below.

The Voluntary Income Tax Assistance (VITA) program is an in-person tax assistance program providing free tax help from IRS-certified volunteers for taxpayers whose adjusted gross income (AGI) was generally less than or equal to $55,000 for Tax Year 2018 (TY2018). VITA offers free basic income tax return preparation, which includes assisting with filing a return for W-2, various Form-1099s, Earned Income Tax Credit (EITC), Child Tax Credit, and Affordable Care Act statements. A full list of included services can be found in IRS Publication 3676-B. Select sites with a "Self-Prep" capability also allow taxpayers to use free Web-based tax preparation software to prepare and electronically file their returns themselves.

Another free tax preparation method is software provided through the Free File (FF) program, a partnership between the IRS and the Free File Alliance. Taxpayers whose AGI was less than or equal to $66,000 for TY2018 could use this software. The Free File Website contains a wizard tool that allows taxpayers to answer questions about their tax situations and then generates a list of commercial online software programs that they can use for free. This includes software provided by companies like TurboTax, H&R Block, TaxAct, and TaxSlayer, among others. These companies allow taxpayers to use their tax preparation software to prepare and file their income taxes electronically.

There are many potential taxpayer benefits to using assisted tax preparation software, including guidance on tax benefits, knowledge of tax laws, and a platform to perform step-by-step calculations. Taxpayers using assistance are also over 30 times less likely to make a math error than those filing on paper unassisted (Javaid *et al.* (2018)). These math errors are typically computational errors and could have been avoided by using an assisted method. Additionally, research suggests that the introduction of electronic filing significantly increased the take-up of the EITC (Kopczuk and Pop-Eleches (2007)). By filing tax returns electronically, most eligible taxpayers receive their refunds at least a week earlier than those filing on paper. Electronic filing also reduces the cost on the IRS, as it costs approximately $5.50 to process a paper return as opposed to only about $0.20 to process an electronic return. By extrapolating this on all the current paper filers, the IRS can potentially save millions of dollars in return processing costs, money that could be reallocated to fund other work.

Related Research

This study is a followup to the outreach experiment that was conducted during Filing Season 2017 by Javaid *et al.* (2018). In that study, a statistical sample of taxpayers who filed a self-prepared return on paper was selected to receive a postcard with information about free assisted tax preparation methods. These methods included VITA and Free File, in addition to another free software called MyFreeTaxes that is offered through the United Way. Taxpayers in the treatment group received one postcard with information about one or more

of the free tax preparation methods. Addresses of the two VITA sites nearest the address reported on the taxpayer's prior-year return, were also included on two of the five treatments. Only one mailing was planned for this experiment, but because of printer issues, a second mailing was conducted to gather additional data. Approximately 640,000 taxpayers were statistically selected to receive the outreach mailings (treatment group) and approximately 1.4 million taxpayers served as a control group. The study found that taxpayers who were sent the outreach postcard were 20 percent more likely to use VITA. They were also 4 percent more likely to use any tax preparation software. Additionally, the use of a paid preparer significantly decreased and overall filing rates significantly increased. The Filing Season 2019 study built on the research from the 2017 study.

Prior research similar to the Filing Season 2019 study presented in this paper include a field experiment conducted by Guyton *et al.* (2017) that examined the effects of sending outreach mailings to nonfilers who appeared to be eligible for the Earned Income Tax Credit (EITC). The first part of that outreach was conducted in Filing Season 2014 and involved persons who had not filed a TY2011 or TY2012 tax return. In those studies, a statistical sample of individuals was sent up to two mailings of either postcards, brochures, or both at different times in the filing season. The results showed that sending any outreach mailing increased filing rates by between one-half of 1 percent to 1 percent. Taxpayers in the treatment group also had a higher rate of EITC claims and were more likely to file returns for multiple tax years when filing.

Another key outreach study by Orlett *et al.* (2017) was conducted in Filing Season 2016 to encourage nonfilers to file their TY2015 returns. The study population consisted of taxpayers who had resolved a nonfiler case through the Automated Substitute for Return (ASFR) process in Calendar Year 2015. Taxpayers in the outreach group were mailed a reminder to file via a postcard or letter, while the control group received no correspondence. Both forms of reminders resulted in significantly higher filing rates, though the letter was more effective.

These prior studies show that encouraging certain tax filing behaviors can be done effectively and economically with outreach interventions. This current study on paper filers and nonfilers extends the research discussed above.

Methodology

In Filing Season 2019, we conducted an outreach study using a randomized control trial design to estimate the impact of sending taxpayers letters with information about VITA and Free File on their filing rates and choice of return preparation methods.

For this study, the following two study subpopulations were created: 1) taxpayers who had filed a paper return without the assistance of a preparer in TY2017, and 2) a 10-percent sample of taxpayers who did not file a return (nonfilers) for TY2017. The adjusted gross income (AGI) for paper filers was based on the taxpayer's TY2017 tax return and the AGI for nonfilers was based on third-party information returns filed for the taxpayer in TY2017.

The additional criteria below were also applied to the sample:

- An AGI greater than $0,
- An AGI of $55,000 or less,
- Lived within 30 miles of two VITA sites, and
- Had not received a Filing Season 2017 treatment postcard.

The initial study population for the paper filers was 1,147,353 taxpayers. Additional characteristics of this group include a median age of 43 and a median household income of $26,780 for TY2017. The prior nonfiler group had an initial study population of 2,033,941 taxpayers with a median age of 31, and median household income of $6,308 for TY2017.

Study Design

The study consisted of one mailing in March 2018 to each of five outreach treatment groups of approximately 25,000 taxpayers split between the subgroups of prior paper filers and prior nonfilers. The letters included information about VITA and Free File. All treatments that included information about VITA also included addresses to the taxpayer's two nearest VITA sites. In addition to providing general information about Free File, half of the treatments included a link to the IRS Free File homepage and the other half included a link to the wizard tool. The Free File Wizard tool allows taxpayers to answer questions about their tax situation and generates a list of free software they are eligible to use.

Outreach treatment group members were sent one of the five letters described below (see Appendix A for copies of the five letters):

- ***Treatment 1:*** VITA addresses and Free File general link
- ***Treatment 2:*** VITA addresses and Free File Wizard link
- ***Treatment 3:*** VITA addresses only
- ***Treatment 4:*** Free File general link only
- ***Treatment 5:*** Free File Wizard link only

Outreach treatment subgroup members were selected through statistical stratified sampling using the following binary fields:

- ***Age:*** age 30 or under or over age 30.
- ***Income:*** $25,000 or under or over $25,000.
- ***Distance:*** At least one VITA site within 5 miles or neither site within 5 miles.
- ***Math error:*** one or more math errors on TY2017 return (for paper filer group only).
- ***Withholding:*** from TY2017 Information Returns (for nonfiler group only).

Taxpayers from the study population who were not selected into the outreach groups were used as the control group and did not receive any letter.

Implementation Issues

There were two key issues that impacted the study. The first was the Federal Government shutdown from December 22, 2018, to January 25, 2019. The shutdown prevented the planned mailing of the outreach letters in January and delayed it to March. Additionally, the shutdown created a delay in data updates for taxpayers in the study population who had already filed their return. These taxpayers could not be determined prior to mailing the outreach letters. To account for this, taxpayers who filed their return on or before March 16, 2019, were excluded from the analysis as they had already chosen their tax preparation method prior to receiving the experimental communication.

The other issue relates to the sampling were the initial sample of paper filers was not constrained to remove "V-coded" taxpayers. These are taxpayers who used software to generate a paper return. These taxpayers were removed from the study population when the analysis was conducted.

Table 1 shows the modified study population and intervention figures based on the two experimental issues.

TABLE 1. Modified Study Population, by Treatment and Filer Group

Group	Paper Filer	Nonfiler
Control	**89,560**	**1,660,817**
Treatment 1	1,039	9,434
Treatment 2	1,018	9,406
Treatment 3	1,018	9,466
Treatment 4	1,024	9,461
Treatment 5	1,056	9,411
All treated	**5,155**	**47,178**
Population	**94,715**	**1,707,995**

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data extracted April 2020.

Analysis Methodology

In this section, we will discuss the methodology for our analysis. Our outcomes of interest include whether the taxpayers filed their TY2018 return and what preparation method they used. We used a logistic regression to model these two outcomes for paper filers and nonfilers separately, since both groups have distinct characteristics. The regression for paper filers included explanatory variables based on information from TY2017 income tax returns. For the nonfilers, the explanatory indicator variables were based on third-party TY2017 information returns.

Because of the Federal Government shutdown, only taxpayers whose returns were recorded to IRS's administrative databases on or after March 17, 2019, were included in the model. Additionally, for the purposes of the preliminary analysis, all taxpayers in the outreach samples whose letter was undeliverable were also included in the model as "treated." The results presented in the next section can be adjusted by the undeliverable rate of 6.6 percent for paper filers and 30.5 percent for nonfilers to account for this factor.

Undeliverable mail was collected by a contractor and all other data was obtained from IRS administrative data files.

Preliminary Results for the Filing Season 2019 Experiment

This section will discuss the preliminary results of the Filing Season 2019 experiment. We will examine the paper filer and nonfiler groups separately as they may have had different responses. A full list of effects on filing and return preparation method by individual treatments, and groups of treatments, can be found in Appendix B for paper filers and Appendix C for nonfilers.

Primary Findings

We first examine the effect of receiving any outreach letter on return filing and tax preparation method. Table 2 shows that both paper filers and nonfilers who were sent outreach letters had higher filing rates, though the increase was statistically significant for only nonfilers. VITA usage in the control group for paper filers was 1.03 percent, but those who received treatment used VITA at a rate of 1.71 percent, a 67 percent difference.

The nonfiler group also had a significant increase when treated, with VITA usage higher than for the control group by 24 percent. The impact on the use of Free File software also yielded significant increases for paper filers in the treatment group by about 20 percent, and for nonfilers by approximately 14 percent. There was no impact of treatment on the use of a paid preparer for nonfilers; however, we saw a little under a 7-percent difference in the use of a preparer for outreach paper filers versus the control group.

TABLE 2. Effects of Any Outreach Intervention, by Item and Filer Group

Item	Paper Filer		Nonfiler	
	Control	All treated	Control	All treated
Filing rates	83.22	83.39	19.71	20.15 **
VITA usage	1.03	1.71 ***	0.53	0.66 ***
Free File usage	2.12	2.55 **	0.64	0.73 **
Paid preparer usage	8.66	8.07 *	8.85	8.96

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data Extracted July 2020.
Outcomes are indicator variables with a scale of 0-100.
Significance based on model coefficient: *90 percent significance | **95 percent significance | ***99 percent significance.

VITA Outreach Interventions

Next, we look at the impact of providing information about VITA on the likelihood of using it. In the previous TY2017 study, only some of the outreach postcards that had VITA information also included addresses to the nearest sites. Because providing addresses proved to be more effective in the usage of VITA, all treatments in this experiment included addresses of the two nearest sites to the taxpayer.

Table 3 shows that each of the VITA outreach interventions significantly increased the likelihood of using it for paper filers, with the largest effect coming from Treatment 2, which was 125 percent higher than the control group. Surprisingly, Treatment 3 had the smallest effect of the three, despite that it only provided information about VITA and no information about Free File. The treatments with VITA information were so effective on prior paper filers that it increased the likelihood of using VITA to more than twice that of the control group.

Only 0.53 percent of nonfilers in the control group used VITA, however, 0.70 percent of nonfilers who received any VITA treatment used VITA, an increase of over 31 percent. Treatment 3 had the largest effect with approximately a 59-percent increase over the control group, while Treatment 2 did not yield a statistically significant increase.

Overall, these results show that both paper filers and nonfilers benefitted from being treated by receiving letters with VITA information. By providing addresses to the nearest sites, these taxpayers may have been able to take advantage of available assistance—assistance that they may not have been aware of previously.

TABLE 3. Effects of VITA Treatments, by Treatment and Filer Group

Treatment	Paper Filer - VITA usage		Nonfiler - VITA usage	
	Control	Treatment	Control	Treatment
T1: VITA + Free File General	1.03	2.14 ***	0.53	0.71 **
T2: VITA + Free File Wizard		2.31 ***		0.55
T3: VITA only		1.98 **		0.85 ***
VITA Treatments: T1, T2, T3		2.14 **		0.70 ***

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data Extracted July 2020.
Outcomes are indicator variables with a scale of 0-100.
Significance based on model coefficient: *90 percent significance | **95 percent significance | ***99 percent significance.

Free File Outreach Interventions

All of the treatments, except Treatment 3, included information about Free File. Table 4 examines the effect of each treatment on the usage of Free File. Paper filers treated with Treatment 1 resulted in a significant increase of approximately 48 percent in the use of Free File over the control group. Treatment 4 also had a significant increase, but by a lower magnitude of about 35 percent. Treatments 2 and 5, which both included the Free File Wizard link, produced no significant effects for paper filers.

Unlike the paper filers, none of the individual treatments with Free File information produced significant effects for the nonfilers.

TABLE 4. Effects of Free File Treatments, by Treatment and Filer Group

Treatment	Paper Filer - Free File usage		Nonfiler - Free File usage	
	Control	Treatment	Control	Treatment
T1: VITA + Free File general	2.12	3.13 **	0.64	0.76
T2: VITA + Free File Wizard		2.14		0.67
T4: Free File general only		2.87 *		0.75
T5: Free File Wizard only		2.72		0.78

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data Extracted July 2020.
Outcomes are indicator variables with a scale of 0-100.
Significance based on model coefficient: *90 percent significance | **95 percent significance | ***99 percent significance.

To test for differences in treatment effects between letters that included a general Free File link versus those that directed taxpayers to the Free File Wizard, we estimated the differences between a combined Free File general link group (Treatments 1 and 4) and a Free File Wizard link group (Treatments 2 and 5). Table 5 shows that only the combined Free File general link intervention was effective in producing a significant response for both paper filers and nonfilers. The prior paper filer group exhibited a difference of about 41 percent, while the nonfilers difference in the use of Free file was approximately 17 percent higher in the intervention group than in the control group.

TABLE 5. Impact of Free File General and Wizard Links, by Treatment Category and Filer Group

Treatment category	Paper filer - Free File usage		Nonfiler - Free File usage	
	Control	Treatment	Control	Treatment
Free File general treatments: T1, T4	2.12	3.00 ***	0.64	0.75 *
Free File Wizard treatments: T2, T5		2.52		0.72

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data Extracted July 2020.
Outcomes are indicator variables with a scale of 0-100.
Significance based on model coefficient: *90 percent significance | **95 percent significance | ***99 percent significance.

Age Analysis

The Filing Season 2017 study showed that younger taxpayers (age 30 or under) were more likely to increase their use of software in response to receiving a treatment postcard while older taxpayers (over age 30) were more likely to increase their use of VITA. In Table 6 of our analysis for the Filing Season 2019 study, we will consider the two age groups and its impact on the usage of VITA in response to the interventions.

Table 6 shows that treatment among younger filers who filed prior paper returns can significantly increase the use of both VITA and Free File when compared to the respective control group. The use of VITA increased by approximately 78 percent and the use of Free File increased by about 27 percent. There was no significant impact on the use of either assisted preparation method among the older filers who filed prior paper returns.

The table also shows that treating prior nonfilers in either age group resulted in a significant increase in the use of VITA. Only the nonfilers over 30 years old were yielded a significant increase in the likelihood of using Free File when treated.

TABLE 6. Impact of Age on Taxpayer Using VITA and Free File, by Filer Group

Age	Paper Filer -VITA usage		Nonfiler - VITA usage	
	Control	Treatment	Control	Treatment
Age 30 or under	1.20	2.13 ***	0.61	0.74 ***
Over age 30	0.46	0.36	0.43	0.56 ***
Age	**Paper Filer - Free File usage**		**Nonfiler - Free File usage**	
	Control	**Treatment**	**Control**	**Treatment**
Age 30 or under	1.59	2.02 **	0.41	0.43
Over age 30	3.80	4.26	0.98	1.16 **

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data Extracted July 2020.
Outcomes are indicator variables with a scale of 0-100.
Significance based on model coefficient: *90 percent significance | **95 percent significance | ***99 percent significance.

Income Analysis

We stratified the study design on income using an income threshold of $25,000 to ensure that we were also able to analyze the impact of income level on the use of a free assisted tax preparation method. Table 7 shows that treating prior paper filers in either income group resulted in significant increases in the use of VITA, with the higher income group resulting in the larger magnitude of increase of approximately 83 percent. For the prior nonfiler group that was treated, there were only significant increases in the use of either VITA or Free File for those in the $25,000 or under income category.

TABLE 7. Impact of Income on Taxpayers Using VITA and Free File, by Filer Group

Income	Paper Filer - VITA usage		Nonfilers - VITA usage	
	Control	Treatment	Control	Treatment
$25,000 or under	1.11	1.71 ***	0.54	0.68 ***
Over $25,000	0.94	1.72 ***	0.53	0.60
Income	**Paper Filer - Free File Usage**		**Nonfilers - Free File usage**	
	Control	**Treatment**	**Control**	**Treatment**
$25,000 or under	2.62	3.05	0.69	0.78 **
Over $25,000	1.61	2.03 *	0.49	0.55

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data Extracted July 2020.
Outcomes are indicator variables with a scale of 0-100.
Significance based on model coefficient: *90 percent significance | **95 percent significance | ***99 percent significance.

Conclusions and Future Research

Based on the overall results of this outreach, we can conclude that providing taxpayers with information about both VITA and Free File was effective in not only increasing filing rates, but also nudging prior paper filers and prior nonfilers to use the two free assisted tax preparation methods. With VITA sites continuing to operate well under capacity, there is tremendous potential to scale up the number of taxpayers that it currently serves. For Free File, there are presently no limitations on scalability, considering that it operates electronically. We will continue our research on this topic by analyzing the data from the Filing Season 2020 experiment. A larger sample size and earlier mailing date should boost the effectiveness of the treatment, though it is unknown what impact the COVID-19 pandemic will have on our results.

References

Gunter, Samara. 2016. "Your Biggest Refund, Guaranteed? Internet Access, Tax Filing Method, and Reported Tax Liability." Waterville, ME: Colby College, Department of Economics.

Guyton, John L., Adam K. Korobow, Peter S. Lee, and Eric J. Toder. 2005. "The Effects of Tax Software and Paid Preparers on Compliance Costs." *National Tax Journal* 58(3).

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 Research." *American Economic Review:* Papers and Proceedings 107(5), 471–475.

Javaid, Rizwan, Brenda Schafer, Jacob Goldin, Tatiana Homonoff, and Adam Isen. 2018. "Can IRS Move Paper Filers to Assisted Tax Preparation?" 2018 IRS Research Bulletin, *8th Annual Joint Research Conference on Tax Administration.*

Kopczuk, Wojciech, and Cristian Pop-Eleches. 2007. "Electronic Filing, Tax Preparers, and Participation in the Earned Income Tax Credit." *Journal of Public Econometrics* 91(7–8), 1351–1367.

Orlett, Stacy, Rizwan Javaid, Vicki Koranda, Maryamm Muzikir, and Alex Turk. 2017. "Impact of Filing Reminder Outreach on Voluntary Filing Compliance for Taxpayers with a Prior Filing Delinquency." *2017 IRS Research Bulletin, 7th Annual Joint Research Conference on Tax Administration,* 83–98.

Appendix A: Samples of Treatments

Address Block of Letter (L6168, L6169, L6170, L6171, L6172)

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

Letter: 6168
Date: [DATE]

[BARCODE] [RECID] [NDC CODE]
[TAXPAYER NAME]
[ADDRESS LINE 1] [ADDRESS LINE 2]
[CITY], [STATE] [ZIP]

TREATMENT 1: VITA Addresses and Free File General Info (L6168)

According to our records, you may qualify for free tax preparation

What you need to know

Two out of three taxpayers qualify for free in-person or online tax preparation through an IRS-sponsored program.

Benefits you may receive from assisted tax preparation:

- Getting your refund in as few as three business days.
- Access to free commercial software for federal and state returns.
- Less chance of making a mistake on your tax return or missing a tax benefit.

Read below for information about these free IRS-sponsored programs.

VITA/TCE programs

- The Volunteer Income Tax Assistance (VITA) and Tax Counseling for the Elderly (TCE) programs provide free in-person tax preparation assistance by IRS-certified volunteers, regardless of a taxpayer's age.
- Most taxpayers qualify if they earned $55,000 or less in 2018.
- Help is available near you. Call for hours of operation:

Cranes Mill Retirement Comm…
459 Passaic Ave
West Caldwell, NJ 07006
(973) 372-2077

GPMBC - Montclair United Wa…
60 S Fullerton Ave
Montclair, NJ 07042
(800) 906-9887

- Be sure to bring photo identification, a copy of your last year's return, Social Security cards, and your tax documents (e.g., Forms W-2 and 1099-MISC).
- **For more information, visit www.irs.gov/VITA or call 800-906-9887.**

Free File program

- Free File provides free commercial software to help prepare your return online.
- Most taxpayers qualify if they earned $66,000 or less in 2018.
- You will need your 2017 tax return, 2018 tax documents, and a valid email address to begin.
- **For more information, visit www.irs.gov/FreeFile.**

Frequently asked questions

- If you have questions about this letter, you can call 855-421-8641 (toll-free).
- You don't need to respond to this letter.

Letter 6168 (02-2019)
Catalog Number 72135K

TREATMENT 2: VITA Addresses and Free File Wizard Info (L6169)

Cranes Mill Retirement Comm…
459 Passaic Ave
West Caldwell, NJ 07006
(973) 372-2077

GPMBC - Montclair United Wa…
60 S Fullerton Ave
Montclair, NJ 07042
(800) 906-9887

According to our records, you may qualify for free tax preparation

What you need to know

Two out of three taxpayers qualify for free in-person or online tax preparation through an IRS-sponsored program.

Benefits you may receive from assisted tax preparation:

- Getting your refund in as few as three business days.
- Access to free commercial software for federal and state returns.
- Less chance of making a mistake on your tax return or missing a tax benefit.

Read below for information about free IRS-sponsored programs.

VITA/ TCE programs

- The Volunteer Income Tax Assistance (VITA) and Tax Counseling for the Elderly (TCE) programs provide free in-person tax preparation assistance by IRS-certified volunteers, regardless of a taxpayer's age.
- Most taxpayers qualify if they earned $55,000 or less in 2018.
- Help is available near you. Call for hours of operation:

AARP Evergreen Library	AARP Golden Library (Ad-Hoc…
5000 Highway 73	1019 10th St
Evergreen, CO 80439	Golden, CO 80401
(303) 235-5275	(800) 906-9887

- Be sure to bring photo identification, a copy of your last year's return, Social Security cards, and your tax documents (e.g., Forms W-2 and 1099-MISC).
- **For more information, visit www.irs.gov/VITA or call 800-906-9887.**

Free File program

- Free File provides free commercial software to help prepare your return online.
- Most taxpayers qualify if they earned $66,000 or less in 2018.
- You will need your 2017 tax return, 2018 tax documents, and a valid email address to begin.
- **For more information, visit www.irs.gov/WizardFreeFile.**

Frequently asked questions

- If you have questions about this letter, you can call 855-421-8641 (toll-free).
- You don't need to respond to this letter.

Letter 6169 (02-2019)
Catalog Number 72136V

TREATMENT 3: VITA Addresses Only (L6170)

Cranes Mill Retirement Comm…
459 Passaic Ave
West Caldwell, NJ 07006
(973) 372-2077

GPMBC - Montclair United Wa…
60 S Fullerton Ave
Montclair, NJ 07042
(800) 906-9887

According to our records, you may qualify for free tax preparation

What you need to know

Two out of three taxpayers qualify for free in-person tax preparation through an IRS-sponsored program.

Benefits you may receive from assisted tax preparation:

- Getting your refund in as few as three business days.
- Access to free commercial software for federal and state returns.
- Less chance of making a mistake on your tax return or missing a tax benefit.

Read below for information about free IRS-sponsored programs.

VITA/ TCE programs

- The Volunteer Income Tax Assistance (VITA) and Tax Counseling for the Elderly (TCE) programs provide free in-person tax preparation assistance by IRS-certified volunteers, regardless of a taxpayer's age.
- Most taxpayers qualify if they earned $55,000 or less in 2018.
- Help is available near you. Call for hours of operation:

Waukegan Community Church	Waukegan Park Senior Center
1016 Grand Ave	412 S Lewis Ave
Waukegan, IL 60085	Waukegan, IL 60085
(847) 360-1008	(847) 244-9242

- Be sure to bring photo identification, a copy of your last year's return, Social Security cards, and your tax documents (e.g., Forms W-2 and 1099-MISC).
- **For more information, visit www.irs.gov/VITA or call 800-906-9887.**

Frequently asked questions

- If you have questions about this letter, you can call 855-421-8641 (toll-free).
- You don't need to respond to this letter.

Letter 6170 (02-2019)
Catalog Number 72137G

TREATMENT 4: Free File General Info Only (L6171)

According to our records, you may qualify for free tax preparation

What you need to know	**Two out of three taxpayers may qualify for free online tax preparation through an IRS-sponsored program.** Benefits you may receive from assisted tax preparation: • Getting your refund in as few as three business days. • Access to free commercial software for federal and state returns. • Less chance of making a mistake on your tax return or missing a tax benefit. Read below for information about this free IRS-sponsored program.
Free File program	• Free File provides free commercial software to help prepare your return online. • Most taxpayers qualify if they earned $66,000 or less in 2018. • You will need your 2017 tax return, 2018 tax documents, and a valid email address to begin. • **For more information, visit www.irs.gov/FreeFile.**
Frequently asked questions	• If you have questions about this letter, you can call 855-421-8641 (toll-free). • You don't need to respond to this letter.

Letter 6171 (02-2019)
Catalog Number 72138R

TREATMENT 5: Free File Wizard Info Only (L6172)

According to our records, you may qualify for free tax preparation

What you need to know	**Two out of three taxpayers qualify for free online tax preparation through an IRS-sponsored program.** Benefits you may receive from assisted tax preparation: • Getting your refund in as few as three business days. • Access to free commercial software for federal and state returns. • Less chance of making a mistake on your tax return or missing a tax benefit. Read below for information about this free IRS-sponsored program.
Free File program	• Free File provides free commercial software to help prepare your return online. • Most taxpayers qualify if they earned $66,000 or less in 2018. • You will need your 2017 tax return, 2018 tax documents, and a valid email address to begin. • **For more information, visit www.irs.gov/WizardFreeFile.**
Frequently asked questions	• If you have questions about this letter, you can call 855-421-8641 (toll-free). • You don't need to respond to this letter.

Letter 6172 (02-2019)
Catalog Number 72139C

Appendix B. Paper Filer Raw Effects vs. Model Effects Table

Treatment	Raw Data				Model			
	Control	Treated	Change	Percent Change	Control	Treated	Change (S.E.)	Percent Change
All Paper Filers								
1: VITA Addresses and FF General Info		76.90	-0.33	-0.43%		82.14	-1.09 (1.03)	-1.30%
2: VITA Addresses and FF Wizard Info		77.80	0.57	0.74%		82.97	-0.25 (1.02)	-0.30%
3: VITA Addresses Only		80.35	3.12	4.04%		84.83	1.61 (0.96)	1.93%*
4: FF General Info Only		77.34	0.11	0.14%		83.00	-0.23 (1.00)	-0.27%
5: FF Wizard Info Only	77.23	79.83	2.60	3.37%	83.22	83.97	0.75 (0.97)	0.90%
1,2,3: All VITA Treatments		78.34	1.11	1.44%		83.32	0.10 (0.59)	0.12%
1,4: All FF Gen Treatments		77.12	-0.11	-0.14%		82.57	-0.65 (0.72)	-0.78%
2,5: All FF Wizard Treatments		78.83	1.60	2.07%		83.48	0.26 (0.71)	0.31%
All Treatments		78.45	1.22	1.58%		83.39	0.17 (0.46)	0.20%
Used VITA								
1 VITA Addresses and FF General Info		2.79	1.51	117.97%		2.14	1.11 (0.39)	108.29% ***
2 VITA Addresses and FF Wizard Info		3.05	1.77	138.28%		2.31	1.28 (0.40)	125.27% ***
3 VITA Addresses Only		2.55	1.27	99.22%		1.98	0.95 (0.37)	92.78% **
4 FF General Info Only		1.76	0.48	37.50%		1.40	0.37 (0.31)	36.49%
5 FF Wizard Info Only	1.28	0.85	-0.43	-33.59%	1.03	0.77	-0.26 (0.23)	-25.27%
1,2,3: All VITA Treatments		2.80	1.52	118.75%		2.14	1.11 (0.23)	108.68% ***
1,4: All FF Gen Treatments		2.28	1.00	78.13%		1.76	0.74 (0.06)	72.10% ***
2,5: All FF Wizard Treatments		1.93	0.65	50.78%		1.52	0.50 (0.05)	48.68% **
All Treatments		2.19	0.91	71.09%		1.71	0.69 (0.16)	66.93% ***
Used FreeFile								
1 VITA Addresses and FF General Info		4.04	1.36	50.75%		3.13	1.01 (0.47)	47.69% **
2 VITA Addresses and FF Wizard Info		3.04	0.36	13.43%		2.14	0.02 (0.41)	0.89%
3 VITA Addresses Only		2.36	-0.32	-11.94%		1.69	-0.43 (0.35)	-20.14%
4 FF General Info Only		3.61	0.93	34.70%		2.87	0.75 (0.44)	35.33% *
5 FF Wizard Info Only	2.68	3.50	0.82	30.60%	2.12	2.72	0.60 (0.43)	28.16%
1,2,3: All VITA Treatments		3.15	0.47	17.54%		2.38	0.26 (0.24)	12.08%
1,4: All FF Gen Treatments		3.83	1.15	42.91%		3.00	0.88 (0.32)	41.46% ***
2,5: All FF Wizard Treatments		3.28	0.60	22.39%		2.52	0.40 (0.30)	18.68%
All Treatments		3.31	0.63	23.51%		2.55	0.43 (0.19)	20.05% **
Used Preparer								
1 VITA Addresses and FF General Info		10.30	1.14	12.45%		9.40	0.73 (0.78)	8.46%
2 VITA Addresses and FF Wizard Info		6.97	-2.19	-23.91%		6.86	-1.81 (0.68)	-20.86% ***
3 VITA Addresses Only		9.04	-0.12	-1.31%		8.47	-0.19 (0.74)	-2.23%
4 FF General Info Only		8.00	-1.16	-12.66%		7.84	-0.82 (0.72)	-9.51%
5 FF Wizard Info Only	9.16	8.52	-0.64	-6.99%	8.66	7.81	-0.86 (0.71)	-9.89%
1,2,3: All VITA Treatments		8.78	-0.38	-4.15%		8.25	-0.41 (0.43)	-4.79%
1,4: All FF Gen Treatments		9.16	0.00	0.00%		8.61	-0.05 (0.53)	-0.60%
2,5: All FF Wizard Treatments		7.76	-1.40	-15.28%		7.34	-1.32 (0.50)	-15.29% ***
All Treatments		8.57	-0.59	-6.44%		8.07	-0.59 (0.33)	-6.79% *

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data Extracted July 2020.
Outcomes are indicator variables with a scale of 0-100. Standard error in parenthesis.
Significance based on model coefficient: *90 percent significance | **95 percent significance | ***99 percent significance

Appendix C. Nonfiler Raw Effects vs. Model Effects Table

Treatment	Raw Data				Model			
	Control	Treated	Change	Percent Change	Control	Treated	Change (S.E.)	Percent Change
All Nonfilers								
1: VITA Addresses and FF General Info		16.02	0.86	5.67%		19.69	-0.02 (0.38)	-0.12%
2: VITA Addresses and FF Wizard Info		16.06	0.90	5.94%		19.91	0.19 (0.38)	0.97%
3: VITA Addresses Only		16.79	1.63	10.75%		20.30	0.59 (0.38)	2.98%
4: FF General Info Only		16.85	1.69	11.15%		20.78	1.07 (0.38)	5.43% ***
5: FF Wizard Info Only	15.16	16.28	1.12	7.39%	19.71	20.08	0.37 (0.38)	1.86%
1,2,3: All VITA Treatments		16.29	1.13	7.45%		19.97	0.25 (0.22)	1.28%
1,4: All FF Gen Treatments		16.43	1.27	8.38%		20.24	0.52 (0.27)	2.66% *
2,5: All FF Wizard Treatments		16.17	1.01	6.66%		19.99	0.28 (0.27)	1.42%
All Treatments		16.40	1.24	8.18%		20.15	0.44 (0.17)	2.23% **
Used VITA								
1 VITA Addresses and FF General Info		0.71	0.25	54.35%		0.71	0.18 (0.08)	33.71% **
2 VITA Addresses and FF Wizard Info		0.53	0.07	15.22%		0.55	0.01 (0.07)	2.25%
3 VITA Addresses Only		0.88	0.42	91.30%		0.85	0.31 (0.09)	58.61% ***
4 FF General Info Only		0.67	0.21	45.65%		0.70	0.17 (0.08)	31.65% **
5 FF Wizard Info Only	0.46	0.45	-0.01	-2.17%	0.53	0.51	-0.03 (0.07)	-4.68%
1,2,3: All VITA Treatments		0.71	0.25	54.35%		0.70	0.17 (0.05)	31.46% ***
1,4: All FF Gen Treatments		0.69	0.23	50.00%		0.71	0.17 (0.06)	32.58% ***
2,5: All FF Wizard Treatments		0.49	0.03	6.52%		0.53	-0.01 (0.05)	-1.31%
All Treatments		0.65	0.19	41.30%		0.66	0.13 (0.04)	24.34% ***
Used FreeFile								
1 VITA Addresses and FF General Info		0.89	0.24	36.92%		0.76	0.12(0.08)	17.88%
2 VITA Addresses and FF Wizard Info		0.71	0.06	9.23%		0.67	0.03 (0.08)	4.35%
3 VITA Addresses Only		0.72	0.07	10.77%		0.70	0.05 (0.08)	8.09%
4 FF General Info Only		0.80	0.15	23.08%		0.75	0.11 (0.08)	16.33%
5 FF Wizard Info Only	0.65	0.96	0.31	47.69%	0.64	0.78	0.13 (0.08)	20.84%
1,2,3: All VITA Treatments		0.77	0.12	18.46%		0.71	0.06 (0.05)	10.11%
1,4: All FF Gen Treatments		0.85	0.20	30.77%		0.75	0.11 (0.06)	17.11% *
2,5: All FF Wizard Treatments		0.83	0.18	27.69%		0.72	0.08 (0.06)	12.60%
All Treatments		0.82	0.17	26.15%		0.73	0.09 (0.04)	13.53% **
Used Preparer								
1 VITA Addresses and FF General Info		6.24	0.13	2.13%		8.76	-0.09 (0.27)	-1.02%
2 VITA Addresses and FF Wizard Info		6.46	0.35	5.73%		9.13	0.28 (0.27)	3.15%
3 VITA Addresses Only		6.78	0.67	10.97%		8.95	0.10 (0.27)	1.08%
4 FF General Info Only		6.35	0.24	3.93%		8.98	0.12 (0.27)	1.38%
5 FF Wizard Info Only	6.11	6.34	0.23	3.76%	8.85	8.98	0.12 (0.27)	1.38%
1,2,3: All VITA Treatments		6.50	0.39	6.38%		8.95	0.10 (0.16)	1.07%
1,4: All FF Gen Treatments		6.30	0.19	3.11%		8.87	0.02 (0.19)	0.18%
2,5: All FF Wizard Treatments		6.40	0.29	4.75%		9.05	0.20 (0.19)	2.26%
All Treatments		6.44	0.33	5.40%		8.96	0.11 (0.12)	1.20%

Source: IRS, Compliance Data Warehouse. Individual Return Transaction File. TY2018 Returns. Data Extracted July 2020.
Outcomes are indicator variables with a scale of 0-100. Standard error in parenthesis.
Significance based on model coefficient: *90 percent significance | **95 percent significance | ***99 percent significance.

Enforcement vs. Outreach: Impacts on Time-To-File, Penalties, and Call Volume

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

Introduction

In July 2019, the Taxpayer First Act was signed into law. A main goal of this Act was to make it easier for taxpayers to interact with the Internal Revenue Service (IRS). This is a broad goal, involving a many-pronged approach. In this paper we expand upon a prior outreach pilot to consider how the IRS can aid taxpayers in resolving tax issues earlier than they would absent IRS communication or through a traditional enforcement process.

In 2017, the IRS began a randomized control trial to compare the impact of tax enforcement via notices regarding delinquent tax returns to outreach encouraging taxpayers to file their current and/or delinquent tax returns. The experimental treatment contacted taxpayers earlier than the enforcement notices would typically be issued and aligned this contact to periods in which tax would be naturally salient (e.g., around the April filing deadline for individual taxpayers). The results of that study were presented in a research paper at the 2019 IRS Research Conference (Herlache *et al.* (2020)). In this extension to the work done by Herlache *et al.*, we focus on the impact of treatments on outcomes related to taxpayer burden. These include the reduction in penalties accumulated due to taxpayers resolving issues sooner and the impact the treatments had on IRS call volume.

Background and Related Research

Nonfilers

"Nonfilers," as it is used in this paper, refers to taxpayers who were identified via the IRS Individual Case Creation Nonfiler Identification Process (CCNIP) as having an unmet filing requirement for Tax Year (TY) 2016, i.e., *known nonfilers*. Known nonfilers are those the IRS can identify through third-party reporting; *unknown nonfilers* are those whose incomes are not reported to the IRS, such as with cash-only arrangements, where they would not be identified during case creation.

The case creation process relies on third-party and other information provided on information returns, like the Form W-2, to determine who is likely to have failed to file a required individual income tax return. Potential nonfilers enter the Return Delinquency (RD) Notice Process, which begins with a mailed notice. Depending on the route within the RD treatment, some taxpayers may receive an additional notice. Typically, taxpayers have up to 14 weeks to respond during the notice process. Those who do not file in response to the mailed treatment may enter a Taxpayer Delinquency Investigation (TDI) status and subsequently face enforcement action (e.g., Automated Collection System, Field Collection, Automated Substitute for Return).

Factors Influencing Individual Taxpayers' (Non)Filing

The Practitioner View on Barriers to Filing

The IRS Nationwide Tax Forums provide an opportunity for tax professionals to exchange information with IRS representatives. Tax practitioners can receive guidance at these forums on how to resolve some of their most difficult cases, and they provide an opportunity for the IRS to showcase their new initiatives and receive valuable feedback. One such feedback mechanism is a series of focus groups in which the IRS can hear from

tax practitioners. At the 2019 IRS Nationwide Tax Forum, one such series of focus groups centered on practitioners with clients who had previously not filed their taxes on time. The results of these focus groups indicated that practitioners felt that they needed to educate their clients on the consequences of not filing an individual income tax return. Such clients seemed unaware of the need to file and/or the ramifications of not filing. This is highlighted in the anecdote of a client assuming they would receive a notice from the IRS if they had to file, and if they hadn't received such a notice, they could assume that they didn't have to file a tax return for the year (IRS (2019)).

Furthermore, fear of the necessary steps to return to filing compliance can weigh heavily upon individual taxpayers. IRS impersonation scams abound and reaching out to the IRS on your own can be both intimidating and time consuming. A common theme from noncompliant clients is that filing in a current year could trigger unwanted attention on prior noncompliant years, which may trigger penalties and other consequences. If you add in reporting requirements around nontraditional employment, such as the gig economy, the landscape becomes more confusing and challenging to navigate, even with the aid of a tax practitioner.

The 2019 Nationwide Tax Forums practitioners indicated that it is often an unrelated issue that prompts nonfilers to return to compliance (e.g., Free Application for Federal Student Aid (FASFA), applying for a mortgage, etc.). Absent that, they suggested that outreach from the IRS can trigger movement toward voluntary compliance. In fact, practitioners specifically mentioned that, in their opinions, contacting nonfilers earlier (i.e., before the traditional enforcement process) and sending reminders to file could help address nonfilers' issues and improve their filing compliance. That is precisely what was studied in the original pilot from which this paper is derived; the extension herein focuses on how earlier treatment and combinations of treatments can reduce the time it takes prior nonfilers to file their current return and reduce the amount of penalties they owe, all while having little to no impact on IRS call volume.

Penalties, Calls, and Taxpayer Burden

Several sources echo some of the themes extracted from the Tax Forums and indicate that a major barrier to prompt filing compliance is a lack of understanding, whether that lack of understanding is around the exact requirements for filing or the consequences of failing to file (Guyton *et al.* (2003); De La Matta *et al.* (2017); Erard and Ho (2003)). Frequently, taxes are perceived as an onerous task, and as the perceived complexity of that task increases, it follows that frustration and anxiety are also likely to increase (Erard and Ho (2003); De La Matta *et al.* (2017)). Likewise, the perception of taxes is often divorced from the benefits of timely filing, for example, how taxpayer dollars are used and penalties avoided (Congdon *et al.* (2009)).

Despite some taxpayers not readily making the connection between tax penalties and the personal ramifications of failing to file, tax penalties play an important function in promoting and defining tax compliance. As an instrument of deterrence, penalties can encourage taxpayers to comply or file earlier than they would have otherwise. However, the accumulation of penalties carries a financial burden for taxpayers, one that they may attempt to sidestep by trying to avoid interacting with the IRS entirely, that is, by not filing a required tax return. As that avoidance and procrastination adds up over time, the perceived burden of filing could become ever more onerous, and potentially lower the incentives for returning to filing compliance. Eventually that obligation may catch up to them and the accumulated penalties would be far less burdensome if they had filed sooner rather than later. This paper investigates how different treatment paths can lessen the time to file, and therefore reduce the penalties (and associated burden) prior nonfilers face in moving toward tax compliance.

The IRS often defines burden in terms of money and time spent on filing taxes. Penalties speak to the former; we turn to phone calls to the IRS as a partial proxy for the latter. While we do not track call *time* in this paper, we consider incoming call *volume* by treatment path as an indication of the burden stemming from IRS contact that taxpayers face to resolve issues. This analysis also provides insight regarding the burden assumed by the IRS, meaning a better understanding of which treatments are more likely to generate (or not generate) a telephone call to the IRS can help the IRS to allocate resources more effectively.

Analysis and Results

Overview of the Original Pilot

The analyses presented here build on a randomized control trial conducted during Calendar Year 2018, which assessed filing behavior for Tax Years 2016, 2017, and 2018. A full description of the study's methodology can be found in Herlache *et al.* (2020). Briefly, the study involved three waves of mailed outreach sent in April, October, and December of 2018. Contact included mailed outreach prior to the filing deadline to remind taxpayers to file their returns, "soft notices" sent near the filing extension deadline and/or near the end of the calendar year, and starting the return delinquency notice process in either TY 2016 or TY 2017 (see Tables 16 and 17 in the Appendix for the experimental design). In an earlier paper on this study, the authors (Herlache *et al.* (2020)) focused on how different treatment paths impacted filing. Table 1 displays the marginal effects for selected treatment paths from that study (see Tables A3-A5 in the Appendix for model results at Wave 3).

TABLE 1. Selected Nonfiler Marginal Effects at Wave 3, 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 (RD) notice process	.067*	.029*	.096	.022*	.118
Simple letter	.019‡	.036*	.055	.041*	.096
Plus Wave 2 soft notice	.030	.052	.082	.033	.115
Plus Wave 3 soft notice	.018	.048	.066	.033	.099
Plus Wave 3 RD process	.026	.072	.098	.047	.145
Simple postcard	-.004	.016	.012	.025*	.037
Plus Wave 2 soft notice	-.003	.017	.014	.020	.034
Plus Wave 3 soft notice	.004	.035	.039	.040	.079
Plus Wave 3 RD process	.006	.046*	.052	.043‡	.095
Soft notice only (Wave 2)	.015‡	.036*	.051	.033*	.084
Soft notice only (Wave 3)	.003	.042*	.045	.026*	.071
TY 2017 RD notice process only (Wave 3)	.013‡	.034*	.047	.023*	.070

NOTES: TY 2016 and TY 2017 outcomes are filing; TY 2018 refers to filing or filing for an extension. * Indicates significance at the 95 percent level; ‡ indicates significance at the 90 percent level. Note that significance indicated for second and third treatments refers to the additional impact of those treatments.

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.

In this paper, we begin by extending the original analysis to consider how different treatment paths influenced the timing of filing. That is, did certain treatments prompt taxpayers to file sooner than they would have otherwise? We investigate this in terms of the average reduction in time to file a TY 2017 return and its associated impact on the average penalty savings. We focus on the most promising alternative treatment path identified in the 2019 study: receiving a letter (either version) in Wave 1, a soft notice in Wave 2 (to those who had not yet filed), followed by a start in the TY 2017 return delinquency (RD) notice process in Wave 3 (among those who had not yet filed). For the sake of brevity, we'll hereafter refer to this treatment path as LtsftN23. We compare this treatment path to the control condition, to a start in the TY 2016 RD notice process, and to a start in the TY 2017 return delinquency notice process.

We also consider the main treatment paths presented in Appendix Table A1 in terms of their impact on calls to the IRS. In this analysis, we collapse across the various Wave 1 reminders. This allows us to focus more

on the number of contacts and the presence or absence of a start in the RD notice process, as compared to the control condition.

Time-To-File (Hazard) Analysis

To estimate the average reduction in time to filing, we used survival analysis to evaluate time-to-event (i.e., time to file after contact). This is a natural extension to our logistic regression analysis of probability of filing. Modeling time-to-file data using survival models will enable us to not only answer whether the treatments have an impact on the probability of filing, but also on the promptness of filing.

The quantity that is typically modeled is the hazard of filing a return; that is, the instantaneous rate of filing at any given time,

$$\lambda(t) = \lim_{\Delta t \to 0} P[t < T < t + \Delta t \mid T > t]$$

where T denotes the time-to-file. The hazard λ(t) is to be interpreted as the instantaneous filing rate. Given that a taxpayer has not filed until time t, what is the probability that he/she will file within another small (Δt) time window? There is an exact relation between the hazard of filing and the expected filing time:

$$E(T) = \int_0^{\infty} \exp\left(-\int_0^t \lambda(u)du\right) dt. \qquad (1)$$

where *E(T)* denotes the expected time to filing. This relation allows us to study the expected time to filing, the quantity of our interest, by studying the hazard of filing.

The *Cox Proportional Hazard* model is appropriate for modeling someone's changes to hazard of filing from his or her own baseline hazard once exposed to treatments (in the case of this paper, receiving a letter or a start in the RD process) (Cox (1972)). We estimate the time-to-file distribution of the nonfiler population by fitting a Cox Proportional Hazard to estimate the effects of the various treatments while controlling for the propensity for the taxpayer to file.

It is important to note that the experimental design is complex because the treatment protocols start at different times during the study and some treatments are sequential (e.g., individuals assigned to treatment at Wave 1 received contact before individuals assigned to begin treatment at Waves 2 or 3, some treatment paths involved contact at more than one wave; see Table A1 in the Appendix for the pilot experimental design). For proportional hazard analysis we used right censoring to accommodate observations censored by the end-of-study date (May 31, 2019) and left truncation to adjust for the delayed entry due to the differing time of administration for the three waves.

Time-To-File (Hazard) Results

The estimates for select treatment effects on time-to-file are given in Table 2 for TY 2017 filing (see Table A6 in the Appendix for the full model results).

TABLE 2. Estimates of Select Treatments on Time-To-File and Hazard Rate, TY 2017

Treatment	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
TY 2016 RD notice process	0.099	0.035	7.996	0.0047	1.104
Simple letter	0.137	0.047	8.386	0.0038	1.147
Additional from Wave 2 soft notice after Wave 1 letter (either version)	0.040	0.057	0.500	0.4797	1.041
Additional from Wave 3 TY 2017 RD notice process after Wave 1 letter (either version) and Wave 2 soft notice	0.080	0.049	2.645	0.1039	1.083
TY 2017 RD notice process (Wave 3 only)	0.464	0.065	51.684	<.0001	1.591
Secured return model score (SRMODEL)	1.466	0.062	559.821	<.0001	4.334
Balance due model score (BDMODEL)	-0.713	0.152	21.902	<.0001	0.490

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.

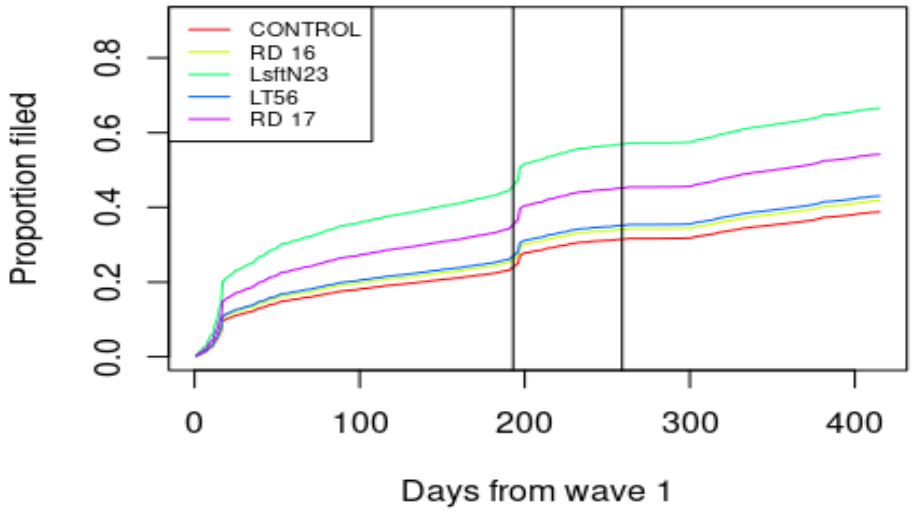
We can see in the regression results that the simple letter and TYs 2016 and 2017 RD processes significantly reduce the time to file a TY 2017 return as compared to the control condition, controlling for the nonfilers' secured return and balance due model scores. It is interesting to consider both the TY 2016 and 2017 RD processes. The TY 2016 process was initiated at Wave 1 during the TY 2017 filing season. This allows us to consider how enforcement directed to past noncompliance impacts current filing compliance. This RD process focused on the prior-year tax return but did prompt taxpayers to file their TY 2017 return sooner than if they were left to their own devices. Likewise, the TY 2017 RD process provides an interesting view into taxpayer behavior, as it was directed toward the TY in question (2017) but was rather removed in time from the filing season, being initiated some 8 months after the filing season. Here we also see the RD process prompting filing earlier than was observed in the control condition.

Sending a simple reminder letter just prior to the filing season prompted prior-year nonfilers to file their current-year return sooner than they would have otherwise. The additional treatments in the LtsftN23 path beyond the initial reminder letter do not rise to the traditional level of statistical significance in terms of reducing the time to file; however, it is important to continue considering this path in terms of its impact on time-to-file and associated penalty reduction, as it was the top-performer in securing additional returns in the original pilot. Extending our understanding to encompass these results provides a more well-rounded view of its performance.

The estimated Cox-regression model will allow estimation of the treatment effect for different risk categories. To see the treatment effect on time to file, we looked at the proportion of taxpayers who filed by time t for the different treatment groups and for different taxpayer risk profiles, in terms of their secured return (SRMODEL) and balance due model (BDMODEL) scores. Specifically, we set the scores to values representing the top, middle, and lower third of the population. This is achieved by determining the 10th, 50th, and 90th percentiles of their respective distributions. The SRMODEL has a positive association with filing behavior and hence taxpayers with higher SRMODEL scores are expected to file earlier than those with lower scores. Conversely, the BDMODEL score has a negative association with filing; taxpayers with higher BDMODEL scores have a higher risk of filing late. Therefore, the percentile combinations are set to either (10, 90), or (90, 10) percentiles for the (SRMODEL, BDMODEL) pair to indicate high- and low-risk groups and "other" to indicate a medium-risk group. It is expected that filing time distribution in the high-risk group will be shifted right toward greater time to file compared to medium- or low-risk groups. Hence, the proportion of people filing by a given time t will be highest for the low-risk group and lowest for the high-risk group.

Based on the estimated model we can perform a counter-factual comparison of treatment effect where we could look at the time-to-file distribution for each treatment assuming the treatments are administered at the same time. The delayed entry (left truncation) analysis assumes that the treatment effect is starting at the time of administration, and hence to compare the estimated treatment effect we align the treatment starting point. For example, the distribution of the group receiving the delinquent notice for TY 2017 is showing the distribution of filing time had they received the treatment at the beginning of Wave 1. This hypothetical comparison is useful in the sense it provides valuable insight to the potential design of early outreach experiments. Figures 1-3 shows the expected proportion filing for the different treatment groups, split by the three risk groups.

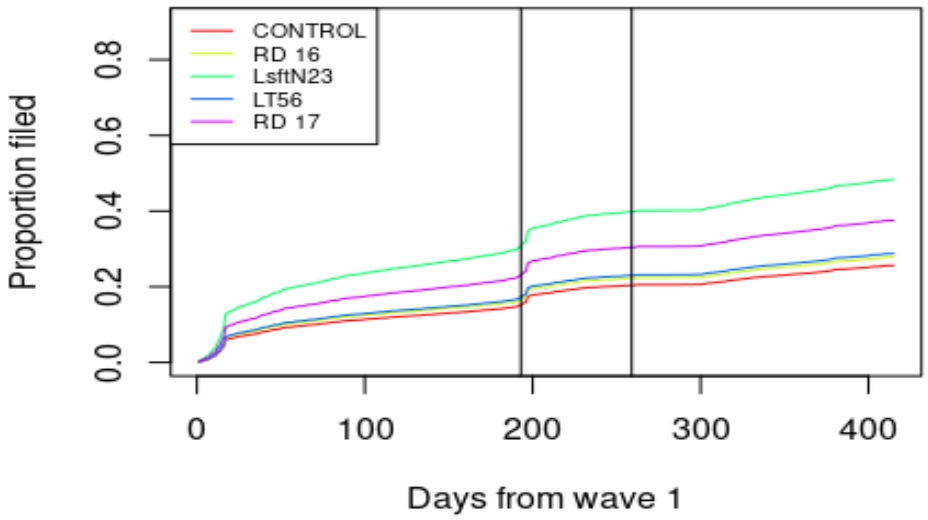
FIGURE 1. Probability of Filing Within a Given Number of Days From Wave 1 Contact Among Low-Risk, Prior-Year Nonfilers, TY 2017



NOTE: The two veritcal lines represent Wave 2 and 3 mailings. RD 16 = TY 2016 RD process start at Wave 1; LsftN23 = strongest treatment path (receiving a letter (either version) in Wave 1, a soft notice in Wave 2 (to those who had not yet filed), followed by a start in the TY 2017 Return Delinquency Notice Process in Wave 3 (among those who had not yet filed)); LT56 = Simple reminder letter at Wave 1; RD 17 = TY 2017 RD process start at Wave 3.

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.

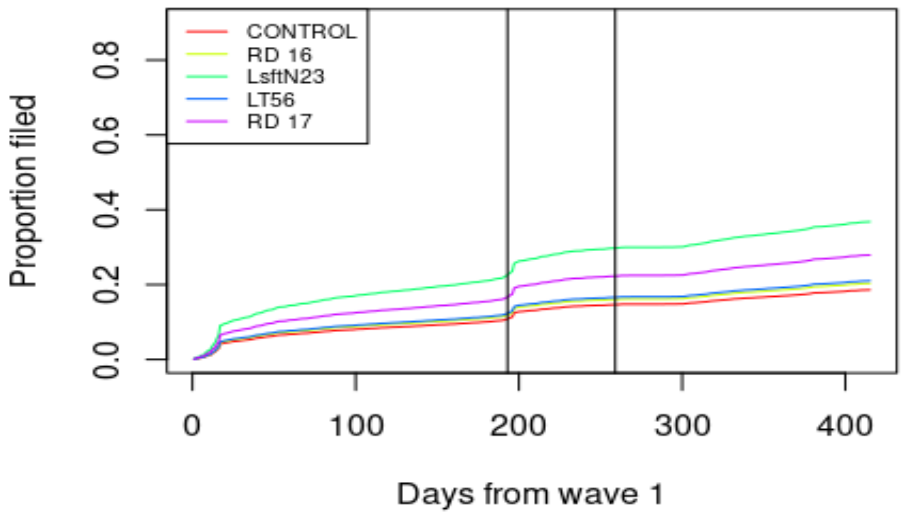
FIGURE 2. Probability of Filing Within a Given Number of Days From Wave 1 Contact Among Medium-Risk, Prior-Year Nonfilers, TY 2017



NOTE: The two veritcal lines represent Wave 2 and 3 mailings. RD 16 = TY 2016 RD process start at Wave 1; LsftN23 = strongest treatment path (receiving a letter (either version) in Wave 1, a soft notice in Wave 2 (to those who had not yet filed), followed by a start in the TY 2017 Return Delinquency Notice Process in Wave 3 (among those who had not yet filed)); LT56 = Simple reminder letter at Wave 1; RD 17 = TY 2017 RD process start at Wave 3.

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. Probability of Filing Within a Given Number of Days From Wave 1 Contact Among High-Risk, Prior-Year Nonfilers, TY 2017



NOTE: The two veritcal lines represent Wave 2 and 3 mailings. RD 16 = TY 2016 RD process start at Wave 1; LsftN23 = strongest treatment path (receiving a letter (either version) in Wave 1, a soft notice in Wave 2 (to those who had not yet filed), followed by a start in the TY 2017 Return Delinquency Notice Process in Wave 3 (among those who had not yet filed)); LT56 = Simple reminder letter at Wave 1; RD 17 = TY 2017 RD process start at Wave 3.

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 figures show the cumulative distribution function (CDF) of the variable time-to-file for different treatments. For example, a value of 0.18 at days = 100 in Figure 1 means roughly 18 taxpayers in the low-risk group for that treatment will file within 100 days of the start of the treatment. This, assuming a homogeneous group, will mean that the probability that a taxpayer from the low-risk group will file within 100 days of the beginning of the treatment is about 0.18.

In general, we could also compute the expected length of time to file from equation (1). Since we are observing T only up to the end-of-study date (May 30, 2019, or 415 days from Wave 1), we can compare the truncated values of the integral in the equation (1) only for different treatment groups. For the three risk groups, the area under the curve from $0 < T < 415$ for the different treatment paths are presented in Table 3 below.

TABLE 3. Expected Time-To-File in Days, by Treatment Path and Risk Group, TYs 2016 and 2017

Treatment	Low-Risk Group	Medium-Risk Group	High-Risk Group
Control (postcard)	310	347	366
Simple letter	296	338	360
TY 2016 RD notice process	300	341	362
TY 2017 RD notice process	262	313	341
LtsftN23	221	281	315

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 actual difference in expected time to file between, for example, the control and simple letters will likely be more than 310 - 296 = 14 days, given that the observations are censored. A conservative estimate is sufficient to indicate the relative gain from the various treatments over the control group, in terms of reduction in expected number of days required to file. Thus, for example, the simple letter treatment will reduce the expected time by about 2 weeks (at least 14 days), compared to the control group which received no tax-related letters or notices.

The treatment path with the greatest impact on increasing the hazard of filing (i.e., prompting filing earlier than the other treatment paths) was LtsftN23 (recall that LtsftN23 began with a letter (either version) in Wave 1, a soft notice in Wave 2 (to those who had not yet filed), followed by a start in the TY 2017 RD notice process in Wave 3 (among those who had not yet filed)). The expected number of days saved in this treatment path compared to the control group are 89, 66, and 51 days for the high-, medium-, and low-risk groups, respectively. We see similar numbers for days saved when comparing the strongest treatment path (LtsftN23) to the TY 2016 RD notice process treatment, with 79, 60, and 46 days saved, and to the TY 2017 RD notice process with 41, 32, and 26 days saved for the high-, medium-, and low-risk groups, respectively.

The Impact of Treatment on Penalties

We extended the results from the time-to-file analyses to assess the impact of the various treatment paths on penalties incurred (failure-to-file and failure-to-pay). To compute a penalty rate per day per dollar balance due we looked at the values of the penalty rate in the given population. For each case where a positive penalty (P) was determined, we noted the number of days over which the penalty was accrued (D) and the total balance due for the case (B) and computed the rate as $R = P/(D * B)$.

From the distribution of R we computed a trimmed mean (there were two extreme values in failure-to-file, which we excluded; for failure-to-pay we excluded the top 0.01 (values) as the average penalty rate E(R). We then computed the expected savings at different percentiles (10, 25, 50, 75, 90, 99) of the total balance due distribution for the strongest treatment path, LtsftN23 (the treatment path starting with a letter (either version) in Wave 1, followed by a soft notice in Wave 2 (to those who had not yet filed), followed by a start in the TY 2017 return delinquency notice process in Wave 3 (among those who had not yet filed)). For example, when the

balance due is approximately around the 25th percentile of the observed balanced due distribution ($1,063) and the case belongs to the low-risk category (i.e., having a high SRMODEL score and a low BDMODEL score), then the expected savings from applying the LtsftN23 over the control is $95. A similar figure for the expected savings at the 75th percentile of balance due for the low-risk category is $776. These numbers are the expected savings with respect to the failure-to-file penalty.

Table 4 gives the expected savings from LtsftN23 treatment path over control, Table 5 gives the expected savings from LtsftN23 treatment path over the TY 2016 RD notice process, and Table 6 gives the expected savings from LtsftN23 treatment path over the TY 2017 RD notice process, all with respect to failure-to-file (left panel) and failure-to-pay (right panel).

TABLE 4. Expected Dollar Savings From the LtsftN23 Treatment Path Over the Control Group

Balance Due	Failure-To-File			Failure-To-Pay		
	Low Risk	Average Risk	High Risk	Low Risk	Average Risk	High Risk
10%	30	23	17	8	6	4
25%	95	71	54	24	18	14
50%	294	219	168	75	56	43
75%	776	578	445	199	148	114
90%	1,799	1,340	1,030	460	343	264
99%	11,745	8,749	6,724	3,007	2,240	1,721

TABLE 5. Expected Dollar Savings From the LtsftN23 Treatment Path Over TY 2016 RD Notice Process

Balance Due	Failure-To-File			Failure-To-Pay		
	Low Risk	Average Risk	High Risk	Low Risk	Average Risk	High Risk
10%	27	20	16	7	5	4
25%	85	64	49	22	16	13
50%	264	198	153	68	51	39
75%	698	524	404	179	134	103
90%	1,616	1,213	935	414	311	239
99%	10,555	7,922	6,109	2,702	2,028	1,564

TABLE 6. Expected Dollar Savings From the LtsftN23 Treatment Path Over TY 2017 RD Notice Process

Balance Due	Failure-To-File			Failure-To-Pay		
	Low Risk	Average Risk	High Risk	Low Risk	Average Risk	High Risk
10%	14	11	9	4	3	2
25%	44	34	27	11	9	7
50%	138	107	84	35	27	21
75%	364	282	221	93	72	57
90%	842	654	512	216	167	131
99%	5,502	4,270	3,345	1,408	1,093	856

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.

Effect of the Taxpayer First Act

While the original pilot took place before the Taxpayer First Act (H.R.3151) was signed into law on July 1, 2019, we have extended our findings to account for the new structure of the failure-to-file penalty to better understand how the tested treatment paths would impact current taxpayers. Under the Taxpayer First Act, the penalty is now a nonlinear function of the balance due and the number of days from tax date before filing, and the expected penalty must be approximated. The expected penalty is now

$$E(P) = E(B^*R^*T^*(T\leq 60))^*P(T\leq 60) + [435^*(B\leq 435)+(B>435)^*\max(435, B^*R^*T^*(T>60))]^*P(T>60)$$

The computation is further complicated by the fact that we observe T only when it is not censored (i.e., the return is filed). Therefore, we need to take censoring into account. This is done by estimating the distribution of T using a survival model, the Cox Proportional Hazard model.

We then approximate E(P) by substituting the number of days with the expected number of days and evaluating the penalty for B at a given balance due percentile.

$$P \approx B^*R^*E[T(T\leq 60)]^*P(T\leq 60)+[435^*(B\leq 435)+(B>435)^*\max(435, B^*R^*E(T^*(T>60))]^*\ P(T > 60)$$

where B is the given balance due percentile (conditional on balance due being positive) obtained empirically from the observed balance due values. R is the per day per dollar rate taken to be 0.0015 and the expected and probability values for the distribution of T are computed based on the survival model at different values of SRMODEL scores for different risk categories.

Table 7 gives the expected savings from the LtsftN23 treatment path over control (left), over TY 2016 Return Delinquency Notice Process (middle), and over the TY 2017 Return Delinquency Notice Process (right) for the current Failure to File penalty structure. Again, we see the LtsftN23 treatment producing greater savings than the control and either return delinquency notice process across both risk and balance due distribution.

TABLE 7. Expected Dollar Savings From the LtsftN23 Treatment Path Over the Control Group, the TY 2016 and TY 2017 RD Notice Processes Recalculated To Account for the Taxpayer First Act Changes to Failure-To-File Penalties

Balance Due	Expected Savings Over Control			Expected Savings Over TY 2016 RD			Expected Savings Over TY 2017 RD		
	Low Risk	Average Risk	High Risk	Low Risk	Average Risk	High Risk	Low Risk	Average Risk	High Risk
10%	52	35	25	47	32	23	26	18	13
25%	103	96	74	90	87	67	35	47	37
50%	394	297	229	354	269	208	183	144	113
75%	1,041	785	606	934	710	550	484	382	301
90%	2,412	1,818	1,405	2,165	1,645	1,276	1,122	884	697
99%	15,753	11,874	9,174	14,141	10,744	8,331	7,331	5,774	4,553

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.

Impact on Call Volume

The efficacy of different enforcement treatments depends not only on the extent to which they induce taxpayers to file their outstanding and future tax returns and pay the taxes that they owe, but also on the amount of resources that the IRS must spend to obtain the increased level of compliance and the amount of burden that results from noncompliance and treatment. The identification of nonfilers and the generation of typical notice treatments is automated, so the initial cost of such treatments is relatively small. However, if a large number of taxpayers call the IRS upon receipt of such notices, then that would significantly increase the cost of treatment,

either because of the additional dollars needed to hire more employees, or the lower level of IRS responsiveness to other taxpayer issues.

To better understand the impact of the pilot treatments on both taxpayers and the IRS, we assess the various treatment paths in relation to the amount of trackable incoming calls (that is, calls from taxpayers that progressed far enough in the call structure to provide their taxpayer identification number (TIN), thus making it possible to link call activity back to the pilot sample). Tables 8 and 9 provide a descriptive overview of the impact of the pilot treatments on incoming calls. Table 8 shows the mean number of calls by treatment group for each wave and across the course of the study. Table 9 shows the percentage of taxpayers making at least one call by treatment group for each wave and across the course of the study. It also provides the percentage point difference from the control group. (Both tables are broken out by general treatment, not specified by type of outreach reminder received in Wave 1.)

TABLE 8. Inbound Call Rates by Treatment Group, TYs 2016 and 2017

Treatment Group	Inbound Call Rate			
	Wave 1	Wave 2	Wave 3	All Waves
1) Wave 1 TY 2016 RD start	0.167	0.063	0.084	0.314
2) Wave 1 reminder only	0.144	0.046	0.068	0.257
3) Wave 2 soft letter only	0.132	0.059	0.065	0.255
4) Wave 3 soft letter only	0.123	0.048	0.068	0.239
5) Wave 3 TY 2017 RD start	0.130	0.048	0.077	0.255
6) Reminder + Wave 2 soft letter	0.146	0.050	0.069	0.266
7) Reminder + Wave 2 soft letter + Wave 3 soft letter	0.129	0.049	0.067	0.246
8) Reminder + Wave 2 soft letter + Wave 3 RD start	0.141	0.057	0.082	0.280
9) Control	0.136	0.044	0.064	0.244

SOURCE: IRS, Compliance Data Warehouse. Collection Accounts Management System Disclosure. Data extracted May 2019.

TABLE 9. Percentage of Taxpayers Making Calls by Treatment Group, Waves 1, 2, and 3

Treatment Group	Percentage of Taxpayers Making at Least One Call				Percentage Point Difference From Control Group			
	Wave 1	Wave 2	Wave 3	All Waves	Wave 1	Wave 2	Wave 3	All Waves
1) Wave 1 TY2016 RD start	10.2%	4.3%	5.3%	15.7%	2.2	1.1	1.2	3.3
2) Wave 1 reminder only	8.5%	3.4%	4.1%	12.9%	0.5	0.2	0.0	0.5
3) Wave 2 soft letter only	7.8%	3.7%	4.2%	12.8%	-0.2	0.5	0.1	0.4
4) Wave 3 soft letter only	7.4%	3.3%	4.2%	12.1%	-0.6	0.1	0.1	-0.3
5) Wave 3 TY2017 RD start	7.5%	3.3%	5.0%	12.8%	-0.5	0.1	0.9	0.4
6) Reminder + Wave 2 soft letter	7.8%	3.3%	4.6%	12.7%	-0.2	0.1	0.5	0.3
7) Reminder + Wave 2 soft letter + Wave 3 soft letter	7.6%	3.6%	4.6%	12.7%	-0.4	0.4	0.5	0.3
8) Reminder + Wave 2 soft letter + Wave 3 RD Start	8.6%	3.9%	5.2%	14.1%	0.6	0.7	1.1	1.7
9) Control	8.0%	3.2%	4.1%	12.4%				

SOURCE: IRS, Compliance Data Warehouse. Collection Accounts Management System Disclosure. Data extracted May 2019.

Based on the above tables, it appears that while there were calls associated with the taxpayers in this study, the taxpayers in the control group called at a relatively high rate. As shown in Table 9, only treatments 1 and 8, which included a start in the RD notice process, increased the number of taxpayers making a call by more than 1 percentage point, compared to the control group. The treatments limited to soft contacts (i.e., reminders and/or soft notices) had little impact on call volume. Furthermore, note the call rate in the control condition; it appears that taxpayers with underlying issues will call the IRS regardless of contact.

We further assess the impact of treatment on calls while controlling for taxpayer characteristics like income and prior IRS action. Table 10 shows the results of a logistical regression estimating the effects of the eight nonfiler treatments on the likelihood of calling the IRS, compared to the control group. The model controls for whether the taxpayer has an existing outstanding balance (TDA) or return (TDI) or both; whether the taxpayer called the IRS in the previous year; the amount of income showing on information returns; the presence of nonemployee compensation, retirement income and mortgage interest; and the number of information returns. Only treatments 1 and 8 had a significant impact on the rate of calling. It is important to note that these treatment paths included a start in the return delinquency notice process. Taxpayers assigned to treatment 1, the start in the TY 2016 RD process, received the initial notice in April of 2018 and would have progressed through that process, receiving additional IRS treatment over time if they did not resolve the issue. Taxpayers assigned to group 8 received soft contacts in April and October of 2018, followed by starting the TY 2017 RD process in December of 2018, thus having several contacts over the course of the 2018 calendar year. Noticeably, taxpayers in treatment group 5, who were assigned to a start in the TY 2017 RD process in December of 2018, without prior soft contact, do not exhibit a statistically significant increase in the likelihood of calling. This may be due to the data being followed through May of 2019—that is, taxpayers in this group would have received only the first notice in the RD process by that point in time. Following this treatment through the summer of 2019 would likely show a pattern similar to treatment group 1 and can be addressed in follow-up analyses. Similarly, a future analysis could investigate how filing and calling covary as a function of treatment.

Treatments with just reminder letters or soft notices had no appreciable impact on calling. The results also suggest that having an existing unpaid assessment with the IRS (TDA) affects the call rate, but the treatments related to the unfiled return have little impact. Even the traditional, more demanding form of nonfiler treatment—return delinquency notices—have a relatively small impact on calling. The regression results suggest that for every 100,000 taxpayers sent a return delinquency notice as a part of automatic case creation, only about 3,700 additional taxpayers will call at least once during the subsequent year.

TABLE 10. Logistic Model of Taxpayer Makes at Least One Call, Wave 1 to Wave 3

Item	Estimate	Standard Error	Marginal Effect
Intercept	-1.518	0.018	
Called in the prior year	0.847	0.021	
TR1 Wave1 TY2016 RD start	0.175	0.026	3.700
TR2 Wave 1 reminder only	0.017	0.027	0.310
TR3 Wave 2 soft letter only	0.033	0.027	0.740
TR4 Wave 3 soft letter only	-0.005	0.027	0.120
TR5 Wave 3 TY2017 RD start	0.035	0.027	0.680
TR6 Wave 1 reminder + Wave 2 soft letter	0.006	0.027	0.120
TR7 Wave 1 reminder + Wave 2 soft letter + Wave 3 soft letter	0.017	0.021	0.330
TR8 Wave 1 reminder + Wave 2 soft letter + Wave 3 RD start	0.086	0.021	1.660
Presence of Social Security or pension income	0.014	0.015	
Presence of nonemployee compensation	-0.001	0.014	
TDA and TDI from return prior to TY 2016	0.269	0.026	
TDA only from return prior to TY 2016	0.322	0.015	
TDI only from return prior to TY 2016	-0.052	0.026	
More than 10 information return documents	0.015	0.020	
Presence of mortgage interest on F1098	0.129	0.016	
Total IRP income > $100,000	0.250	0.018	
Total IRP income between $50,000 and $100,000	0.148	0.016	

SOURCE: IRS, Compliance Data Warehouse. Collection Accounts Management System Disclosure, Individual Masterfile Status History, and Individual Case Creation Nonfiler Identification Process. Data extracted May 2020.

Conclusions

We extend Herlache *et al.* (2020) to look at the impacts of earlier contact for taxpayers at risk of not meeting their filing requirements. We focus on indirect measures of burden relating to filing and payment of taxes. We use penalty avoidance as a proxy for the burden reduction of early treatment. Thus, treatments that help taxpayers avoid penalties, reduce burden. We find that the most proactive early intervention significantly reduces burden in terms of penalty avoidance, especially for lower risk taxpayers. This is particularly important since the Taxpayer First Act increased penalties for not filing. Beginning with a simple reminder during the April filing season and following up around the extension deadline with those who have not yet filed, brings additional taxpayers into compliance. Directing those who are not moved by simpler treatments into the RD notice process saves the more costly intervention for those who need a stronger nudge. This tiered approach provides benefits to the taxpayers, as it encourages them to act sooner than the other treatments studied in the original pilot, including the traditional RD process, thus avoiding accruing greater penalties.

Likewise, studying the impact on different treatments on call volume is an important consideration for both taxpayers and tax administrations alike. The results from this analysis indicate that if a taxpayer has an existing issue that needs to be addressed (i.e., an unpaid assessment), they are more likely to call the IRS, regardless of current treatment. Overall, soft contacts (reminder letters and soft notices) do not seem to spark additional calls. Starts in the RD notice process do prompt taxpayers to call the IRS, likely en route to resolving their tax issues, but the rate at which they do so is fairly low.

Taken together, the results presented in this paper suggest that beginning with soft contact encouraging prior-year nonfilers to file their current-year return and escalating into a formal delinquent return notice, if necessary, saves the taxpayer money in penalties while not prompting a call the IRS. This seems beneficial to both the IRS, as it improves filing compliance and has a relatively low impact on resources, and taxpayers, as they begin to resolve their tax issues and avoid burdensome penalties and time spent in an IRS call system.

References

Congdon, William, Jeffrey R. Kling, and Sendhil Mullainathan (2009). *Behavioral Economics and Tax Policy*. Cambridge, MA: National Bureau of Economic Research (NBER) Working Paper No. 175.

Cox, D.R. (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2): 187–220.

De La Matta, Jose Colon, John L. Guyton, Ronald Hodge II, Patrick Langetieg, Stacy Orlett, Mark Payne, Ahmad Qadri, Lisa Rupert, Brenda Schafer, Alex Turk, and Melissa Vigil (2017). "Understanding the Nonfiler/Late Filer: Preliminary Findings." *2016 IRS Research Bulletin* (Publication 1500), 141–163. Washington DC: Internal Revenue Service, Statistics of Income Division.

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

Guyton, John L., John F. O'Hare, Michael P. Stavrianos, and Eric J. Toder (2003). "Estimating the Compliance Cost of the U.S. Individual Income Tax." *National Tax Journal* LVI(3): 673–688.

Herlache, Anne, Stacy Orlett, Ishani Roy, and Alex Turk (2020). "Enforcement Versus Outreach—Impacts on Tax Filing Compliance." *2019 IRS Research Bulletin* (Publication 1500), 30–71. Washington, DC: Internal Revenue Service, Statistics of Income Division.

Internal Revenue Service (IRS) (2019). *Bringing Taxpayers Back Into Filing Compliance.* 2019 IRS National Tax Forum, Focus Group. Washington, DC: Internal Revenue Service, Small Business/Self-Employed Division. Internal report, unpublished.

Appendix
Original Study Experimental Design and Randomization Notes

TABLE A1. 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 A2.

TABLE A2. 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.

TABLE A3. Nonfiler Regression Results from Original Study With and Without Undeliverable Control for Wave 3, TY 2016

Dependent Variable: Filed return for TY 2016 after treatment

Parameters	Modeling With Undeliverable Control			Modeling Without Undeliverable Control		
	Parameter Estimates		Marginal Effect of Treatment	Parameter Estimates		Marginal Effect of Treatment
Intercept	0.988*	(0.089)		-1.885*	(0.055)	
Simple Letter	0.096	(0.070)	0.015	0.120‡	(0.069)	0.019
Complex Letter	0.066	(0.070)	0.010	0.079	(0.069)	0.013
Simple Postcard	-0.037	(0.072)	-0.006	-0.022	(0.071)	-0.004
Complex Postcard	-0.020	(0.079)	-0.003	-0.014	(0.071)	-0.002
Return Delinquency Notice Process	0.438*	(0.048)	0.069	0.414*	(0.047)	0.067
Secured Return Model Score	-0.763*	(0.140)		3.582*	(0.094)	
Balance Due Model Score	-0.659*	(0.207)		-1.258*	(0.207)	
Balance Due Model Score Squared	0.189	(0.210)		1.072*	(0.211)	
Indicator Taxpayer Filed TY 2017 Return Prior to Outreach	-0.076	(0.051)		0.094‡	(0.050)	
Wave 2 Soft Notice Only	0.089‡	(0.050)	0.014	0.091‡	(0.049)	0.015
Add—Wv2 Soft Letter After Reminder Letter	0.087	(0.086)	0.014	0.071	(0.084)	0.011
Add—Wv2 Soft Letter After Reminder Postcard	0.027	(0.088)	0.004	0.007	(0.086)	0.001
Add—Wv3 Soft Letter After Reminder Letter & Wv2 Soft Letter	-0.078	(0.074)	-0.012	-0.075	(0.073)	-0.012
Add—Wv3 Soft Letter After Reminder Postcard & Wv2 Soft Letter	0.037	(0.076)	0.006	0.041	(0.075)	0.007
Add—Wv3 RD Start After Reminder Letter & Wv2 Soft Letter	-0.026	(0.074)	-0.004	-0.026	(0.073)	-0.004
Add—Wv3 RD Start After Reminder Postcard & Wv2 Soft Letter	0.041	(0.076)	0.006	0.054	(0.075)	0.009
Wave 3 Soft Notice Only	0.004	(0.051)	0.0006	0.021	(0.050)	0.003
Wave 3 Delinquent Return Notice Process Only	0.100‡	(0.051)	0.016	0.084‡	(0.050)	0.013
Probability of Undeliverable	-11.960	(0.302)				
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 A4. Nonfiler Regression Results from Original Study With and Without Undeliverable Control for Wave 3, TY 2017

Dependent Variable: Filed return for TY 2017 after treatment

Parameters	Modeling With Undeliverable Control			Modeling Without Undeliverable Control		
	Parameter Estimates		Marginal Effect of Treatment	Parameter Estimates		Marginal Effect of Treatment
Intercept	1.765*	(0.081)		-1.154*	(0.049)	
Simple Letter	0.175*	(0.064)	0.032	0.187*	(0.062)	0.036
Complex Letter	0.113‡	(0.064)	0.021	0.117‡	(0.062)	0.022
Simple Postcard	0.072	(0.065)	0.013	0.084	(0.063)	0.016
Complex Postcard	0.022	(0.065)	0.004	0.027	(0.063)	0.005
Delinquent Return Notice Process	0.173*	(0.046)	0.032	0.153*	(0.045)	0.029
Secured Return Model Score	-1.695	(0.123)		2.569*	(0.082)	
Balance Due Model Score	-0.669	(0.190)		-1.113*	(0.188)	
Balance Due Model Score Squared	-0.356	(0.193)		0.394*	(0.191)	
Indicator Taxpayer Filed TY 2016 Return Prior to Outreach	1.944	(0.048)		2.275*	(0.047)	
Wave 2 Soft Notice Only	0.194*	(0.045)	0.036	0.189*	(0.044)	0.036
Add - Wv2 Soft Letter After Reminder Letter	0.099	(0.078)	0.018	0.083	(0.076)	0.016
Add - Wv2 Soft Letter After Reminder Postcard	0.038	(0.080)	0.007	0.008	(0.078)	0.001
Add - Wv3 Soft Letter After Reminder Letter & Wv2 Soft Letter	-0.031	(0.067)	-0.006	-0.023	(0.066)	-0.004
Add - Wv3 Soft Letter After Reminder Postcard & Wv2 Soft Letter	0.089	(0.068)	0.017	0.095	(0.067)	0.018
Add - Wv3 RD Start After Reminder Letter & Wv2 Soft Letter	0.107	(0.067)	0.020	0.106	(0.065)	0.020
Add - Wv3 RD Start After Reminder Postcard & Wv2 Soft Letter	0.134	(0.068)	0.025	0.149*	(0.067)	0.029
Wave 3 Soft Notice Only	0.217*	(0.045)	0.040	0.221*	(0.044)	0.042
Wave 3 Delinquent Return Notice Process Only	0.120*	(0.046)	0.037	0.176*	(0.045)	0.034
Probability of Undeliverable	-12.033	(0.266)				
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 A5. Nonfiler Regression Results from Original Study With and Without Undeliverable Control for Wave 3, TY 2018

Dependent Variable: Filed return or extension for TY 2018 after treatment

Parameters	Modeling With Undeliverable Control			Modeling Without Undeliverable Control		
	Parameter Estimates		Marginal Effect of Treatment	Parameter Estimates		Marginal Effect of Treatment
Intercept	0.607*	(0.073)		-1.369*	(0.046)	
Simple Letter	0.184*	(0.058)	0.037	0.199*	(0.058)	0.041
Complex Letter	0.092	(0.058)	0.019	0.102‡	(0.058)	0.021
Simple Postcard	0.113‡	(0.059)	0.023	0.120*	(0.058)	0.025
Complex Postcard	0.012	(0.059)	0.002	0.014	(0.058)	0.003
Delinquent Return Notice Process	0.114*	(0.042)	0.023	0.106*	(0.042)	0.022
Secured Return Model Score	2.940*	(0.119)		5.986*	(0.086)	
Balance Due Model Score	-1.250*	(0.176)		-1.696*	(0.174)	
Balance Due Model Score Squared	0.634*	(0.179)		1.273*	(0.177)	
Wave 2 Soft Notice Only	0.160*	(0.041)	0.032	0.162*	(0.041)	0.033
Add—Wv2 Soft Letter After Reminder Letter	-0.032	(0.072)	-0.006	-0.037	(0.071)	-0.008
Add—Wv2 Soft Letter After Reminder Postcard	-0.014	(0.072)	-0.003	-0.026	(0.071)	-0.005
Add—Wv3 Soft Letter After Reminder Letter & Wv2 Soft Letter	0.003	(0.062)	0.001	-0.001	(0.062)	-0.000
Add—Wv3 Soft Letter After Reminder Postcard & Wv2 Soft Letter	0.096	(0.062)	0.019	0.096	(0.061)	0.020
Add—Wv3 RD Start After Reminder Letter & Wv2 Soft Letter	0.072	(0.062)	0.015	0.069	(0.061)	0.014
Add—Wv3 RD Start After Reminder Postcard & Wv2 Soft Letter	0.104‡	(0.062)	0.021	0.111‡	(0.061)	0.023
Wave 3 Soft Notice Only	0.118*	(0.042)	0.024	0.126*	(0.041)	0.026
Wave 3 Delinquent Return Notice Process Only	0.122*	(0.043)	0.025	0.110*	(0.042)	
Probability of Undeliverable	-8.155	(0.043)				
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 A6. Regression Results—Treatment Effects on Time to File and Hazard Rate, TY 2017

Analysis of Maximum Likelihood Estimates

Treatment	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Simple Letter	1	0.137	0.047	8.386	0.004 ***	1.147
Complex Letter	1	0.088	0.048	3.452	0.063 *	1.092
Simple Postcard	1	0.060	0.049	1.474	0.225	1.062
Complex Postcard	1	0.0157	0.050	0.101	0.751	1.016
TY 2016 Return Delinquency Notice Process	1	0.099	0.035	7.996	0.005 ***	1.104
Secured Return Model Score	1	1.466	0.062	559.821	<0.0001 ***	4.334
Balance Due Model Score	1	-0.713	0.152	21.902	<0.0001 ***	0.490
Balance Due Model Score squared	1	0.173	0.157	1.202	0.273	1.188
Filed TY 2016 early	1	1.212	0.025	2377.580	<0.0001 ***	3.359
Soft Notice (Wave 2 Only)	1	0.160	0.052	9.576	0.002 ***	1.174
Additional from Wave 2 soft notice after Wave 1 letter (either version)	1	0.040	0.057	0.500	0.480	1.041
Additional from Wave 2 soft notice after Wave 1 postcard (either version)	1	-0.006	0.060	0.010	0.921	0.994
Additional from Wave 3 soft notice after Wave 1 letter (either version) and Wave 2 soft notice	1	-0.009	0.050	0.032	0.856	0.991
Additional from Wave 3 soft notice after Wave 1 postcard (either version) and Wave 2 soft notice	1	0.074	0.051	2.090	0.149	1.077
Additional from Wave 3 TY 2017 return delinquency notice process after Wave 1 letter (either version) and Wave 2 soft notice	1	0.080	0.049	2.645	0.104	1.083
Additional from Wave 3 TY 2017 return delinquency notice process after Wave 1 postcard (either version) and Wave 2 soft notice	1	0.111	0.051	4.721	0.030 **	1.118
Soft notice (Wave 3 only)	1	0.210	0.072	8.491	0.004 ***	1.234
TY 2017 return delinquency notice process (Wave 3 only)	1	0.464	0.065	51.684	<0.0001 ***	1.591

*** Significant at 1% ** significant at 5% and * significant at 10%.

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.

Perspectives on New Forms of Remote Identity Proofing and Authentication for IRS Online Services

Becca Scollan, Melanie Shere, and Ronna ten Brink (MITRE)[1]

1. Introduction

In May of 2019 OMB released Memorandum M-19-17, *Enabling Mission Delivery Through Improved Identity, Credential, and Access Management* (OMB (2019)). The memo states that Federal "…agencies must implement National Institute of Standards and Technology (NIST) Special Publication (SP) 800-63-3 and any successive versions (hereafter referred to as NIST SP 800-63)." The NIST SP-800-63 *Digital Identity Guidelines* (includes updates as of 12-01-2017, NIST (2017)) cover identity proofing and authentication of users (such as employees, contractors, or private individuals) interacting with government IT systems over open networks. They define technical requirements in each of the areas of identity proofing, registration, authenticators, management processes, authentication protocols, federation, and related assertions.

The IRS is subject to the digital identity standards developed by NIST, as well as mandates to improve its customer experience. For example, Taxpayer First Act (U.S. Congress (2019 July, 1)), which became law in July 2019, requires the IRS to develop a "comprehensive customer service strategy" and emphasizes providing secure services to taxpayers that meet the best practices for online services in the private sector. The IRS also has a business imperative to provide improved, secure online services to address challenges such as the high cost, up to $41, of an average call to the IRS customer service (Konkel (2018)).

Over the past decade, the IRS has developed a number of services that enable taxpayers to access their personal, sensitive data to meet legislative, business, and user demands. For example, Get Transcript (IRS (n.d.)) provides digital access to transcripts of prior tax returns, and the IRS Account supports taxpayers in looking up a balance owed and viewing payment history. Both services utilize IRS Secure Access (IRS (n.d.)) to provide access to personal information and require personal information to verify those requesting access are who they say they are.

The updated NIST guidelines impact taxpayers seeking to access services using future IRS services at NIST SP 800-63 Identity Assurance Level 2 (IAL2) and Authenticator Assurance Level 2 (AAL2). IAL2 requires remote or in-person identity proofing to verify an applicant. The proofing requires one or more evidence documents, for example, a Real ID or Passport. AAL2 requires that an organization has high confidence that the user controls the authenticator(s) bound to their digital identity. AAL2 also requires proof of possession and control of two distinct authentication factors through secure authentication protocol(s). NIST SP 800-63 also provides guidance for Federal agencies to work with third-party Credential Service Providers (CSPs), who can identity proof, register authenticators and issue credentials to users at both IAL2 and AAL2 (NIST (2017)).

[1] The authors also wish to thank Kim Jarvi in support of this work.

There is little information on how U.S. taxpayers will react to new methods the IRS may choose to implement within IRS Secure Access to comply with NIST SP 800-63, such as new authenticators, remote identity proofing on a smartphone or use of a CSP. To provide more insight into user preferences and perceptions of such methods, MITRE conducted two qualitative user research studies on potential future capabilities of IRS Secure Access on behalf of the IRS.

We conducted two qualitative user research studies with representative users of IRS online services to understand user perceptions and comprehension of these new digital identity concepts, as well as to unearth key usability and accessibility considerations. In the first study, MITRE developed clickable wireframes of an IRS Secure Access application flow compliant with NIST SP 800-63 IAL and AAL 2 requirements and conducted usability testing with 13 tax professionals. The wireframes began on a mocked-up version of IRS.gov and showed the current eAuthentication "create account" process with some additions, before displaying a mocked-up identity proofing flow within IRS2Go. The prototype demonstrated several in-depth features of validating identity documents remotely, such as a "selfie" verification of a license, and liveness testing using voice, text, or physical movement. Through semi-structured interviews and a usability walk-through of the wireframes, we investigated the tax professionals' willingness to remotely identity proof and the usability and accessibility opportunities and concerns of the new process.

In the second study, we prototyped a notional CSP application flow and conducted semi-structured interviews and usability walk-throughs with 19 individual taxpayers who are considered potential users of an IRS Account. The CSP prototype included setting up two-factor authentication and remote identity proofing using one or more identity evidence documents. The research goals of the second study were to understand taxpayers' perceptions of and comprehension of the CSP concept, including topics such as factors that affect willingness to choose to use a CSP, emotional responses to using a CSP, and whether or not taxpayers understand what using a CSP means technically as well as how it can impact them.

Our results suggest that both groups are willing to engage with the new identity management concepts, with some notable reservations and misunderstandings. Based on our observations we developed a series of observations and recommendations on remote identity proofing, authentication, and perceptions on the use of a CSP for the IRS context, and we laid the groundwork for future research and design questions for developing usable and secure account creation and authentication. Through such research, the IRS can ultimately help ensure that taxpayers make informed decisions about their willingness to participate and accept IRS authentication, account management, and privacy practices.

2. Secure Access Study

2.1 Method Overview

We conducted semi-structured interviews and a usability walk-through with 13 Tax Professionals between January 4 through 14, 2019. In the usability walk-through, we displayed a wireframe prototype with a notional workflow of an IAL2 customer journey, and asked participants to "think aloud" and indicate where they would click next.

2.2 Participants

We recruited 13 Tax Professionals and conducted research sessions January 4 through 14, 2019. Participants were recruited using an IRS Secure Access Beta participant listserv. No incentive was offered.

There was no formal screening of participants via a qualifying survey mechanism. Participation was voluntary based on criteria provided by email. Participants received the following information on the study and its requirements for participation: "Participants will be practicing tax professionals (e.g., Enrolled Agents, Certified Public Accountants, Tax Attorneys) who use e-Services currently or have in the past, who are not representatives or employees of State agencies." All voluntary responses to the recruitment email were offered the option to participate in a session.

The majority (10) of participants were age 55 years or older, with 11 out of 13 reporting 6 or more years of experience as a Tax Professional. We also had a fairly even sampling of males to females of those who

participated, with seven females volunteering and six males. Most participants owned a tax services firm with 100 or more clients. The majority reported using the IRS Transcript Delivery Service on a weekly basis. Additionally, 11 of the 12 participants reported they are authorized eFile providers.

FIGURE 1. Age Distribution of Participants

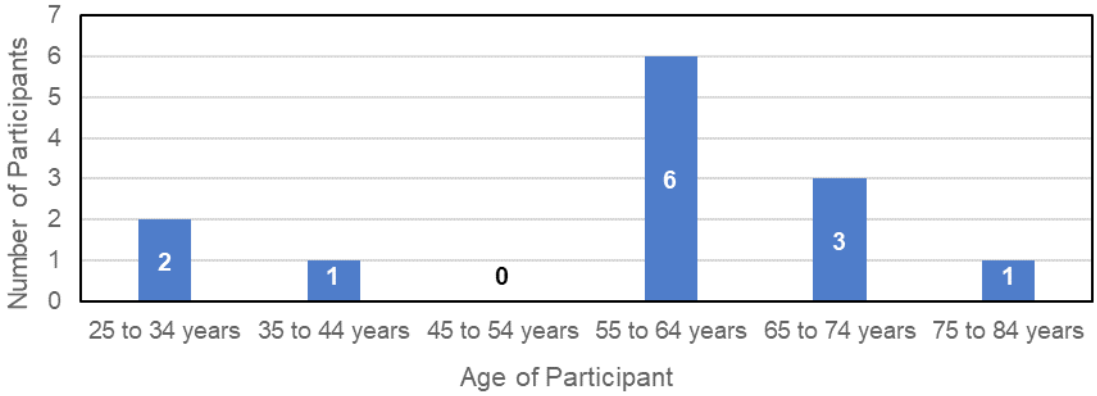
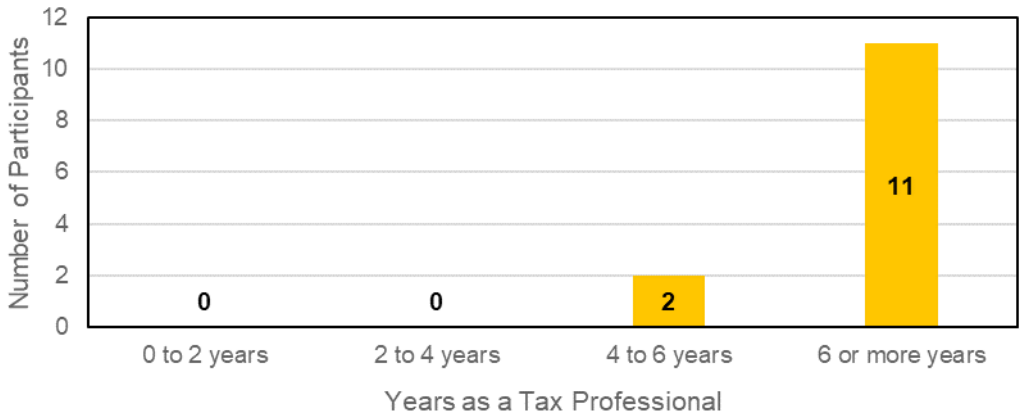


FIGURE 2. Distribution of Years as a Tax Professional



All participants reported that they have used IRS Secure Access to register an account for IRS Tax Professional services; when using Secure Access, five of the participants expressed they went through the current Secure Access process without any issues. The other eight mentioned some challenge or frustration with use, such as difficulties providing financial information, data validation, and only owning a company-issued smartphone (Secure Access requires a smartphone registered in the applicant's name). Otherwise, the exact issue while using the process was unclear to participants.

All 13 participants reported a high level of personal security practice for their smartphone. 75 percent had used their smartphone to access sensitive personal accounts such as a bank account, and 75 percent had used their smart phone to submit documents, like checks or receipts (only one was adamantly against). All of the participants reported their smartphones were up to date on software compliance. 92 percent of participants reported that they "never" share their personal account passwords and pins, compared to 41 percent of adults in a 2017 Pew report on Americans and Cybersecurity who reported they have shared a password to one of their online accounts with a friend or family member (Smith (2017)). The one participant who did not select "never" noted that they only share passwords with their spouse. 46 percent of participants reported that they "never" reuse the same passwords or pin numbers, 38 percent of participants "sometimes" reuse the same passwords or pin numbers (8 percent rarely do, 8 percent frequently do) (versus 39 percent of the general population who

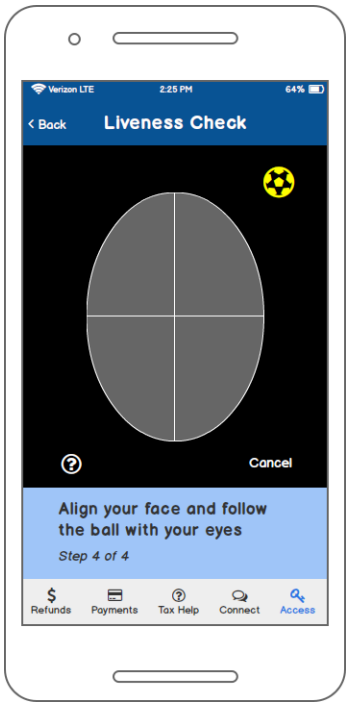
say that they use the same (or very similar) passwords for many of their online accounts, from the 2017 Pew report).

2.3 Procedure

In the interview study, participants joined an online Skype conference room with audio and screen-sharing enabled. After an overview of the study and an informed verbal consent, participants completed a short survey to capture basic demographics, their role as tax professionals, and experience with various identity proofing and authentication methods related to the usability test. They were also asked about their current IRS online services use.

After completing the survey, participants were introduced to a set of wireframes of an account creation workflow of a notional Information Assurance Level 2 (IAL2) customer journey. A website wireframe is a visual representation of user interface elements, for example, layout, content, and interactive elements. Wireframes are intended to communicate the functionality and workflow of a website or software. For the Secure Access study, a set of wireframes was created in a style that resembles a sketch using Balsamiq software. We used the sketch style in order to convey to participants that the design was notional. See Figure 3, Secure Access Wireframe, intended to demonstrate liveness testing.

FIGURE 3. Secure Access Wireframe



Participants viewed and interacted with the click-through wireframes online. They were asked to verbalize their thought process as they clicked through the wireframes with think aloud protocols (Ericsson and Simon (1980)). In addition, participants were asked to state aloud what they would do next or where they might click before advancing to the next screen.

Participants started on a wireframe version of IRS.gov, and then viewed the current wireframes styled after the current eAuthentication "create account" process with minor modifications to align them with the potential IAL2 process. Participants were asked to assume that to ensure security, the IRS was requiring that all individuals must be vetted to access IRS e-Services. Participants then viewed a mocked-up identity proofing flow within an illustrated version of the IRS2Go application. The wireframes demonstrated more indepth

features involved with using a selfie verification like liveness testing (voice, text, physical movement, read random string of words), with feedback on photo quality and multiple retries.

During the walk-though, the session facilitator asked open-ended, neutral questions such as: "what comes to mind" and "how might you respond at this point." They also attempted to reflect back the language the participant used, especially in reference to technological terminology. When the participant encountered new concepts of remote identity proofing, such as using a second device, uploading a State identification card and a selfie check with liveness testing, they were asked followup questions regarding their reaction, expectations and any concerns or frustrations.

After the usability walk-through, participants were asked to provide more indepth feedback on the wireframes in a semistructured interview. Questions were asked for further opinions on remote identity proofing using a smartphone, using a State photo identification card as a key document to use to register for e-Services, facial recognition, and how the IRS might better design initial notification and instructions on how to identity proof.

2.4 Data Analysis

We qualitatively analyzed participants' think aloud and interview responses. Our analysis occurred after completing all 13 sessions with participants. All 13 sessions were recorded and transcribed. We performed qualitative data analysis of the verbal transcripts using a grounded theory approach (Lazar *et al.* (2017)) in an Excel spreadsheet. We generated a set of codes in response to our three primary research questions: usability issues, how willing Tax Professionals are to remotely identity proof, and how the IRS might better design initial notifications and instructions on how to identity proof. A team of three researchers independently coded in one round of analysis, and then met to review and align themes.

2.5 Secure Access Results Overview

Our analysis showed that while skeptical of the new technologies, participants were willing to use more complex remote identity proofing with the IRS. Security was a primary concern of the participants, and some misperceptions, for example, the assumption that the liveness testing procedure captures and stores personal biometrics, fueled skepticism. In the following sections, we first discuss how participants responded to the new procedure. Next, we present potential usability and accessibility opportunities and concerns to address in design. Lastly, we provide insights into how to better communicate and instruct users on these new technologies.

2.6 Response to Remote Identity Proofing

The majority of participants, even those who expressed negative reactions to remote identity proofing, were willing to go through the process (and seven participants made positive statements on the topic). Attitudes ranged from an understanding of and expectation for higher security measures from the IRS…

> *"…Being a tax preparer it's something I've gotten used to … providing all that detailed information to the IRS."* (P05)

to expressing that they have no choice in the matter …

> *"Well, I don't have an option. I mean, what other option do I have if I need this service?…If I'm forced to do it, I guess we have to. But I don't want to."* (P02)

P06 was the only exception:

> *"I will tell you if I have to go through all that to use the IRS2Go, I won't use it. I will use my computer or use whatever other process that might be available where I don't have to go through all this."* (P06)

Security was a top concern for participants, and some participants (six) expressed that no assurances would persuade them that the registration process is secure. A number of concerns were expressed, such as the potential for a State photo identification card to be spoofed, security concerns with SMS, email and entering personal information into a mobile phone application and a fear that the IRS might get hacked. There was also

concern for the general public's security behaviors with their mobile phones. The majority of participants had never downloaded IRS2Go, a key application for the notional registration process. Three of the eight expressed reluctance to ever download the application.

> *"I'm not a big fan* [of downloading mobile apps]. *I think it's too easy to lose a phone, and even if I have it password-ed, I don't want to have a bunch of stuff on it. But, I know everybody else in the world does everything on their phone, so apparently I'm one of those old dogs. I like to do it on the computer, but I'm just not quite ready to move over to phones yet*." (P10)

Approximately half (seven) of the participants assumed the license upload and verification process was capturing their face and voice biometrics. This is likely due to the license verification process using interactions similar to registering biometrics on a smartphone. In addition to concerns that their biometric data were being captured, seven participants also assumed their image was being captured and stored, for example,

> *"There's no way. I'm not going to give you a driver's license, a static picture, and then allow you to match that to my face, because now I'm in a database for sure*." (P01)

There were mixed opinions expressed on the security and quality of facial recognition and use of financial data for verification. Eight of the participants made positive statements on the topic, and five negative. The positive statements included five participants who expressed that the remote identity proofing method was a better way to ensure someone's identity, and three who discussed their feeling that the IRS is a trustworthy institution when it comes to protecting their data. Five of the participants expressed willingness to using biometrics with the IRS, however, one participant had a very negative reaction to being asked to smile during liveness testing.

2.7 Usability and Accessibility Issues of Remote Identity Proofing

When introduced to the task of registering an e-Services account from scratch, participant expectations varied. Some expected the process to take approximately 5–10 minutes, while others anticipated 2, even up to 15, days to complete. One early usability issue observed occurred on the wireframe homepage of IRS2Go. The wireframe was modeled after the existing application at the time, which opened on the "Where's My Refund" page. This confused some participants, with four attempting to fill out the refund form before recalling and continuing the task of registering for Secure Access. While this may be due to participants being recently introduced to the task, it demonstrates that future use of IRS2Go for additional audiences and tasks requires reconsidering the design of the mobile application.

Usability concerns brought up by participants during the walk-through were primarily around identity proofing transactions, such as knowledge of the correct address to enter into the form, license quality, phone positioning during liveness testing and the utility bill. For example, the Tax Professionals who participated were familiar with the need to enter an address correctly when interacting with the IRS and expressed concerns around it for their colleagues and clients. Some also questioned what users would do if they had recently moved. And while five participants made positive statements on the ability to use a utility bill as evidence to confirm their identity, a few expressed concerns. Several pointed out that their utility bill is not in their own name. One was concerned that it would be time-consuming to get a copy of their bill since they do not live in an urban area, claiming it can take 15 days or more to get a copy of a bill (P02).

The key requirements of remote identity proofing illustrated in the wireframes were using your phone to scan the front and back of one's State photo identification card, then verifying you are the same person on the license using a selfie verification, and liveness testing to ensure you are truly holding the phone in that moment. Some participants were concerned about the status and quality of their ID, and whether they would have to do this on each login. However, 10 of the 13 participants made positive statements about using a license as evidence in identity proofing.

> *"My only concern would be the possibility of rejection if I don't have a good photo ID*." (P07)

Some participants requested the image of the license automatically focus and capture, preferring the ease of capture on a mobile phone, however six participants requested a desktop option for submitting documents and ID verification. In selfie verification and liveness testing, some participants were wary of the process.

Several participants questioned whether they could wear their glasses when verifying their ID, concerned that their glasses might prevent the application from completing facial recognition.

Participants expressed concern that Tax Professionals as a group are less likely and less willing to own a personal smartphone, primarily due to what they described as a large number of older tax professionals who are less familiar with the technology. In particular, the process of liveness testing, in which the user must react to instructions on a mobile device while positioning a phone in front of their face, was deemed by participants as inaccessible to their older or less tech-savvy colleagues.

> *"Like I said, I just think that the … is a little challenging. It's asking someone to be very attuned to technology and not everyone has that level of comfortability with technology to maybe go through the entire process."* (P12)

> *"The problem you have is the average age of a tax professional is 65, so I think you're battling age more than anything … So I think as long as you're dealing with people that are millennials, you're fine. You're getting into gen-X, gen-Y, they should be okay. But the boomers are done. They're not willing to do a lot of this."* (P01)

In addition, two participants stated that liveness testing as presented in the wireframes was likely inaccessible to them personally due to a disability that would make it very difficult to complete.

2.8 Communications on Remote Identity Proofing

Participants requested more information on what is being collected, why it's needed and how it is secured, however we did not observe many participants reviewing the detailed information provided. The wireframes themselves provided detailed notifications**.** We used the current eAuthentication step-by-step confirmation of documents required, and added an additional page confirming that the user has their State photo identification card accessible. A reminder of documents required was also included on the IRS2Go Secure Access login screen. During identity proofing, a dashboard listed each step of the process. A statement on the purpose of ID verification with the assurance that no images are captured prior to starting the process was also included. Some participants wanted more. They requested more guidance, clearer expectations, better lists of what is required and more assurance of why this is being done and the security measures around it.

> *"So I mean, apps are usually intuitive I find, so I don't know that I would need specific instructions on how to go through it but I did need instructions at the beginning that I didn't feel that I was getting as to what I need to do to obtain secure access. I didn't think it was … I mean, it told me about what I could get with secure access but it really, I didn't think it was quite clear to me as how do I get it."* (P06)

2.9 Secure Access Discussion

Two primary areas emerged in analyzing participant comments from the walkthrough and interview afterwards. First, clear communications around key parts of the remote identity proofing process are key to a security minded audience such as Tax Professionals. Delivering the content is a key design challenge, as users may choose to scan or ignore lengthy areas of text before or during the process. Second, while remote identity proofing offers the opportunity for more users to access IRS online services more easily from their own home or at work, there are many potential usability and accessibility challenges to uploading documents and remote ID verification that should be addressed.

2.10 Improve Communications

Clear communications on remote identity proofing will benefit from repeated user testing on the messaging style, ordering, and placement of document requirements, security, and instructions on document verification and liveness testing. Despite the time added to the workflow, displaying brief instructions when needed that must be dismissed by users may help ensure critical messages are seen, but regardless of the approach user testing will help to confirm how effective the design is. Participants asked for more, not less, information about the registration process. Provide plenty of assurance of the security of the process before and during the process

> *"I'd like to have a better understanding of why they're asking for this information, why it's necessary, maybe what they're cross-referencing, like a credit report. And then the security measures that they're taking to secure the information."* (P09)

2.11 Transparency About Data

A better understanding of how personal data are handled will help to alleviate concerns we observed from our participants. Clearly identify what type of data are being analyzed, and if transmitted, where they go and what happens to them. Only five of the participants understood the concept of liveness testing. Learn from the way people describe the process to craft clear, plain language descriptions. The IRS can also use the opportunity to clearly notify that no biometric data are being stored, since many participants assumed that they were.

> *"Okay, liveness check. Okay. Make sure I'm a living, breathing thing and not just an Android thing."* (P04)

2.12 Implications for Design

To address potential usability and accessibility challenges to uploading documents and remote identity verification, aim to provide as many options as possible. For example, provide a desktop and web camera option if available. Camera interactions may not be intuitive for all users. Make sure to provide instructions before and during the workflow process. Identify technologies that put less burden on the users in regard to camera positioning. More importantly, further explore the accessibility issues with document upload and camera interactions. Such interactions are currently not accessible to communities with a visual disability (ten Brink and Scollan (2019)). Lastly, provide more assurances on the facial recognition process—what if the license picture is from many years ago? Has your face or hair changed? Explicitly state the capabilities up front, and what the boundaries are for use; for example, whether people can wear scarves or glasses when in use.

The Secure Access study focused on an audience of tax professionals; however, the IRS must provide its Secure Access solution to all taxpayers. We were curious how the needs of a general audience differ from those of tax professionals. In our next study, we investigated what usability considerations emerge for a general audience creating an account in compliance with the NIST Digital Identity Guidelines.

3. CSP Usability Study

The CSP Usability Study sought to explore the implications of a third-party providing identity proofing and authentication for the IRS with a general audience of individual taxpayers, as well as a continuation of seeking feedback on new remote identity proofing concepts from potential users. A Credential Service Provider is a trusted third-party who provides a service for identity proofing, registering authenticators and issuing credentials to users. A CSP may be run commercially or by government but is beholden to the same NIST guidelines of a Federal agency to fall within the "trusted" category. For our CSP study, we developed a new prototype to offer participants a selection between two fictional CSPs, as well as display some of the same concepts of remote identity proofing from the Secure Access study.

3.1 Method

3.1.1 Method Overview

We conducted semi-structured interviews with 19 participants about their perceptions and comprehension of Credential Service Providers and used a prototype walk-through to identify potential usability issues. Participant sessions were conducted during January 2020. In the usability walk-through, we displayed a high-fidelity wireframe prototype with a NIST IAL2 flow and asked participants to "think aloud" while interacting with the prototype. This section first introduces the method and results of the participant interviews. Then, we discuss the implications of our results.

3.1.2 Participants

We recruited a total of 19 individual taxpayers who reported having owed a balance on their Federal taxes within the past 3 years. Interviews were conducted during January 2020. Participants were recruited through a professional recruitment firm, Fieldworks, with a recruitment screener provided by the research team. Sessions were approximately 1 hour long, and participants received a $75 incentive for their time.

Additional recruiting criteria included having U.S. citizenship, being age 18 years or older, having previously filed a Federal income tax return, not being representatives or employees of State or Federal Government agencies and not working in tax services. The participant mix was requested to be balanced across age and gender. Fieldworks collected information on participants' ethnicity, education level, income level, tax-filing experience, experience with online government accounts and services, and internet behavior. Internet behavior questions investigated how participants primarily access the Internet, if they tended to share personal account passwords or Personal Identification Numbers (PINs), and if they tended to reuse passwords or PINs.

The majority of participants expressed familiarity with the concept of a CSP. When asked "Have you used an account such as a Google or Facebook account to log into a different website or app, for example, Spotify, Medium, a mobile phone game, etc.," only two of the participants reported no such experience. Experience with mobile device transactions similar to those used in remote identity proofing was also high. Sixteen of the participants had reported using their smartphone to submit documents such as checks or receipts, and seventeen reported experience taking a selfie.

3.1.3 Procedure

MITRE developed a research protocol for a semi-structured interview and interactive prototype walk-through. In the interview sessions, participants joined an online Skype conference room with audio and screen-sharing enabled. On joining the online conference room, participants accessed the online prototype through a link provided by the interviewer. The prototype also displayed several survey questions at appropriate times to allow participants to both see and hear the questions during the interview.

After providing informed consent verbally, participants completed a short survey to capture baseline perception of CSPs. To set the context of using a CSP to access secured services such as what might be offered on IRS.gov, participants were asked to "…imagine [they] could use [their] online account at [their] bank to log into other sites or applications such as insurance, mortgage or credit cards. Places that contain sensitive information and where it would be important to know that it is really you that is logging in."

We also described the services as one "…where you use one account that protects your sensitive information to create a different account where it is important to know that it is really you logging in. In this example, you would be logging on through your online bank account in order to log in to a different site or application that is not part of your bank." We asked participants a series of questions with Likert scale or open-ended responses to better understand their trust and perception of information security and usability of using federated services, as well as their comprehension of how they work technically. The questions were then slightly modified and asked again after the walk-through, in which participants interacted with a notional CSP to gain access to an IRS Account. See Table 1 for the CSP perceptions survey used before and after the usability walk-through.

TABLE 1. Credential Service Provider (CSP) Perceptions Survey

Question before notional prototype	Question after notional prototype	Topic
I find services like these easy to understand.	I found [chosen option] easy to understand.	Comprehension
In a general sense (meaning, not specific to a certain company or organization), this method of logging in using a third party keeps my information secure.	The method of logging in to an IRS account using a third party keeps my information secure.	Trust Comprehension Info Security
In a general sense (meaning, not specific to a certain company or organization), services such as these are concerned about keeping my data secure.	A service such as [chosen option] is concerned about making sure my data are secure.	Trust Info Security
I find a service such as this useful.	I find a service such as [option] useful.	Usability
A service such as this is easy to use	A service such as [option] is easy to use	Usability
I would feel comfortable using a service like this in the future	Based on my experience using [chosen option] to log in to an IRS account, I would feel comfortable using a similar service again in the future, for IRS or other government websites.	Trust Willingness to use

Response option	Meaning
1	Strongly disagree
2	Disagree
3	Somewhat disagree
4	Neither agree nor disagree
5	Somewhat agree
6	Agree
7	Strongly agree
n/a	Unsure / Don't know
n/a	No opinion

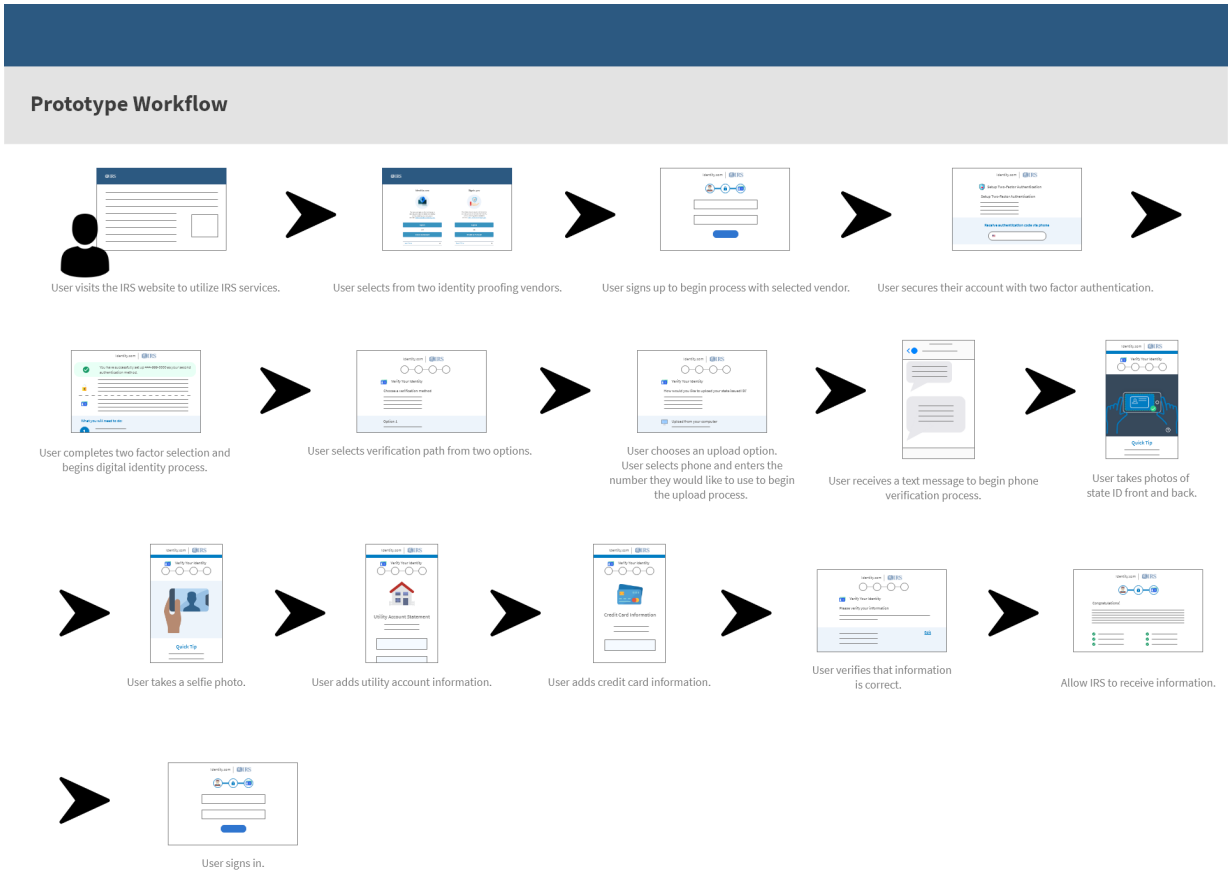
The response options were displayed visually to participants to help them see and rapidly respond to the spoken questions. The n/a options were included to more accurately capture the perspectives of participants who had not used CSPs before, especially in the pre-walk-through survey.

On completing the CSP Perception survey we introduced the usability walk-through by describing an illustrative task of checking a remaining balance and printing out a payment history from an IRS account. We asked them how they might accomplish the task if it was truly their goal that day, and then asked them to provide their thoughts on one potential way to create an online account to do the task online. Participants accessed the prototype on their personal device and shared their screen during the walk-through.

Participants started on a high-fidelity wireframe of the IRS.gov View Your Account page. The next page offered them a choice to select between two CSPs, one service that ends in ".com" and one that ends in ".gov." The motivation behind offering participants a choice was two-fold: first, to see if they have a preference for one over the other, and second, to see how users may react to being offered a choice. Screens for the remainder of the walk-through were the same for both the ".gov" and ".com" flows with these exceptions: the name and icon of the fictional CSP on the screens corresponded to the participant's chosen CSP, and the marketing-style splash image on one screen differed.

The prototype workflow and visual elements were modeled after current implementations of industry-leading CSPs, with a demonstration of setting up two-factor authentication and identity proofing, including some indepth features involved in remote identity proofing, such as taking photos of a license and selfie verification. See Figure 4, Prototype Workflow.

FIGURE 4. Prototype Workflow



Participants were asked to verbalize their thought process as they clicked through the wireframes with traditional think aloud protocols (Ericsson and Simon (1980)). At key points of interest, such as making a selection between the two CSPs or selecting primary evidence documents, the facilitator probed with a standard set of open-ended interview questions, described in Table 2.

TABLE 2. Usability Walk-Through Interview Questions

Topic	Questions
Select between Credential Service Providers to use to access the account	What did you choose? Why did you choose [descriptor] over the other options? Specific: What helped you understand what the options were and which to choose? Specific: Was there anything difficult about choosing a log-in option? Is there anything else you would have liked to know before making a decision? Optional: What do you take into account when choosing to use a service like this? What is most important to you? Optional: What do you think about one option being a ".gov" and another a ".com"? Did this play a part in your decision to choose [option]? Why? What do you think will happen next? How long do you think it will take?
Sign in	What would you do if you ran into difficulty or needed help? How do you keep track of your different accounts, usernames, and passwords? What do you think would happen if you did not need to login again for 1, 2 years from now? Do you think you would remember which option you selected?
Choose more verification methods	Why did you choose those verification options? What makes you un/comfortable with these options vs the other options?
Choose upload method	Have you used your smartphone before to submit documents, like checks or receipts? (yes/no)
Take selfie pic	What do you think is going on technically? Do you think the image is stored anywhere?

In addition to the planned interview questions asked during the walk-though, the session facilitator asked open-ended, neutral questions such as: "What comes to mind?" "How might you respond at this point?" and "How do you feel about [this page/the process you just encountered]?" They also attempted to reflect back the language the participant used, especially in reference to technological terminology.

After the usability walk-through, participants were asked the CSP perceptions survey modified to the illustrative CSP from the walk-through and were asked for more indepth feedback on the prototype through a semistructured interview. The interview explored topics such as help preferences, motivating factors behind willingness or unwillingness to use a third party to register for an IRS account, and specific prior experience with online IRS accounts.

3.1.4 Data Analysis

Audio and video of all 19 sessions were recorded. Recordings were edited to remove unnecessary details and then transcribed by a professional transcription service (Rev.com). Quantitative and qualitative data from researcher notes and transcriptions were captured in an Excel spreadsheet. We performed qualitative data analysis of the verbal transcripts of the open-ended questions and usability walk-through using a grounded theory approach (Lazar *et al.* (2017)). We generated a set of codes in response to our three primary research themes: user perceptions of CSPs, user comprehension of CSPs, and potential usability issues. A team of three researchers independently coded in one round of analysis, and then met to review and align themes.

3.2 Results

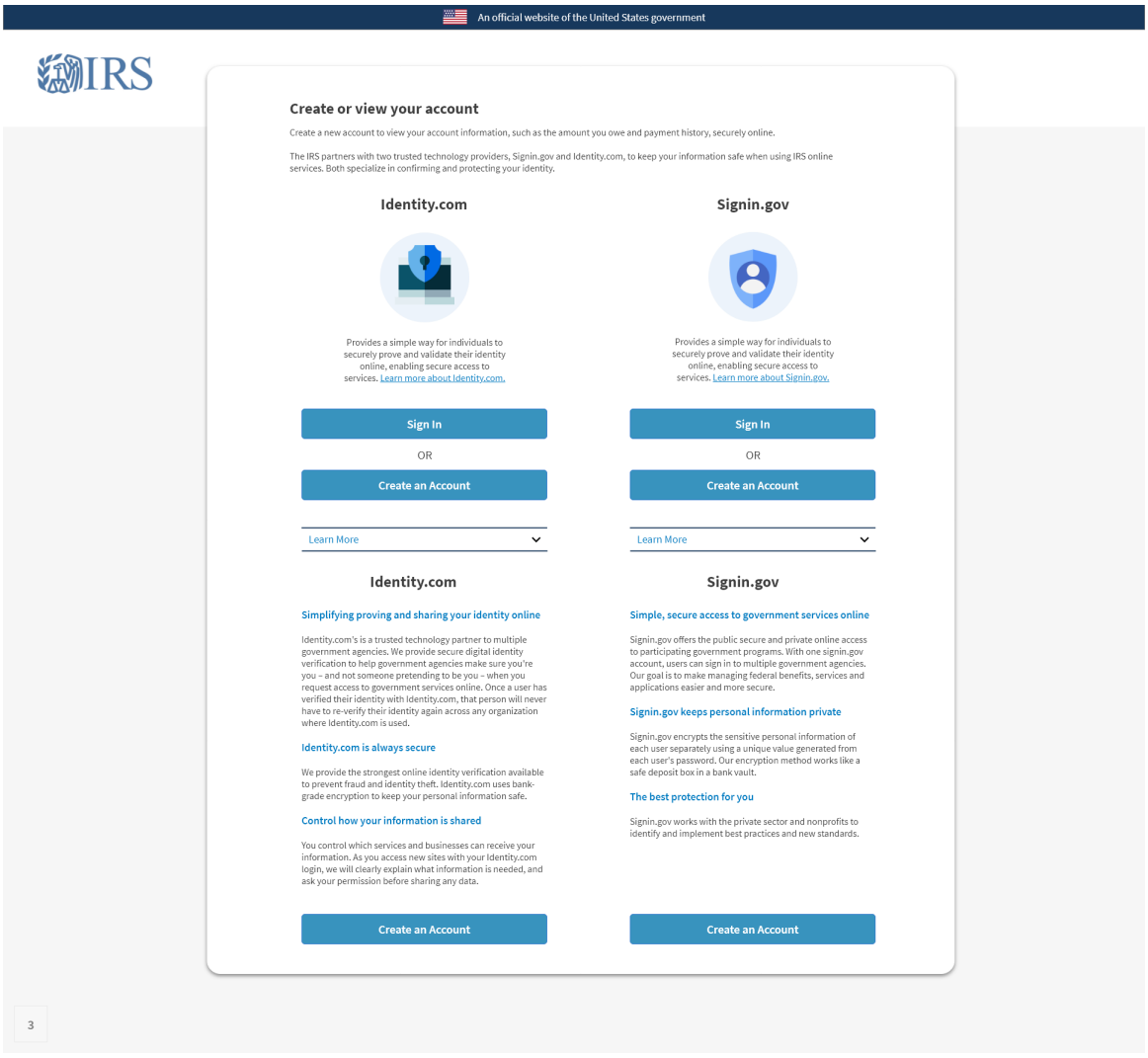
3.2.1 Results Overview

Our survey findings and comment analysis showed that participants are willing to use third-party Credential Service Providers (CSPs) to register for an online account, however they favor a ".gov" for both its familiarity and assumed security. Participants who selected the .gov did not necessarily understand it was a third party to the IRS, however, and many participants preferred registering directly with the IRS. In the following sections, we first discuss how participants perceive the use of Credential Service Providers with the IRS. Next, we present what users think CSPs do, and whether their comprehension improves after exposure to the prototype. Lastly, we show the response to document selection, upload, and verification.

3.2.2 User Perceptions of a Government Run Versus Commercial CSP

A key question we explored with the participants was their perception of government-run versus a commercial CSP. To capture their preference, the second page of the CSP prototype offered participants a selection between two fictional CSPs partnered with the IRS. See the wireframe with the two selections in Figure 5. After participants encountered the options and made their selection, we paused them to ask, "What did you choose?" and "Why did you choose [selected CSP] over the other option?". Approximately two-thirds of participants opted for the fictional Signin.gov service (63 percent).

FIGURE 5. Credential Service Provider Selection Page



We saw some split in CSP selections by income range and education level demographic groups. The mid and higher income groups leaned strongly to Signin.gov, while the lower income group leaned toward Identity.com. In this analysis, the lower income group is defined as a reported income of $50K or less, mid-income as $50K to less than $100k, and high income as $100K to less than $1 million. The master's education group largely chose Signin.gov, while most other education groups were close to evenly split. We also looked at the selection from the perspective of participants' experiences with online government and online IRS accounts. Participants with no prior online government accounts tended to choose Signin.gov (two-thirds of participants). Participants with prior government online accounts were near-evenly split.

FIGURE 6. Credential Service Provider Choice by Income Range

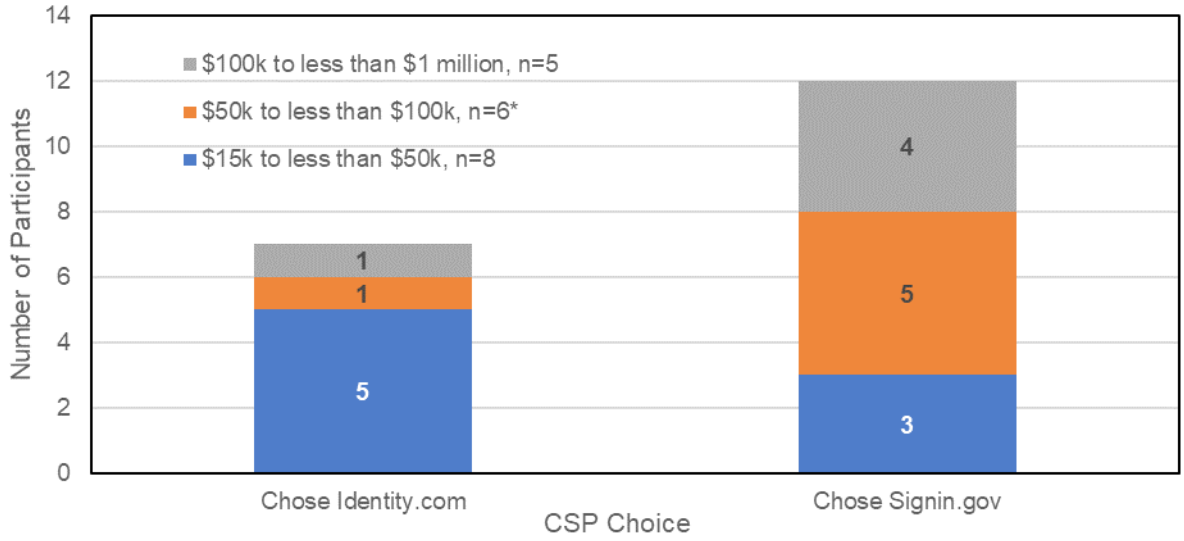
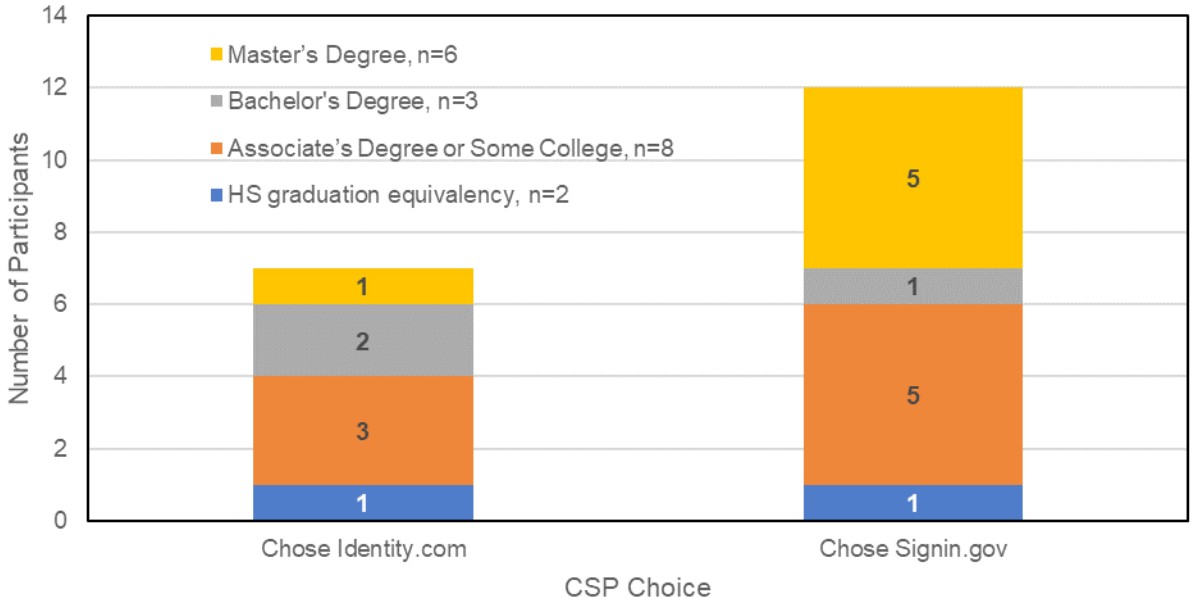
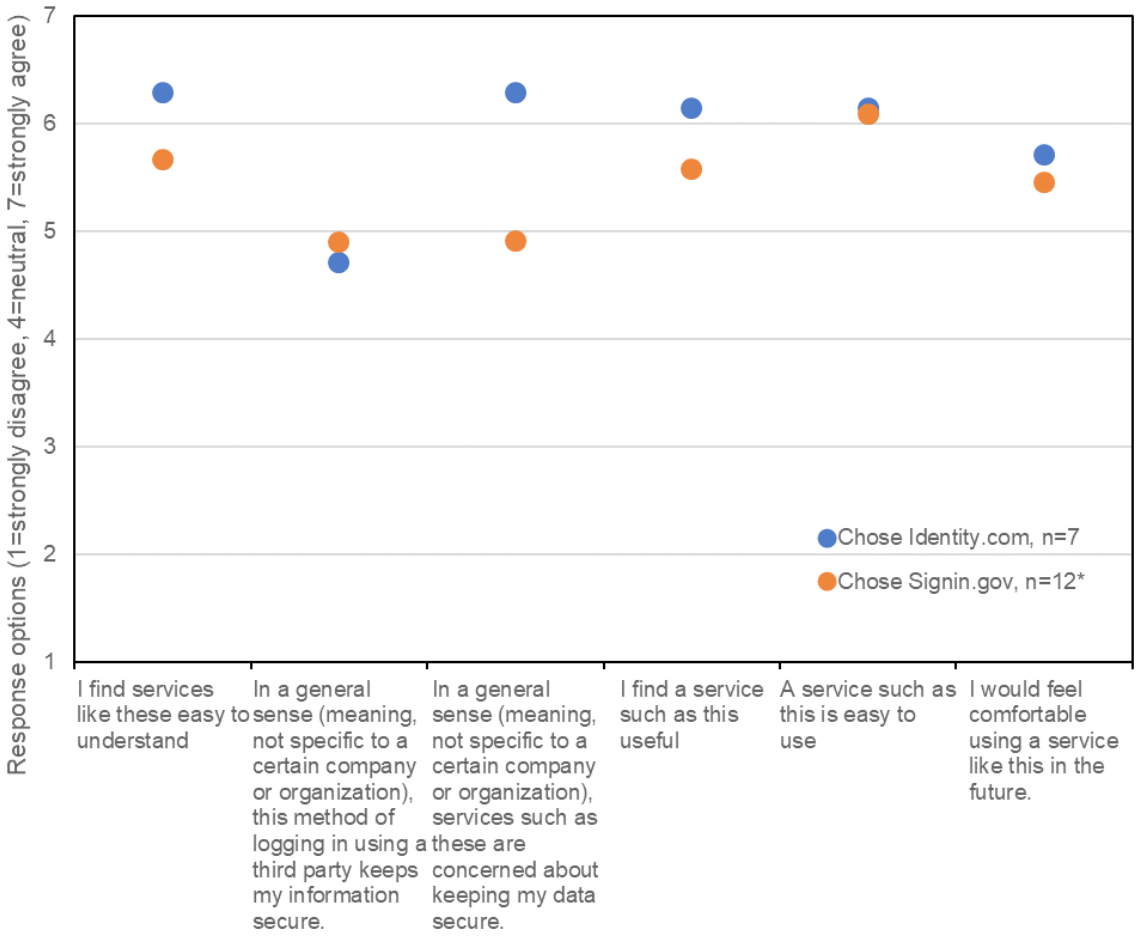


FIGURE 7. Credential Service Provider Choices by Education Level



Trust in CSPs, specifically perceptions of the CSP's priorities, may impact CSP selection. In the CSP Perceptions survey before the prototype walk-through, participants who selected Signin.gov reported lower average trust in a CSP's concern about keeping its data secure (close to "Somewhat agree") than did participants who selected Identity.com (close to "Agree"). Both groups showed middling trust that the methodology of how a CSP works keeps their information secure (close to "Somewhat Agree").

FIGURE 8. Average Credential Service Provider Opinions Before Notional CSP Experience



In an analysis of the comments from participants when asked "Why did you choose [selected CSP] over the other option," we found that five out of the seven participants who selected Identity.com made statements indicating they did not hold a strong preference for either the .com or .gov option. On the other hand, only one of the twelve participants who selected Signin.gov did not indicate a preference (P02).

> *"It doesn't really seem either one is too different from the other…It was more so on a whim. Like I said, they're not too different from each other. Left versus right, I just went left…"* (P06, selected Identity.com)

When probed further, the seven participants who selected Identity.com cited reasoning such as improved usability, stronger identity protection, the ability to use existing credentials and familiarity in general with "things with .com on the end" (P17). There were also assumptions that a commercial entity is less likely to be targeted by hackers than the government, as well as a sentiment that if the commercial entity is trusted by government, that it is secure. Only one participant (P14) did not cite any reasons for their selection.

The participants that selected Signin.gov perceived the option as more secure, more direct to the Federal service they were tasked with accessing, and more familiar. Eight of the twelve participants that selected Signin.gov stated a government option is safer. Participants spoke of a government option as "protected" and "more secure." It's important to note that participants perception of Signin.gov providing more "protection" may have been affected by a misunderstanding that Signin.gov was not a third party to IRS. In analyzing comments, it was sometimes unclear whether participants were able to distinguish that the service was operated by a government entity separate from the IRS. Some participants noted a .gov option is more "direct," "correct," or "official." Familiarity was also a reason cited for selecting the .gov option. Two of the participants who selected Signin.gov did not provide a reason why (P02, P03), one of whom was confused by the two options (P03). Finally, one participant (P18) did not want to mix a .com account with their "official government business."

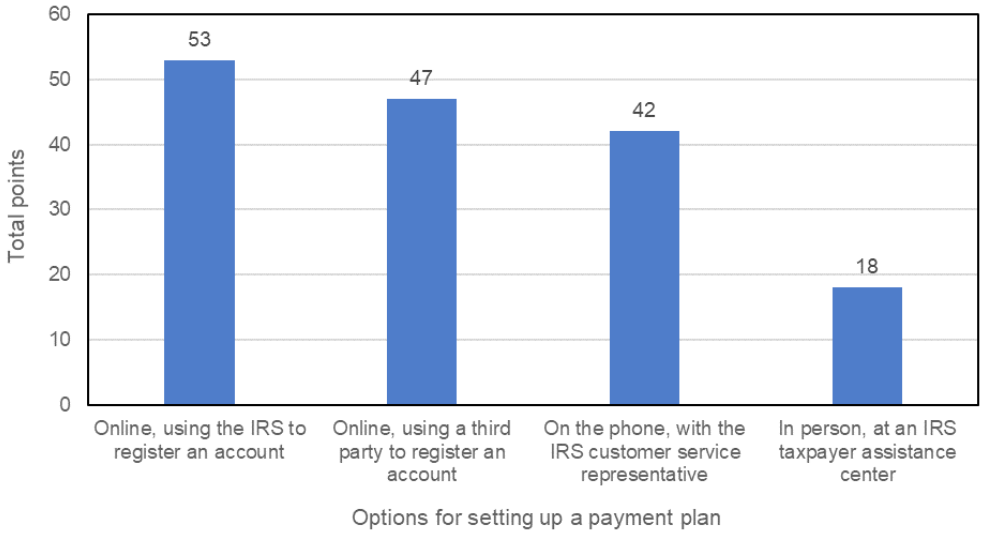
> *"I guess, at first glance I'm more comfortable with a .gov than I am with a .com … It seems more direct to what I want to do."* (P04)

> *"Just because of the .gov. There's absolutely no other reason. It just seems more official … Yeah I have absolutely no idea which one would be the one to go with, but I'm drawn towards this one just because it has the .gov instead of the .com."* (P20)

> *"I've never heard of identity.com before."* (P08)

Despite comments from participants expressing willingness to use a third party to identity proof and authenticate into an IRS online service, there was still a strong preference amongst the group to work directly with the IRS. Confusion about Signin.gov may have affected this: eight of the participants very closely associated Signin.gov with the IRS, considering it a part of the IRS or "less of a third party" (P08). After experiencing the prototype, participants were asked: "If you found yourself in a scenario like we described in our task today…what would be your priority of the following options to set up a payment plan?" Responses were weighted as the following: rank 1 = 4 points, rank 2 = 3 points, rank 3 = 2 points, rank 4 = 1 point. Responses were then summed to produce a weighted ranking. Nine participants (approximately half of participants) expressed a preference to work directly with the IRS.

FIGURE 9. Weighted Rank of Payment Plan Options



Weighted rank response to: "If you found yourself in a scenario like we described in our task today, setting up a payment plan so that you can pay off federal taxes owed in small payments over a year, what would be your priority of the following options to set up a payment plan?"

Responses were weighted as follows: rank 1 = 4 points, rank 2 = 3 points, rank 3 = 2 points, rank 4 = 1 point.

Note: Higher score is a higher ranking

3.2.3 Usability Concerns to Offering More Than One CSP

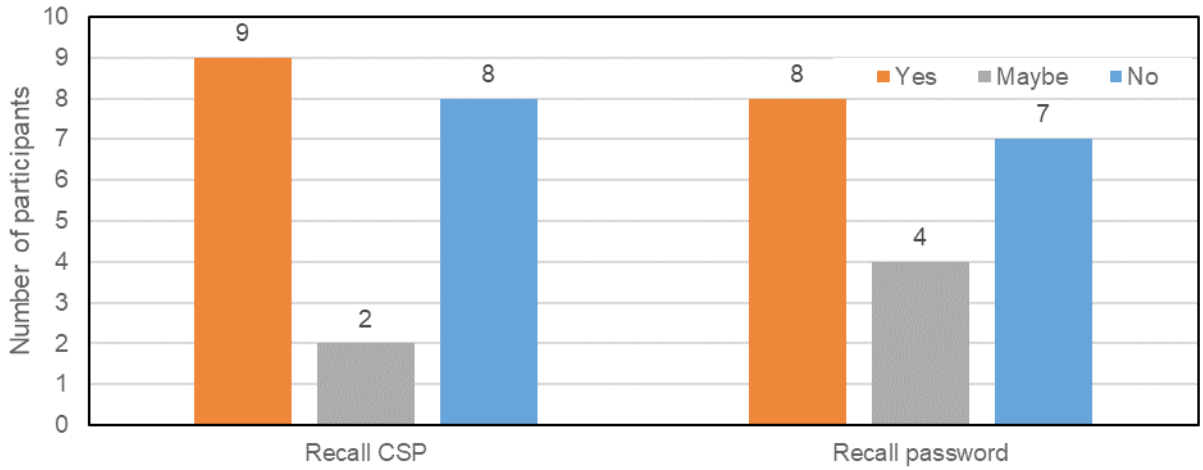
In addition to understanding how the participants perceive government-run and commercial CSPs, we also looked at comments made during CSP selection to identify any potential usability concerns. We found that when presented a choice to select between two CSPs on IRS.gov, 9 out of the 19 participants made statement we coded as confused or negative.

> *"… like I said, it confused me right away. I'm like, why are there two sign-ins and two creates? I didn't know which one to choose … From the very beginning, I'm like, okay, where am I supposed to go?"* (P03)

Three of the participants who responded in such a confused or negative way asked to be shown the differences between the two more clearly. Six of the participants made comments indicating they relied on logos, headers, or the name of the CSP to make their selection. One participant (P02) did not make any confused statements, however, did state the CSP offered access to a different IRS service, a misunderstanding of the service.

Once an account is created, we wondered whether users would be able to recall their selection and account details. Individual taxpayers may not have a need to access IRS online services frequently, resulting in a higher likelihood that users will lose or forget their account details. While it is beyond the scope of this study to determine whether users can truly recall their selection at a later date, we did ask participants, "*What do you think would happen if you did not need to login again for 1, 2 years from now? Do you think you would remember which option you selected?"* Participants were split on whether they would remember their CSP selection and username and password in the future. We followed up by asking what their strategies to recall their selection might be, which included strategies such as using the "forgot password" functionality, signing into both CSPs using their recalled username and password, reusing familiar passwords, and researching solutions on Google.

FIGURE 10. Participant Estimations of Whether They Would Recall



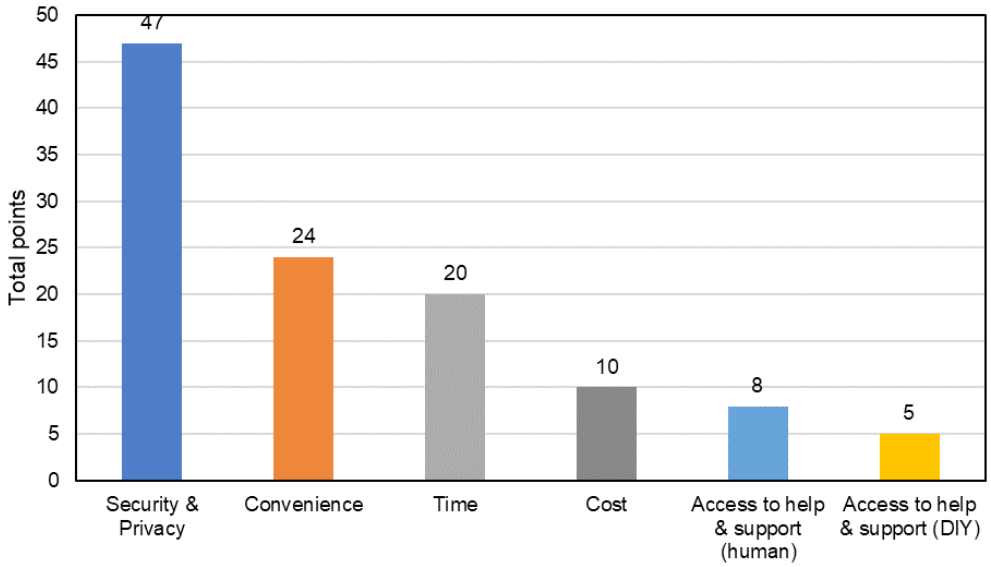
Note: If a participant referred to only Credential Service Provider or only password, their response was recorded for both CSP and password.

3.2.4 Factors That Affect Willingness to Use a CSP

Participants have choices when using government online services. If they encounter an issue, they can try to troubleshoot using information available online, call a help phone line, or opt out of using the service. Frustration with online technical support can lead to expensive calls to customer services (Konkel (2018)) or tax noncompliance (as a result of opting out). Additionally, there is legislation calling for the IRS and other Federal agencies to improve their online services. Understanding citizens' motivations and improving incentives to move to online services are critical. Motivated by this idea, we asked participants to rank their top three options for the question, "What is most important to you when getting your IRS balance and payment

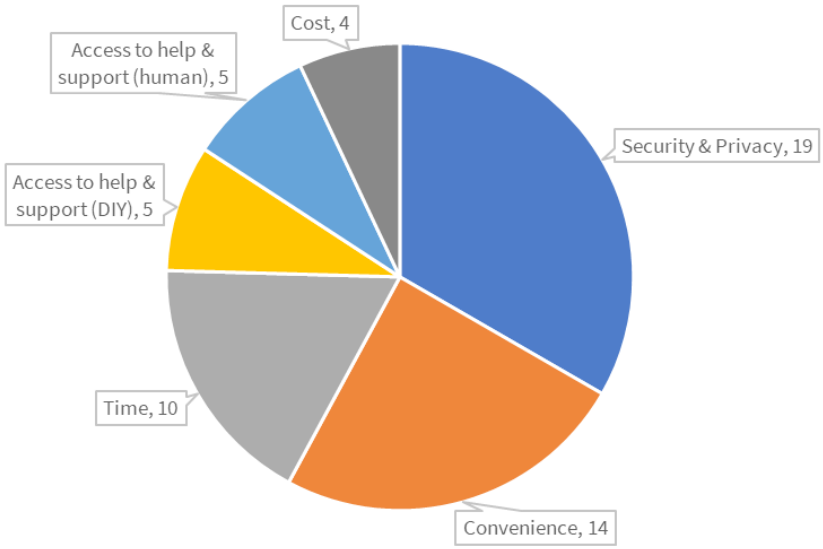
history online?" after the usability walk-through. The options offered were time, Cost, Convenience, Time, Access to help and support (do it yourself), Access to help and support (human), Privacy, Security, and Other (free response). In analysis, we weighted the responses as follows: rank 1 = 3 points, rank 2 = 2 points, rank 3 = 1 point, no rank = 0 points. We chose to combine Security and Privacy in analysis because participants expressed confusion at the difference between the two terms, often combining them in their response. Overall, participants ranked Security and Privacy highest for the task of accessing personal information online with the IRS, followed by Convenience and Time. Security and Privacy occurred 19 times in top-3 ranking responses. Convenience appeared 14 times, and Time, 10 times. The remaining 3 factors each appeared 4 to 5 times. These findings suggest that future IRS tools should be designed to reassure users they are secure and private. However, Convenience and Time are still important to users and should not be highly compromised.

FIGURE 11. Weighted Rank of Factors Affecting Online Payment-Related Queries



Response to "What is most important to you when getting your IRS balance and payment history online? Please rank your top 3."

FIGURE 12. Number of Times Each Appeared in Top 3 Factors



To better understand how the participants defined convenience and time, we reviewed comments made during the usability walk-through and open-ended question responses and coded positive or negative statements made on both subjects. Comments were grouped by subject and counted. The subjects that participants cited as a convenience are shown in Table 3, and subjects deemed either time-saving or time-consuming are in Table 4. The primary concerns for this group of participants were: easily accessible identity documents; the ability to conduct their business from home; and technical features that improved the ease of providing identity evidence.

TABLE 3. Conveniences Cited by Participants

Number of participants	Convenience cited
9	Use of identity documents that are easily accessible
5	Mobile phone document upload
8	Conducting business at home without need to travel
3	Not needing to remember passwords
3	Features such as auto-populate and text extraction
3	Use of passport for identity verification (not requiring additional documents)
2	Concept of a Credential Service Provider (accessing multiple sites)

Also mentioned: password rule transparency, the number of steps of account creation, not waiting in line, mobile phone options (text 2FA), ability to select from multiple identity documents.

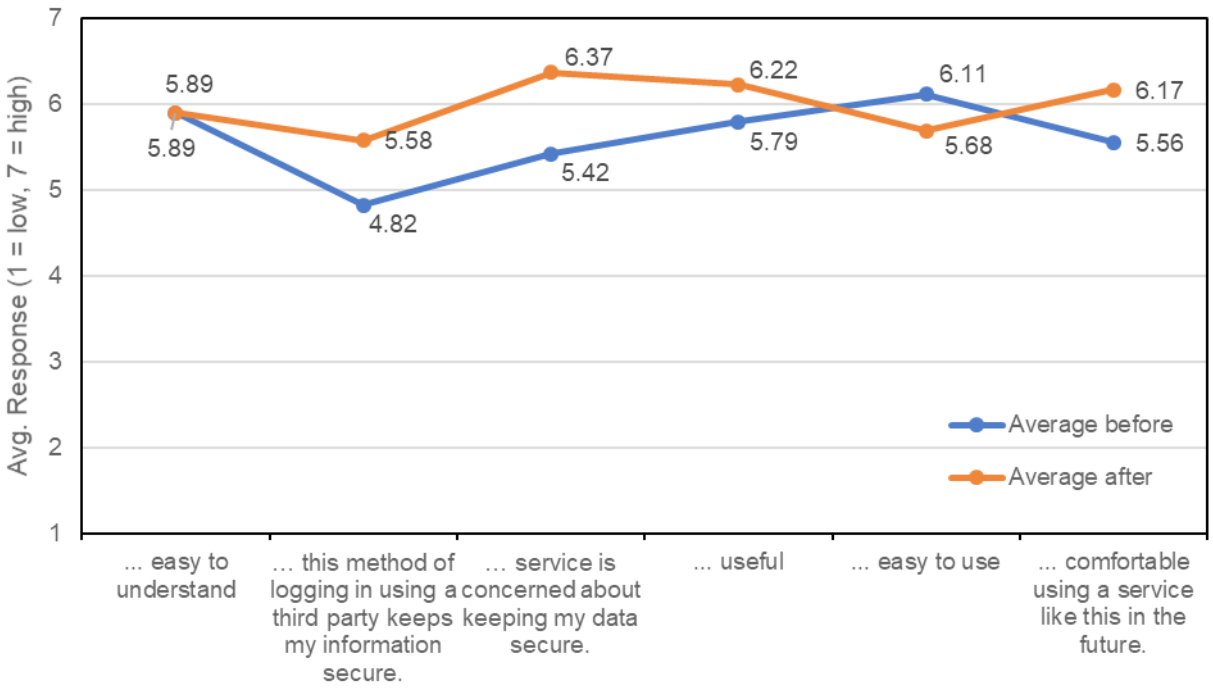
TABLE 4. Time-Saving or Time-Consuming Topics Cited

Number of participants	Time-saving / Time-consuming
3	Number of steps to create an account—too time-consuming
2	Number of steps to create an account—more secure (positive)
2	The trade-off between learning more about options and making a quick selection
2	Calling the IRS takes too long
2	Time is too valuable to spend creating accounts
1	Willingness to travel (familiarity of post office is faster)
2	Not willing to travel (to an IRS location or post office)

Also mentioned: assumed identity.com is faster, online video call is faster, determining the fastest identity document, concern that switching between devices slows technology down. "We're waiting for your photos. Click on the link, follow the instructions. Automatically direct you to the next page once your photos are received. All right. It's getting a little cumbersome…" (P11)

We also wondered if directly experiencing a CSP would change participants' willingness to use a CSP. To better understand if experience affects willingness, we looked at the responses to the CSP Perceptions survey before and after the usability walk-through. We found that after experiencing the prototype, participants felt higher trust on average in CSPs. In particular the item "keeps my information secure," "concerned about my data security," and "comfortable using in future" improved. However, while perceptions of usefulness increased, their perception of ease of use decreased. This may be due to the walk-through being slower than many participants expected and the surprise seen in their comments about identity verification requirements. Perception of understandability remained constant.

FIGURE 13. Average Credential Service Provider Opinion Responses Before and After Walk-Through



We also found that less-willing participants became more willing after interacting with the CSP prototype. Among the nine participants who were less willing to use a CSP, based on their response to the statement "…I would feel comfortable using a similar service again in the future…" on the CSP Perceptions survey, seven of them increased their score after interacting with the prototype. Only one participant out of the total group that expressed willingness to use a CSP prior to viewing the prototype changed their view after use (P10, who wanted more detail on how their information is handled by the CSP). Those who were less willing to use a CSP expressed a desire for more explanation and assurance, and/or more typical security practices like email two-factor authentication, security questions, and notification to change the password regularly. Some simply did not know what would make them more comfortable using a CSP. Only one of the unwilling participants cited the selfie as a reason.

TABLE 5. Pre- and Post-Task Comfort With Using a Credential Service Provider

Pre-task: "I would feel comfortable using a service like this in the future."

Post-task: "Based on my experience using [chosen option] to log in to an IRS account, I would feel comfortable using a similar service again in the future, for IRS or other government websites."

Response options	Pre-task count	Pre-task participants	Post-task count	Post-task participants
(5) Somewhat agree	5	1, 5, 8, 13, 19	2	5, 10
(4) Neither agree nor disagree	3	3, 12, 14	1	12
(3) Somewhat disagree	1	10	0	n/a

The average responses between different income range groups tended to align more after the experiencing the prototype than before, as shown in Charts X and Y. Experiencing the prototype seemed to bring perceptions between these different groups closer together. It may be that users in different income-range groups had different background knowledge, experiences, and/or expectations before the walk-through, but receiving more information (through experiencing the prototype) brought attitudes closer together. Our findings indicate an opportunity to use more effective communication and education about CSPs to compensate for different experience backgrounds, and to give users more realistic expectations and attitudes before they employ the technology. This could potentially enable users to make better-informed decisions as well as experience less confusion and disappointment during the experience.

FIGURE 14. Average Credential Service Provider Opinion Responses *Before* Walk-Through (By Income)

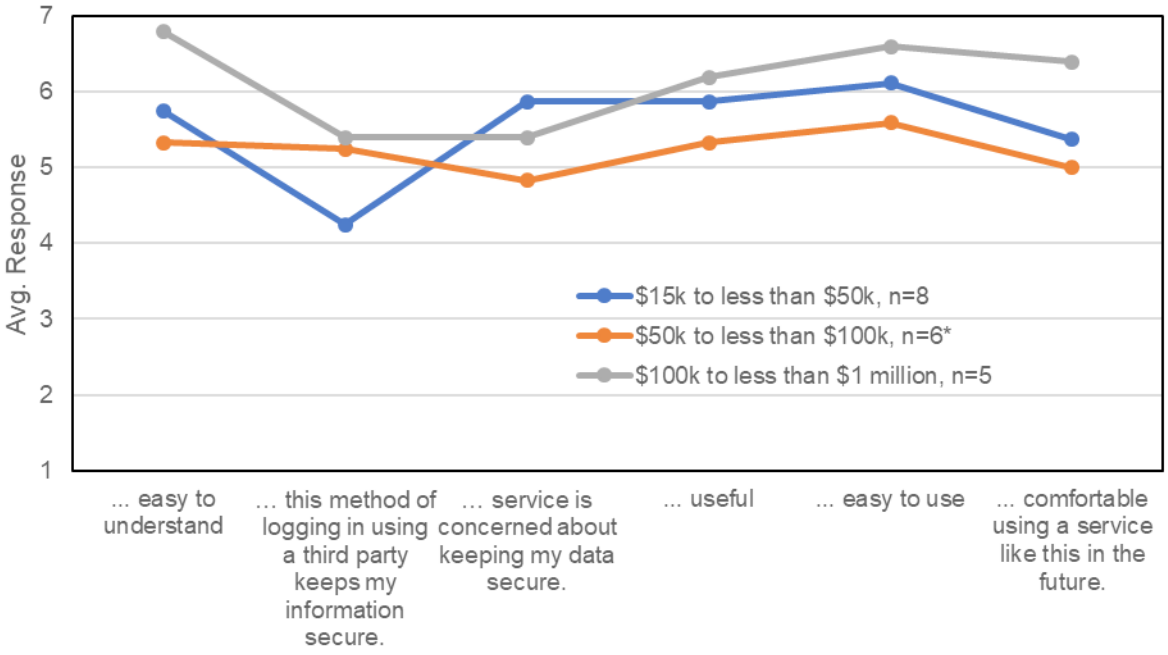
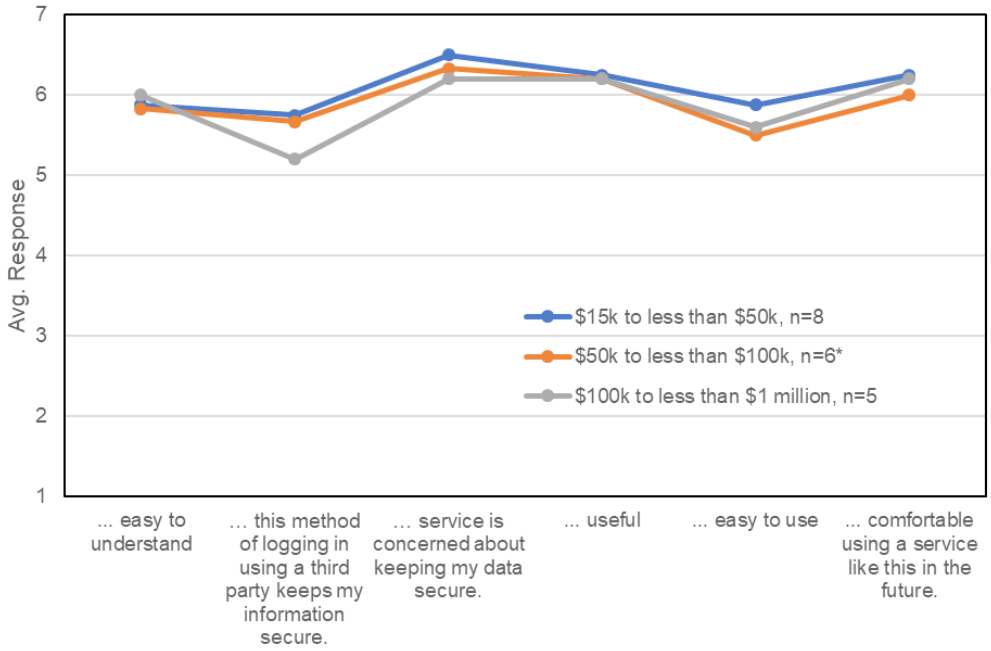


FIGURE 15. Average Credential Service Provider Opinion Responses *After* Walk-Through (By Income)



3.2.3 Comprehension of CSPs

Identity management using federation and Credential Service Providers are challenging technical concepts to understand, but some comprehension may be key to citizens willingness to use such services. A Credential Service Provider is a trusted entity that performs enrollment and identity proofing of an applicant. It issues a credential to the user after successful verification. But what do participants believe that a CSP does with limited exposure? To better understand this, we asked participants "*When you use a service like this to login or create an account for a different service, what do you think is happening?*" before and after the usability walk-through. We first asked two MITRE Identity Credential Access Management SMEs (subject matter experts) what would be considered an appropriate, high-level understanding of what a CSP is. Acceptable responses were deemed to be that a CSP verifies the applicants' identity and provides a credential so that the applicant may authenticate into the service. The CSP can also act as a single credential to multiple services or websites. We analyzed responses to "…*what do you think is happening?*" and coded and counted any comments that aligned with the basic activities of a CSP. See counts of the coded responses in Table 6. We also identified and counted several misperceptions, such as a CSP providing a dashboard that pulls services from other organizations, and a general lack of understanding ("I don't know"). Overall, based on the qualitative analysis, we feel that comprehension of what a CSP does is low, but improves after use. For example, more people understood that a CSP verifies their identity after using the prototype, and less participants stated they did not know what a CSP does. The number of people who stated that a CSP authenticates them into a service dropped; however we attribute this to not showing participants the login after completing identity verification in the prototype (the additional pages were omitted for time).

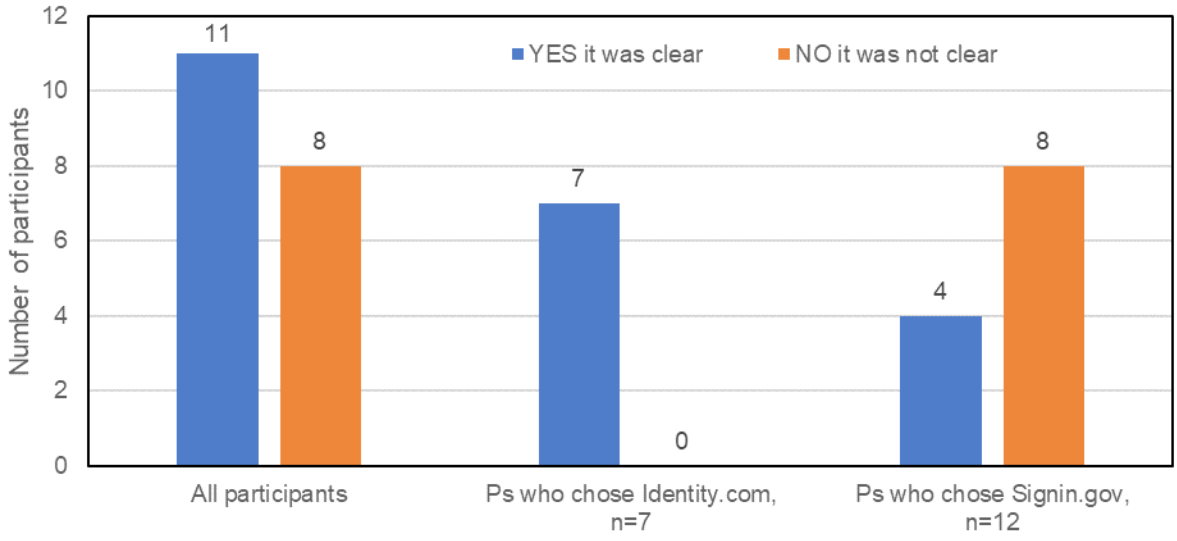
> *"I was logging into my account for identity.com, it has all my information so they can verify so they know it's really me. … I would figure I was logging onto my identity.com account that has all of my information that verifies who I am and that it's really me … [What did you think was going on technically?] Basically, this information that identity.com is collecting is going to be stored, so if I need to use, if I have to create another login to another government sites, I can." (P01, who cites logging in (authentication) and identity verification)*

TABLE 6. Coded Response Count for "When you use a service like this to login or create an account for a different service, what do you think is happening?"

Coded response	Before using prototype	After using prototype
'I don't know'	5	3
Verifies my identity	6	11
Authenticates into service	6	4
Stores or contains PII	4	2
Provides access to multiple services	1	3
Acts as a dashboard to multiple services	2	1
Auto-populates form fields	1	0
Acts as a single credential	4	1

Another important concept is that the CSP is a trusted third party to the organization offering the service. We asked all participants, "*Was it clear that [their selected CSP] is a third party to IRS?*" after the usability walk-through. Two-thirds of participants who chose Signin.gov replied "no" to the question, while all participants who selected Identity.com understood its third-party status. We feel that ".com" is a clear indicator that the CSP is a third party to a Government entity, and participants are less clear on differing entities within the Federal Government itself. After hearing their response to whether it was clear their selected CSP was a third party to the IRS, we asked participants, "*What do you think about using a third party to log into an IRS account?*" The majority of participants (16) made positive statements about using a third party, but surfaced concerns such as increased assurances of security and privacy and more transparency on how personal information is handled.

FIGURE 16. Participants' Understanding of Whether the Selected Credential Service Provider Was 3rd Party to IRS



3.2.4 Document Selection, Upload and Verification

During the usability walk-through, we asked participants to think-aloud as they encountered each page of the prototype. We also asked each participant several standard questions regarding their expectations at key moments of their interaction with the prototype. We analyzed the notes taken during the sessions as well as the

transcripts of the walk-through for potential usability issues users may face when using a CSP similar to what we presented. The prototype itself was not interactive, so usability issues identified are limited to user expectations and any positive or negative sentiment expressed throughout. Based on our review, we found a need to provide more information up front to set clear expectations, as well as improved layout and content around setting up two-factor authentication and selecting identity evidence.

Most participants expected a quick experience setting up their account with IRS, estimating 5 minutes or less when asked, "*How long do you think it will take?*" after CSP selection. Participants were not aware that they would be asked to identity proof or did not know what identity verification would entail. Many participants made comments that indicated their expectation was for account creation only, or that they expected verification through "something you know" such as Knowledge-Based Authentication using financial information.

> *"I put in the email address and the password, it'll just send an email to my account, so [it will take] as long as it takes me to open up my email account, take that password and type it in the next box. I don't think it would take long at all."* (P10)

The workflow prototype demonstrated the following workflow: create a password (first factor), select a second factor (with text message and phone as the default options, and other options listed lower on the page).

In the prototype, participants first viewed a screen on which they set up a password, the first factor of authentication. Once created, they are asked to set up second factor authentication. Text message and phone are offered as the default options, with additional options such as a code generator and a security token offered below. Most participants (14) selected text message. When asked, participants cited prior knowledge as the primary reason for their decision to select text message. Participants who selected the code generator (2) said they interpreted the offering description to mean that code generator was "something that's more secure," and made their decision based on that, "to add an extra layer of security." Device access, likely coupled with familiarity, led to phone call and security token selections.

After setting up two-factor authentication the prototype walks participants through two tiers of selecting identity evidence documents. Participants first selected between a Passport, Real-ID compliant Drivers' License or ID card, Permanent Resident Card and a Uniformed Services ID. Participants made a selection aloud and were then informed that the License was the selection option in the prototype. Ten participants selected the license, and eight selected the passport. One participant said they would use a passport or a military ID (P04). In a review of answers to the question, "Why did you choose those verification options? What makes you un/comfortable with these options' vs the other options?" we found that participants factored in access to documents, ease of fetching the document, perceived security, and the ease of the process. Some participants remarked unprompted that the options were reasonable (P04, P18). No participant said they wanted another option. We included a requirement that a Driver's License required two additional forms of identity evidence documents. The prototype offered a Credit Card, Bank Statement, School ID Card with Photo, and a Utility Account Statement. Several participants thought they needed to select only one document until prompted by the facilitator for a second choice. The majority of participants selected a Credit Card (12) and a Bank Statement (11). Eight selected a utility statement, and one selected a School ID, however noted it was several years out of date. Ease and convenience, a participant's perception of the data security for the document, and confusion over requirements all affected their selections.

After selecting their documents, participants were asked to choose between uploading their license on a mobile phone or a computer. The majority of participants (14) chose to upload using a mobile phone. In general, participants seemed to consider the options typical and acceptable. Convenience was the main driver behind their selection**.** During first-tier document selection, three participants explained their document selection choice was influenced by already having digital documents on their computer.

Participants were informed that the prototype used the mobile phone upload option after making their selection. The prototype demonstrated switching from computer to mobile phone by first entering a mobile phone number, receiving a text with a URL, clicking the URL that then opens a web browser page to continue the process. After viewing the prototype demonstrate taking a photo of the front and back of a license, participants landed on a screen requesting "selfie verification." Selfie verification is facial recognition used to confirm

that the applicant matches the photo identification card. The majority of participants (16) were comfortable with selfie verification. We saw two primary attitudes in our comment analysis: either approaching it in a matter of fact manner or making positive comments. The three participants who did make negative comments expressed willingness to use it despite their concerns.

We noted several concerns and misunderstandings of selfie verification. Some participants (five) expressed concern that their license photo was out of date, mentioning beards, weight loss or gain, and age. One participant claimed their license photo was 20 years old. Some participants (five) voiced reluctance to use selfie verification over concerns on their current appearance.

> *"…people get IDs taken at different points in their life. For example, I have a large beard, but had I not had that beard I looked totally different than I may have on my ID when I got that picture."* (P05)

> *"I hate to admit this, but I don't like pictures of myself so I don't want to have to take the selfie of myself, especially, if my hair's a mess and just don't look very—I don't look my best, so I cringe at having to take a selfie of myself.* (P01)

A few participants expressed concerns about the ability to fool the selfie verification process using a photograph**.** One participant (P12) claimed they would subvert the verification process themselves by putting a peace sign in front of their face or holding up a magazine image of a model. One misunderstanding expressed by one participant (P14) was that the selfie verification would be used to create a profile photo for their account**.** We also asked participants, **"**Do you think the image is stored anywhere?" based on the misunderstanding expressed in the Secure Access study. Approximately 12 participants felt their selfie image would be stored in some way for use in future verification. In addition, there were approximately 11 references to facial recognition when viewing the selfie verification process. Reactions ranged from approval of the higher security to skepticism of government surveillance and inconsistencies in software performance for different skin colors or features.

3.3 Discussion

Our study suggests that taxpayers will accept a CSP to register for an online account with the IRS, however communications and design will be critical for user satisfaction. Ultimately either a government entity or commercial entity may work, however a government-owned CSP has the advantage of familiarity and trust to new users. Excellent usability and an expanded list of options for identity proofing will aid in user acceptance.

3.3.1 Willing to work with a third party

Most participants did ultimately choose creating an account directly with the IRS over other options when asked in the post-walk-through interview, but due to the heightened average responses on trust, comprehension, and satisfaction after viewing a notional CSP, as well as participants' stated willingness to use a third party, we feel that users will accept using a CSP with the IRS. Almost two-thirds of participants selected the government-run CSP. We saw differences in selection of commercial versus government-run CSPs by income range. However, due to our small sample size, we do not offer this as a finding and instead suggest this topic as a future research question, especially if online IRS account services are intended to cater to specific income demographics. Generally, preference trends by demographic (income, education, experience) could be used to target communication and advertising to specific communities, or to adapt messaging for specific communities, to increase adoption of CSPs for online government services. However, our sample sizes for demographic groups were small, so further research should be conducted to investigate trends in .gov versus .com preference.

Many participants viewed a government-run CSP as inherently more secure than a commercially run option, and those who were less trustful of CSPs preferred the .gov option. But while our analysis of the survey and comments found that a .gov is often considered more secure than a .com, participants did not necessarily understand that the .gov was a third party to the IRS. Ultimately, we believe a commercial entity will need more time and clear communications to build trust and familiarity with users of IRS.gov but will be quickly

recognized by users as a third party. IRS.gov users will strongly associate a ".gov" CSP option with the IRS and the government as a whole but will be less likely to recognize it as a third party.

3.4 Choice can degrade user experience

Half of the participants had a confused or negative reaction to being offered a choice between two CSPs. Some also decided based only on the CSP name or logo; this is worrying for what should be an important decision for a sensitive online procedure. If the IRS were to ever provide a choice between two or more CSPs, the choice presentation would need to be very carefully designed to improve user acceptance and satisfaction and be understandable to users.

The user experience of account recovery becomes critical if more than one CSP is offered. Online services offered on IRS.gov vary in frequency of use. Tax professionals are frequent users, sometimes visiting daily, but individuals may only visit once per year or less. Infrequent users are more likely to lose their account details, including which CSP they selected.

3.4.1 Image verification breeds confusion

Selfie verification was generally accepted and understood, but we believe there were enough questions and concerns to motivate the IRS to ensure the experience is well explained, makes use of usability best practices and offers technical support for errors and concerns. Since it is a new technology for many users, it is also important to be very clear on what kind of input is acceptable. Participants were concerned that having a slightly different appearance than that on their photo ID, such as having a new beard, glasses, or a change in weight, would affect the accuracy of the verification. If the facial recognition system is not accurate enough to compensate for variations like these, then such concerns should be addressed and clearly explained up front. Understanding why failures happen may help prevent users from abandoning the task when they encounter issues. Facial recognition was also a general topic of interest and concern from our participants. A clear, plain language description of how it works and what is done with the data may help alleviate those concerns.

3.4.2 Implications for Design

Whether the IRS chooses to partner with a CSP or develop its own remote identity proofing system, it will be important to clearly display the time it takes to register, the documents required, and what the steps are in the process. In addition to communicating what documents are required, a description of how they are entered into the system (photo upload versus the last four digits of a credit card, for example) is useful to reduce assumptions users may have of what they will be required to provide. We also found that key terms used in buttons and headers lead to some misperceptions. For example, "Create Account" seemed to imply a quick sign-up process to our participants, and "Selfie" implied a stored image to share with family and friends.

4. Implications

Both our Secure Access and CSP studies suggest that taxpayers and Tax Professionals are willing to use the new digital identity concepts of remote identity proofing and two-factor authentication. In addition, individual taxpayers show willingness to identity proof and authenticate using a third party CSP.

4.1 Limitations and future work

Both the Secure Access and CSP studies were qualitative, and due to the level of effort behind recruiting and conducting individual interviews, had a small sample size. Our findings were also influenced by the questions we asked participants. Any observations or recommendations are intended solely to improve the design of IRS Secure Access and to not draw conclusions about any particular group. Any observations may be used to inspire larger sample surveys to provide more statistical power in order to understand the attitudes of key demographics on two-factor authentication, remote identity proofing methods, and third-party CSPs.

We identified three limitations of our research protocol. One key limitation in our data collection on the demographics of our participants was that income was self-reported, and we did not clarify whether it was before or after taxes. We also did not collect household size and therefore could not accurately group participants by income. When we asked, "Who do you think would have access to your personal information?" we did not

make it clear if "personal information" meant authentication details or the information held by the IRS. We also included a driver's license twice in the prototype; a Real-ID compliant license as primary evidence; and a non-Real-ID license as secondary evidence. It is more likely that a system would remove the non-Real-ID option after such a selection. Several participants selected the driver's license as both primary and secondary identity verification documents, so we countered this by asking them to choose a third option.

There is much more to learn about user perceptions and the usability of these new digital identity technologies. Additional qualitative research with key individual demographic groups as well as larger-sample surveys will help to validate that users are willing and able to use these new technologies. We feel that this is especially important for identity proofing methods like selfie verification and liveness testing. The IRS will play an important role in communicating key requirements to users either through communicating them on IRS.gov or ensuing a CSP does. Future design research on how communications style, content order, and visual layout improves comprehension and awareness of identity and authentication requirements will improve acceptance and satisfaction with IRS Secure Access.

4.2 Discussion

Both individual taxpayers and Tax Professionals were willing to use the prototypes they engaged with. The Tax Professional group prioritized security and expressed low trust in devices like smartphones. The majority of individuals selected a government-run CSP, and those who selected the fictitious ".gov" option presented were both less trusting of CSPs and more likely to choose the ".gov" due to their perception of its being more secure. This suggests that Tax Professionals may also prefer a government-run CSP, however both groups may also be willing to engage with a commercially run CSP if there are strong assurances of security, convenient identity evidence document selections, and a high-quality user experience.

Individual participants were not familiar with "jargon" and technical concepts such as CSPs or code generators. Any system design will have to strike a balance between offering clear and thorough information about the process, while grappling with the likelihood that users will only quickly scan or even skip over most content, as observed in both studies. Both studies showed misunderstandings and potential usability issues with selfie verification, and the Secure Access study showed potential accessibility issues. While we did not explore liveness testing with individuals, we anticipate similar misperceptions as seen with the Tax Professional audience. Selfie verification and liveness testing are still very new concepts and may be met with skepticism and uncertainty from users. Our participants were willing to try the process, but the finding is best viewed skeptically until true usability testing can be conducted.

5. Conclusion

As the IRS continues to grow its online services, including those that offer taxpayers access to their personal, sensitive information, they face a challenge of offering a digital identity solution that is both highly secure, usable, and accessible to its wide audience. We conducted a qualitative study to investigate individual taxpayer and Tax Professional perceptions of new digital identity technologies: remote identity proofing, two-factor authentication, and CSPs. We examined their trust, comprehension, and satisfaction around these new concepts, and looked to identify potential usability and accessibility issues. Our analysis revealed that both audiences are willing to use these services, despite their concerns. The majority of Tax Professionals interviewed were very concerned about security and had a false assumption that their image and biometrics were being captured in the process. Individuals had the same false assumption that their image was being stored in the selfie verification process. Individuals preferred not to have a choice of CSPs, but despite prioritizing registering directly with the IRS, they were willing to use a third party. Our work suggests a need for more research into improving comprehension and awareness through the design of future versions of IRS Secure Access, and further user research on new digital identity concepts such as selfie verification and liveness testing.

References

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. Psychological review, 87, 3, 215–251.

Internal Revenue Service (IRS). (n.d.) Welcome to Get Transcript. Retrieved May 20, 2020, from https://www.irs.gov/individuals/get-transcript.

---. (n.d.). Secure Access: How to Register for Certain Online Self-Help Tools. Retrieved May 20, 2020, from https://www.irs.gov/individuals/secure-access-how-to-register-for-certain-online-self-help-tools.

---. (2018, April). Internal Revenue Service Strategic Plan FY2018-2022 (Publication 3744). Retrieved May 20, 2020, from https://www.irs.gov/pub/irs-pdf/p3744.pdf.

Konkel, F. (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/.

Lazar, J., Feng J. H., and Hochheiser, H. (2017). Research methods in human-computer interaction. Morgan Kaufmann, 2017, 303–313.

National Institute of Standards and Technology (NIST). (2017, June). Digital Identity Guidelines (NIST Special Publication 800-63-3). Retrieved May 20, 2020, from https://pages.nist.gov/800-63-3/sp800-63-3.html.

Office of Management and Budget (OMB). (2019, May 21). Enabling Mission Delivery through Improved Identity, Credential, and Access Management (M-19-17). Retrieved May 20, 2020, from https://www.whitehouse.gov/wp-content/uploads/2019/05/M-19-17.pdf.

Smith, A. (2017, January 26). Americans and Cybersecurity. Retrieved from Pew Research Center for Internet & Technology May 20, 2020, from https://www.pewresearch.org/internet/2017/01/26/americans-and-cybersecurity/.

ten Brink, R., & Scollan, R. (2019, September). Usability of Biometric Authentication Methods for Citizens with Disabilities. Retrieved from MITRE Technical Papers May 20, 2020, from https://www.mitre.org/publications/technical-papers/usability-of-biometric-authentication-methods-citizens-disabilities.

U.S. Congress. House. Committee on Ways and Means; Budget; Financial Services. (2019 July, 1). H.R.3151 - Taxpayer First Act. Washington, DC: 116th Congress. Retrieved May 20, 2020, from https://www.congress.gov/bill/116th-congress/house-bill/3151.

4

Doing More With Less

Howard ◆ Lykke ◆ Pinski ◆ Plumley

Collum ◆ Jurney ◆ Marshall

Decal ◆ Lee ◆ Reinhart ◆ Shipley

Swanson ◆ Wooten

Can Machine Learning Improve Correspondence Audit Case Selection?
Considerations for Algorithm Selection, Validation, and Experimentation

Ben Howard, Lucia Lykke, and David Pinski (MITRE Corporation), and Alan Plumley (IRS, RAAS)

Abstract

The Internal Revenue Service (IRS) houses large volumes of data on taxpayer reporting and attributes. Using machine learning techniques, the IRS can use these data to improve its operational processes and make data-driven decisions. We consider the application of these techniques to correspondence audit case selection with the objective of improving revenue while limiting taxpayer burden.

The IRS conducts correspondence audits in a variety of categories, with each category focusing on potential misreporting on one or more lines of individual tax returns. Typically, returns are prioritized for correspondence audit based on one or more features of the return in question. We demonstrate how machine learning methods that incorporate many features of taxpayer reporting and other characteristics may improve audit results through focusing on two model outcomes. The two approaches we consider in developing machine learning models for audit case selection (i.e., prioritizing among the returns that meet all of the selection criteria) are:

1. Find the returns with highest expected assessed and/or collected tax revenue that will be generated if they were audited;[1] and
2. Avoid returns likely to have no-change audit outcomes if they were audited (that is, audits that result in little or no tax adjustment).

The first outcome is a continuous measure of value to the IRS. The second outcome is a binary measure of operational efficiency, as no-change audits produce no revenue, are unnecessarily costly for the IRS, and burdensome to compliant taxpayers. If we could perfectly predict before auditing a return whether it will produce a tax change, and if so, how much, then these two approaches would select the very same cases, and we would rank the returns in declining order of their tax adjustment. However, because we do not have perfect knowledge of audit outcomes before the fact, our pre-audit estimates of outcomes are subject to much uncertainty. This means that a modeling approach that seeks to find high-value cases is likely to yield somewhat different results from an approach that seeks to avoid low-value cases. The question is: which approach is likely to produce the largest aggregate revenue, given that neither approach will yield a perfect selection of returns for audit?

This study has two parts. In Part 1, we discuss the implementation of several iterations of machine learning model experiments for correspondence audit case selection for individual tax returns filed for Tax Years (TYs) 2013–2016. These operational experiments were conducted in collaboration with the correspondence audit program over the past several years and serve as a proof of concept for using machine learning techniques to rank tax returns for audit. Results show that for one type of correspondence audit, the machine learning

[1] In principle, we want to select returns that will be the *most cost-effective* to audit—not those with the largest predicted tax change. However, as a practical matter, there is not much variation in cost among correspondence audits, particularly within a given category. There is far more variation in revenue, so predicting that is our task in this paper.

selection algorithm resulted in an increase of 29 percent in assessed revenue in TY2014 compared to the status quo method, while TY2015 results showed little difference between the machine learning algorithm and the status quo selection method. For another category of correspondence audit, the machine learning algorithm results were mixed in terms of revenue, generally producing slightly less revenue compared to the status quo method, but machine learning selection methods resulted in a decrease in the no-change rate of approximately 7 percent in TY2014 and a decrease of 6 percent in TY2015 compared to the status quo selection method.

In Part 2, we focus on refining our models and research agenda based on lessons learned from the experiments in Part 1. For several categories of correspondence audits, we apply a variety of machine learning techniques, including regression algorithms, ensembles between classification (seeking to minimize the no-change rate) and regression (seeking to maximize the dollar outcome), and learning-to-rank algorithms. We validate our results during model development using diagnostic visualizations, and we show that strictly minimizing the audit no-change rate may come at the cost of collecting less revenue, and vice versa.

Introduction

Among the many responsibilities of Federal tax agencies are the obligation to enforce the tax code, encourage voluntary taxpayer compliance, and reduce the tax gap. For Tax Years 2011–2013, the annual gross tax gap in the U.S. was $441 billion. Underreporting tax liability—that is, filing one's tax return on time but underreporting how much tax one owes—comprises the largest component of the tax gap, accounting for $352 billion (Internal Revenue Service (2019)). Mitigating this phenomenon is a challenge that all tax administration agencies face (Webley *et al.* (2001)). A key aspect of closing the tax gap is reducing tax underreporting, such as by auditing taxpayers who are suspected of noncompliance to prompt them to pay the correct amount of tax. Ideally, the IRS would audit only those taxpayers who are indeed noncompliant, and whose returns yield the largest adjustments. The IRS and other tax agencies have vast amounts of tax return data spanning many years at their disposal. The focus of this paper is to make full use of these data for the purposes of detecting noncompliant, high-value returns, which requires integrating modern data analytics methods into audit operations.

In this study, we do two things (summarized in Table 1). First, in Part 1, we describe pilot studies conducted by our team in which the IRS correspondence audit program conducted operational audits to test how machine learning techniques to rank returns for correspondence audit performed against status quo ranking methods for ranking these audits. This experimentation, first implemented on Tax Year (TY) 2013 returns for two audit categories dealing with Schedule A and Schedule C expenses, respectively, was initially launched as a proof of concept. Could the IRS use machine learning techniques to rank and select returns for correspondence audits? What would the results look like? Through conducting this initial experimentation, we learned about several facets of model specification and training that warranted further exploration and investigation. This provided the motivation and the foundation for Part 2 of this study.

In Part 2, we take the insights from the initial pilot experiments to develop a research agenda focused on model specification and validation. Our pilot experiment results suggest that there may be a trade-off between seeking individual tax returns with the highest potential for generating revenue from audits—a measure of high value—and avoiding no-change audits, which result in no tax adjustment—a measure of low value and taxpayer burden. We present three modeling approaches trained and tested on historical audit data from one category of correspondence audit that examines some Schedule C (nonfarm sole proprietor business) line items, and we evaluate each approach in terms of predicted aggregate revenue and aggregate no-change rate.

TABLE 1. Summary of Study Parts 1 and 2

Study Part	Objective	Scope	Approach	Conclusion
1	To share results and lessons learned from 4 years of operational experimentation that tested the performance of machine learning methods for audit selection against status quo methods.	Two categories of correspondence audit: Category 1: Schedule C expenses; Category 2: Schedule A deductions. Operational experimentation conducted on tax returns from TYs 2013–2016.	Apply machine learning techniques to provide an alternative method to current (status quo) audit prioritization methods. Train and test models using out-of-time validation with completed audit records.	Operational experiments show mixed results with regard to revenue and no-change rate audit outcomes.
2	To take a deeper dive into machine learning modeling approaches for the correspondence audit use case, and present three different model types and predicted results.	One category of correspondence audit: Category 1: Schedule C expenses. Used training data from TYs 2012–2013, tested on data from TY2016.	Evaluate measures of audit value (revenue) and audit outcome (no-change rate) between three different applications of a gradient-boosted model (GBM). Train and test models using out-of-time validation with completed audit records.	All alternative machine learning approaches show a predicted improvement in no-change audit outcomes, compared to the status quo selection method. Some machine learning approaches predict improved revenue, compared to status quo selection.

Background

Applying data mining and machine learning techniques to large bodies of financial, accounting, and tax data is not a new concept for researchers. The application of these techniques to classification and prediction problems can help facilitate business decision-making and enhance operational efficiencies in the contexts of banking, stock exchanges, and taxation (Kirkos and Manolopoulos (2004)). However, using these techniques can be expensive and burdensome for organizations; they require maintaining large warehouses of good quality data and the software infrastructure required to make predictions or classifications with those data (Bots and Lohman (2003); Cleary (2011)). As such, it is crucial to identify how data mining techniques align with business decisions and to test whether data mining methods produce information patterns that are actionable (Bots and Lohman (2003)). In this paper, we show how a data mining approach can align to correspondence audit operational procedures and how different approaches may yield different results in terms of revenue and no-change rates.

In the tax domain, some prior studies focus on whether or not these methods—including approaches such as decision trees and neural networks—result in more favorable predicted outcomes, compared to traditional approaches (e.g., Gupta and Nagadevara (2007)), whereas others put data mining techniques into experimentation via pilot studies to see whether these methods indeed can outperform traditional approaches (e.g., Micci-Barrecca and Ramachandran (2004)). When analyzing historical audits without the use of operational pilot tests to determine whether data mining techniques improve audit selection in real settings, multiple studies have investigated how machine learning models might predict Value-Added Tax (VAT) noncompliance. In a study of VAT in Chile, researchers found that using a combination of neural network and decision tree models was most effective at detecting noncompliance among medium and large sized companies (Gonzalez and Velasquez (2013)). Similarly, a variety of supervised machine learning modeling techniques have shown promise predicting the occurrence of noncompliance for VAT in India (Gupta and Nagadevara (2007)). Notably, the authors of this study acknowledge a tradeoff between two model objectives; they were unable to attain optimal "strike rate" (percentage of cases that are true positives) and "performance efficiency" (percentage of true positive cases predicted by the model) simultaneously in any single model. We also compare multiple models in order to weigh the tradeoff between revenue and no-change cases.

Other countries and domains have put data mining techniques into use for audit selection. For example, researchers in Ireland developed a neural network model to score tax returns for their probability of "yielding" (that is, resulting in any revenue). This model showed promising results in testing, and therefore Revenue Irish Tax and Customs put the results into production by making predictions available to auditors (Cleary (2011)). In the U.S., two State pilot programs have shown promising results using data mining techniques for audits. Using predictive modeling to generate scores that represented risk of noncompliance on sales tax filings, Micci-Barrecca and Ramachandran (2004) found that these scores resulted in a 16-percent increase in audit adjustments compared to status quo prioritization methods used by auditors. Recently, the Minnesota Department of Revenue deployed a pilot program to use a supervised machine learning approach for sales and use tax audits; the machine learning technique showed improvement in predicting which were good cases (resulting in adjustment) and in predicting revenue from a case compared to status quo methods used previously (Hsu *et al.* (2015)). Note that in the Minnesota study, results from the machine learning pilot were compared against audit results from prior years (conducted on a different population of taxpayers); in this study, we advance that design by comparing results from machine learning tests against results from a control group of taxpayers audited using status quo methods randomly drawn from the same population.

Although machine learning algorithms are the predictive analytics techniques we focus on in this study, it is noteworthy that recent research has also used other data-driven techniques, including simulation techniques, to predict noncompliance and fraud. In the Australian tax context, Yang *et al.* (2011) advocate for an approach that simulates a distribution of "notional peers" to use unsupervised methods to detect noncompliance. Agent-based modeling that simulates taxpayer behavior and attitudes, including occupation choice and following social norms, has also shown promise in predicting increased audit revenue compared to revenue from randomly selected audits (Hashimzade *et al.* (2016)).

Data-Driven Selection Techniques at the IRS

In this study, we apply data-driven machine learning selection techniques to correspondence audits, which have not, to our knowledge, previously been the subject of this type of selection method. These are audits conducted primarily through the mail (though taxpayers may communicate with the IRS about their audit via phone or online). Correspondence audits focus on narrowly defined segments of the individual taxpayer population, and only examine one to three line items on the return where noncompliance is suspected; for example, overstating certain expenses or deductions to reap a tax benefit. There are many categories of correspondence, each with different taxpayer populations of interest; in this study, we focus on pilot program results from two categories, and explore the performance of several modeling approaches with one single category focused on Schedule C expenses.

The IRS has long used one particular type of data-driven classification technique, discriminant function (DIF), to select tax returns for field examinations (comprehensive audits of a wide range of tax reporting) (Wedick (1983)). DIF scores use many variables to predict potential noncompliance and identify returns to consider for audit (Rettig (2016)). First developed using data from the Taxpayer Compliance Measurement Program (TCMP), research has shown these scores to closely mimic human classifiers' selections and to predict unreported income (Cyr *et al.* (2002)), and to perform better compared to neural networks for the purpose of selecting certain types of audits (Asner (1993)). DIF scores are now calculated using National Research Program (NRP) data, a nationally representative sample of taxpayers where the entire individual tax return is subject to audit (Brown and Mazur (2003); Luttati (2006)).

However, the IRS currently does not apply the DIF approach to correspondence audits because these audits are issue-focused rather than looking at the entire return; instead, correspondence audits use what Rettig (2016) calls "user-developed criteria," or business rules developed to identify narrow taxpayer populations with suspected noncompliance on one or a small number of specific line items on the tax return. DIF scores as they exist currently would not work for correspondence audits, because they are a measure of the potential misreporting on the *full* individual tax return. Additionally, DIF is derived using NRP data, which represent the full taxpayer population, whereas correspondence audits examine small, specific subsets of taxpayers who meet certain criteria. As such, using a data mining method for correspondence audit selection needs to be tailored to the attributes of the correspondence audit program. We do this by training and validating our predictive

models on data from previous correspondence audits from the same category, rather than from the full taxpayer population. We also advance the state of the field by using an experimental design that includes a control group that is randomly drawn from the same taxpayer population in the same year to compare the outcome of our alternative machine learning approach to current status quo selection methods.

Research Objectives

As such, this study addresses two research objectives, which are addressed in turn in the two major parts of this paper:

1. To evaluate from experimentation the value (in terms of no-change rates and two definitions of revenue) of using machine learning techniques to select returns for correspondence audit; and
2. To refine the machine learning methodology used in future experimentation by assessing the potential results of three different machine learning modeling approaches, in terms of the same outcome metrics.

Part 1. Machine Learning for Correspondence Audits: Proof of Concept Experiments

In this study, we first describe operational experiments to test the effectiveness of machine learning methods on correspondence audits conducted for TYs 2013 through 2016. For each tax year, two categories of audit were considered, which we denote by Audit Category 1 and Audit Category 2. Experimentation continues for Audit Category 1 post TY 2016. For Audit Category 2, the IRS transitioned to using the supervised learning method for all correspondence audits for TY 2017.[2] In this section, we present more details on these experiments and the results. We use the findings from these pilot experiments as motivation for our deeper dive into refining and improving our methodology in Part 2 of this study.

Experiment Design

Each year, the IRS down-selects a subset of the general U.S. taxpayer population to consider for correspondence audit using business rule filters. This subset represents the candidate population for audit; that is, the potential pool. In TYs 2013–2016, we used an experimental design where 50 percent of returns were assigned to be selected from this potential pool using traditional user-defined criteria methods ("status quo" method), and 50 percent of returns were selected to be audited using the machine learning algorithm ("alternative" method). That is, half of the audits are ranked using the status quo method and audited in that order, and the other half are ranked using the alternative method and audited in that order. It is important to note that not all of the returns in the candidate population are audited. Therefore, both the status quo and alternative method will leave behind a remainder of unaudited returns. Although there is typically significant overlap in the populations that were selected by either method, this allows for the alternative method to select returns that otherwise would not be selected by the status quo method.

Experiment Approach

Below, we describe the two audit categories that were used in the pilot data mining studies for returns audited from TYs 2013–2016, and the machine learning models used to generate the alternative rankings for the pilot study experiments.

Audit Category 1. This correspondence audit category examines some Schedule C business expenses. For this pilot experiment to test the efficacy of machine learning techniques for audit selection, we applied a Two Stage Support Vector Machine (SVM), trained on collected revenue from audits conducted in prior years.[3] The two stages refer to a binary classification SVM model trained to identify no-changes, and a regression SVM trained

[2] We do not describe results from TY2017 for Audit Category 2 in this paper because there was no "control" group for this year—that is, the IRS used the alternative selection method with machine learning algorithms to rank all correspondence audits for Audit Category 2. As such, there is no point of comparison for the method's performance.

[3] See Analytical Approaches section for more details on how we train and validate models using out-of-time validation methods.

to predict revenue. Binary SVMs attempt to learn a hyperplane that maximizes the boundary between the two classes (Boser *et al.* (1992)); regression SVMs operate similarly, instead minimizing the residuals between training points and the points predicted by the hyperplane.

Audit Category 2. This correspondence audit category examines some Schedule A deductions. For this pilot experiment testing the use of machine learning techniques for selection, we used Gradient Boosted Machine (GBM) trained on collected revenue, with no consideration for no-change rate. Boosting algorithms are based on the idea of combining weak models additively, where subsequent models learn from errors of previous models; GBMs extend this idea to decision trees, with the decision trees being combined additively using gradient descent (Friedman (2001)).

Experiment Results

For the experiments, we report the following outcome metrics to evaluate the success of the machine learning alternative ranking approach compared to the status quo ranking method. Tables 2 and 3 display the results for the pilot experiments for Audit Categories 1 and 2. Figure 1 shows these same metrics in bar chart format.

Total Assessed Revenue. This is the cumulative assessed revenue from the entire set of returns audited for a given tax year and a given prioritization method. Note that not all revenue assessed from an audit will ultimately be collected from the taxpayer.

Total Collected Revenue. This is the cumulative collected revenue from the entire set of returns audited for a given tax year and a given prioritization method by a certain point in time. This represents real dollars returned to the IRS. Note that there is often a time lag in revenue collection—some taxpayers may slowly remit payments over time, so collected revenue may continue to trickle in for years after an audit.

No-Change Rate (Assessed). This measure is defined as the percentage of audits that resulted in an assessed tax liability adjustment of $100 or less.

Average Revenue on Changes. This measures the average revenue assessed for audits that resulted in an adjustment of $100 or more.

In Table 2, we show results for experiments conducted with TYs 2013–2016 audited returns for Audit Category 1. The alternative methods developed using machine learning methods are highlighted in orange while the status quo methods are highlighted in green. We observe a consistently lower no-change rate for the alternative method compared to the status quo for Audit Category 1: the no-change rate, as measured by any tax adjustment less than $100 resulting from the audit, ranges from 5 to 8 percentage points lower for the alternative method than the status quo selection method. On the other hand, for Audit Category 1, the status quo method returned more assessed revenue compared to the alternative method in the range of about $2,000,000 to $6,000,000. However, the differences in collected revenue between the two ranking methods is much smaller, and in fact the TY2016 alternative ranking method yielded $1,254,757 *more* collected revenue compared to the status quo ranking method despite yielding less assessed revenue. This suggests that the alternative method may have better prioritized taxpayers who represent different potential for revenue collectability—their ability or willingness to actually pay additional tax liability.

Table 3 displays results for the audit experiments conducted in TYs 2013–2015 for Audit Category 2. With regard to revenue, the results are inconsistent across years:[4] the alternative ranking method did not perform better compared with the status quo method for TY2013 (the status quo method resulted in $1,232,778 more in assessed revenue and $889,726 more in collected revenue for TY2013), but the alternative ranking resulted in more revenue in both TYs 2014 and 2015. The difference is especially notable for TY2014: the alternative method returned $16,774,305 more in assessed revenue and $13,259,820 more in collected revenue compared to the status quo method. As displayed in Figure 1, the no-change rates are either equivalent across ranking methods, or 2 percentage points higher for the alternative method in TYs 2013 and 2014.

[4] Although outside the scope of this paper, additional investigation to better understand why results are inconsistent from one year to the next could be fruitful. To our knowledge, the status quo selection methods remained consistent throughout our study years. However, the specifications of the alternative models changed slightly year over year; additionally, the taxpayer population underlying both the models and the experimental treatment and control populations changes from one year to the next.

Overall, the results from these preliminary experiments show that machine learning methods can be fruitful for correspondence audit selection, but the magnitude of the improvement is variable and dependent on the type of audit category. Further, in Figure 1, we observe a trade-off between improving revenue versus no-change rate: for Audit Category 1, the alternative method was more effective for decreasing the no-change rate, whereas for Audit Category 2, the alternative method brought in more revenue while slightly increasing the no-change rate. Observing these results in the preliminary pilot experiments provides the motivation for the focus on finding ways to achieve predictive performance on both outcomes—revenue and no-change rates—simultaneously.

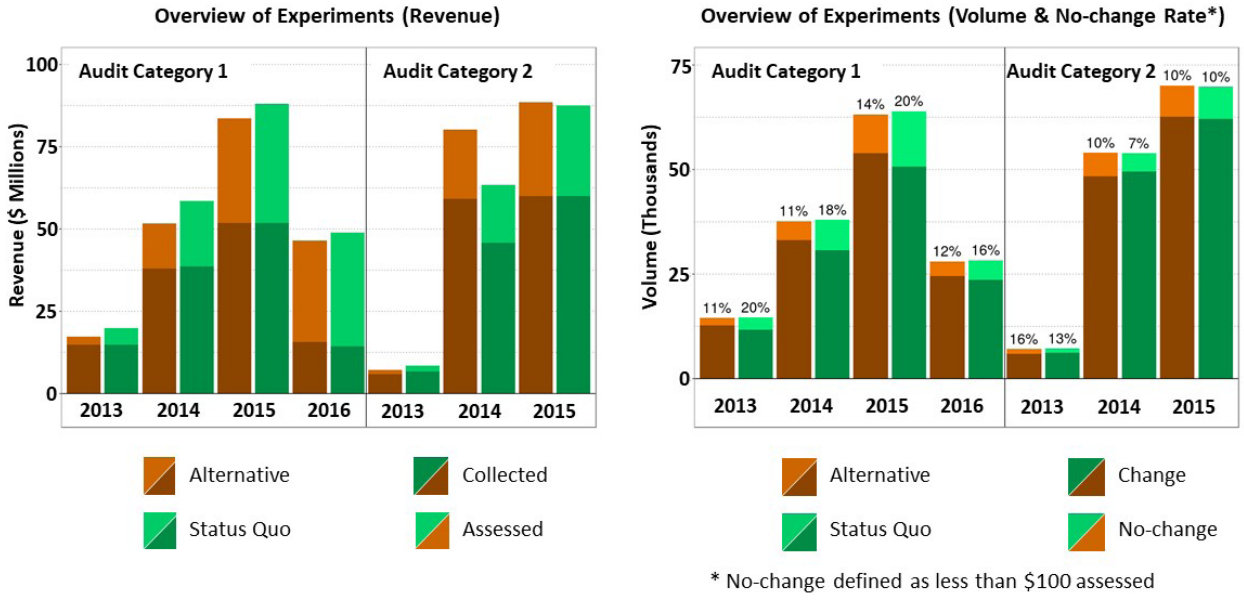
TABLE 2. Revenue and No-Change Rates for Experiment Results, Audit Category 1 (Schedule C Expenses), TYs 2013–2016

Tax Year	Method	Count	Total Assessed Revenue	Total Collected Revenue	No-Change Rate (Assessed)	Average Revenue on Changes
2013	Alternative	5,791	$ 17,360,429	$ 14,775,204	12%	$ 3,405.34
2013	Status Quo	5,865	$ 19,959,013	$ 14,898,792	20%	$ 4,272.96
2014	Alternative	15,062	$ 51,660,978	$ 37,920,470	12%	$ 3,894.24
2014	Status Quo	15,184	$ 58,515,068	$ 38,688,618	19%	$ 4,756.94
2015	Alternative	25,280	$ 83,629,971	$ 51,868,231	15%	$ 3,874.09
2015	Status Quo	25,578	$ 87,968,337	$ 51,897,451	21%	$ 4,341.76
2016	Alternative	11,201	$ 46,459,194	$ 15,659,677	12%	$ 4,728.67
2016	Status Quo	11,305	$ 48,891,693	$ 14,404,920	17%	$ 5,182.50

TABLE 3. Revenue and No-Change Rates for Experiment Results, Audit Category 2 (Schedule A Deductions), TYs 2013–2015

Tax Year	Method	Count	Total Assessed Revenue	Total Collected Revenue	Assessed No-Change Rate	Average Revenue on Changes
2013	Alternative	2,833	$ 7,190,287	$ 5,878,440	16%	$ 3,037.72
2013	Status Quo	2,882	$ 8,423,065	$ 6,768,166	14%	$ 3,397.77
2014	Alternative	21,591	$ 80,165,646	$ 59,102,912	10%	$ 4,146.36
2014	Status Quo	21,536	$ 63,391,341	$ 45,843,092	8%	$ 3,199.00
2015	Alternative	28,042	$ 88,581,377	$ 60,024,966	11%	$ 3,539.01
2015	Status Quo	27,919	$ 87,511,260	$ 59,999,856	11%	$ 3,521.72

FIGURE 1. Visualization of Audit Revenue and Change Rates by Selection Method and Tax Year, Audit Categories 1 and 2.



Part 2. Testing Three Analytical Approaches to Machine Learning for Correspondence Audit Selection

In Part 2, our objective is to refine our modeling approach for correspondence audit selection based on the lessons learned and insights derived from the pilot experiments described in Part 1. We compare three modeling approaches that can be used for case selection of correspondence audits for Audit Category 1 moving forward. Audit Category 1 continues to be a top priority type of correspondence audit for the IRS, so tuning the selection approach to maximize revenue is valuable to IRS operations.

For each of the three modeling approaches described below, we consider the potential benefits and pitfalls arising from how the approach handles the binary outcome (change versus no-change audit result) versus the continuous outcome (revenue). We validate the models using other historical audit data to project how well they would perform in a real experimental setting compared to the status quo selection method. For each model, we evaluate assessed revenue, collected revenue, and no-change rate. (See the description of these outcomes in Part 1.)

Analytical Approaches

Table 4 shows a summary of the three modeling approaches used for the analysis in Part 2. We also include whether or not each modeling approach includes a binary classifier in addition to a continuous regressor—that is, whether the approach explicitly accounts for no-change audit outcomes.

TABLE 4. Modeling Approaches Tested

Approach	Description	Include binary classifier?	Include continuous regressor?
Pairwise	Compares returns by ranking pairs against each other in terms of revenue.	No	Yes
Hurdle (2-stage regression)	Combines two trained models: binary classifier (no-change) with regression model (revenue).	Yes	Yes
Penalized regression	Takes probability of no-change into account when training a second regression model.	Yes	Yes

Approach 1: Pairwise Ranking

Pairwise ranking attempts to learn a ranking function f over all returns $\mathbf{X} = \{\overrightarrow{x_1}, \cdots, \overrightarrow{x_n}\}$. Each return $\overrightarrow{x_i}$ has an associated label (e.g., audit revenue) l_i; if $l_i > l_j$, then $\overrightarrow{x_i}$ should be ranked before $\overrightarrow{x_j}$.

To learn a ranking function the algorithm compares the label for every pair of returns with different labels The pairwise loss function $\mathcal{L}$ is defined in Chen *et al.* (2009) as:

$$\mathcal{L}(f; \boldsymbol{X}) = \sum_{s=1}^{n-1} \sum_{i=1; l_i < l_s}^{n} \phi\left(f(x_s) - f(x_i)\right),$$

where

- f is the pairwise ranking function
- $\boldsymbol{X}$ is a collection of returns and x is an n dimensional vector representing a taxpayer's return
- ϕ is the logistic function: $\phi(z) = log(1 + e^{-z})$.

Since the loss function compares returns where $l_s > l_i$, substantial loss occurs only when a pair of objects is ordered incorrectly – that is, $f(x_s) - f(x_i) < 0$. When the pair of returns is ordered correctly, there

is still loss; however, as the difference in predicted rankings for correctly ordered pairs increases, that loss for the pair approaches zero.

$$\lim_{[f(x_s)-f(x_i)]\to\infty} \mathcal{L}(f;\boldsymbol{x}) = 0$$

Furthermore, this loss has an upper bound of log(2). In practice, this loss will be inconsequential compared to the loss incurred by incorrectly ordered pairs.

Pairwise ranking weighs all errors equally, regardless of the difference in the underlying revenue values. For example, the model will be indifferent between ranking a return that yields $5,000 before a return that yields $1,000 and ranking a return that yields $1,000 before a return that yields $0. A model that is more focused on maximizing revenue would focus on minimizing the first kind of error, while a model that is more focused on reducing no-changes would focus on minimizing the second kind of error.

Approach 2: Hurdle Model

Drawing from Cragg (1971), the hurdle model is defined below:

$$h(x) = I(R(\boldsymbol{x}) > 0) \cdot f_c(\boldsymbol{x}) + I(R(\boldsymbol{x}) = 0) \cdot f_{nc}(\boldsymbol{x}),$$

where

- x_i is an n dimensional vector representing a taxpayer's return
- R(x) with range [0,) is the function for the actual revenue received.
- $I(\cdot)$ is the indicator function which, when the condition is met, is 1 and otherwise is 0.
- $f_c(x)$ is the function approximating value assuming the return is a change
- $f_{nc}(x)$ is the function approximating value assuming the return is a no-change.

In our use case, since no-change audits have no value, $f_{nc}(\boldsymbol{x}) = 0$, so the hurdle model simplifies to:

$$h(\boldsymbol{x}) = I(R(\boldsymbol{x}) > 0) \cdot f_c(\boldsymbol{x}).$$

From here, the hurdle model can be implemented in two ways, which in this paper will be referred to as the "two-stage" model and the "expected value" model. For each implementation, a binary classification model is trained to predict whether a return will result in no-change. The function $f_c(\boldsymbol{x})$ is usually a supervised model to predict revenue trained on actual change cases, but it can be potentially other forms such as a learning-to-rank model, an unsupervised algorithm, or even another ensemble.

The two-stage model chooses a cutoff point in the binary classifier's output to label a result as a change or no-change that can be chosen by heuristic methods, such as wanting to declassify the top k returns in the candidate pool, or data-driven methods such as from a precision-recall plot.

$$H(x) = I(g_c(x) > t) \cdot f_c(x),$$

where

- $g_c(x)$ is a binary classifier model predicting whether a return x will result in a change, usually the range is [0,1] expressing a probability.
- t is the cutoff value in the range of g_c(x).

The expected value of the hurdle model is defined as

$$E[H(\boldsymbol{x})] = g_c(\boldsymbol{x}) \cdot f_c(\boldsymbol{x}).$$

As an example, consider a return x_1 with $f_c(x_1) = 1000$, $g_c(x) = 0.6$, and two cutoffs $t_1 = 0.7$, $t_2 = 0.5$. The result from the expected value model would be $0.6 \cdot 1000 = 600$. The two-stage model with cutoff t_1 would be $I(0.6 > 0.7) \cdot 1000 = 0 \cdot 1000 = 0$. The two-stage model with cutoff t_1 would be $I(0.6 > 0.5) \cdot 1000 = 1 \cdot 1000 = 1000$.

Approach 3: Penalized Regression

The penalized regression method modifies the mean squared error loss function commonly used in regression in order to incorporate the predicted probability of no-change from a binary classifier. To train a penalized regression, a binary classifier $g(\boldsymbol{x})$ must first be trained. The no-change probability predictions for the classifier on the training data are saved; the penalized regression is then trained on the same data, now incorporating the no-change probability predictions from the binary classifier.

The loss function $\mathcal{L}$ is defined as:

$$\mathcal{L}(f;\ \lambda;\ \boldsymbol{x}) \ = \ \frac{\sum_{i=1}^{N} \quad (f(x_i) - y_i)^2\lambda + (1-\lambda)\ \mathrm{g_{nc}}(x_i)\cdot f(x_i)}{N} \ .$$

where

- x_i is an n dimensional vector representing a taxpayer's return
- $f(x)$ is the function approximating revenue
- λ is a parameter in the range of [0,1] that must be tuned to balance the two objectives
- $\mathrm{g_{nc}}(\mathrm{x})$ is a binary classifier model predicting whether a return x will result in a no-change, usually the range is [0,1] expressing a probability.

The first term in the loss function is mean squared error; the second term accounts for the probability of no-change. The loss function seeks to balance cases where the no-change probability prediction is at odds with the regression estimate.

Penalized regression can be seen as inserting bias into the training procedure, similar to what regularization does to avoid overfitting to training data. In this case, the bias serves to account for a particular type of overfitting driven by complex returns, which can be high-value when resulting in a change, but which also have a relatively high probability of resulting in a no-change. For other examples of regularization, we recommend looking into elastic net regression (Zou and Hastie (2005)), dropout in neural networks (Srivastava *et al.* (2014)), and early stopping in gradient-based methods (Prechelt (1998)).

Data and Methods

We use out-of-time validation for Audit Category 1 to illustrate results from our approaches. Out-of-time validation is out-of-sample validation on a later dataset than the dataset used to train the model. For this illustration, the training set consists of audit data from TYs 2012 and 2013 and validated on TY 2016 audit data.

We use the 'xgboost' package (Chen and Guestrin (2016)) to train gradient boosted machines (GBMs) for all models. Boosting algorithms are based on the idea of combining weak models additively, where subsequent models learn from errors of previous models; GBMs extend this idea to decision trees, with the decision trees being combined additively using gradient descent (Friedman (2001)). For the optimization of hyperparameters while training the gradient boosted trees, we have found a lower learning rate (about 0.1), a maximum tree depth of 8, and a relatively large number of trees (at least 200) to work well.

The models use 515 features related to the subject of the audit. Examples of these features include the change in total tax owed over the 4 years of tax returns prior to audit, the profit or loss amount from a Schedule C, and whether the taxpayer files as "married filing jointly" or not. To assess the generalizability of the model, we employ five-fold cross-validation; this is an out-of-sample testing technique where the model is trained on a random 80-percent subset of all observations in the dataset and evaluated on the remaining 20 percent.

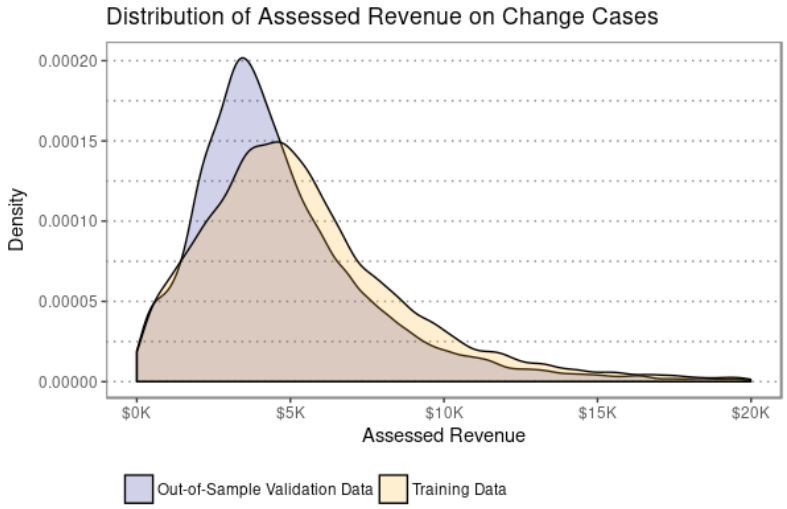
For the two-stage model, we train a regression GBM that minimized the root mean squared error (RMSE) on the log value of assessed tax on only the changes within the training data. We have found that training revenue-specialized models on only changes makes a substantial improvement compared to models trained on the both changes and no-changes. We use the top 40$^{\text{th}}$ percentile of predicted no-change probabilities as our cutoff for classifying a return for no-change.

Table 5 and Figure 2 display sample sizes, no-changes rates, and the distribution of revenue for the training set and the out-of-time validation set in order to gauge how parallel the data sets are on key attributes. We note the stark difference in no-change rate between the two sets. This difference is due to an operational change in the selection of cases between 2013 and 2016. Nevertheless, we do not observe that this difference adversely affects prediction; we present the prediction results in the next section. Figure 2 shows the distribution of assessed revenue for only the audits that resulted in a tax adjustment between the training data and out-of-time validation data. The two data sets have similar distributions on assessed revenue, though the out-of-sample set has a thinner tail, with more returns yielding values from $1,000–$5,000.

TABLE 5. Summary of Training Dataset and Out-of-Time Validation Dataset

Dataset	Tax Year(s)	Sample Size	No-change rate
Training	2012, 2013	48,141	34.4%
Out-of-time validation	2016	29,995	16.4%

FIGURE 2. Distribution of Assessed Revenue on Change Cases in the Training Data and Out-of-Time Validation Data



Evaluating Results

In this section, we present a comparison of the approaches detailed in the Analytic Approaches section. These results were produced on completed audit data from TY 2016 with models trained on TYs 2012 and 2013 to project how each method will fare in an actual scenario. Since each method is applied to the same dataset, constraints on the selection process for this dataset can impact the performance of models on the out-of-time validation set. Thus, an operational experiment with no constraints on how returns are selected is needed to verify model performance.

To facilitate presentation of results, we introduce the following terms:

The Perfect Knowledge Ranking: The "Perfect Knowledge" ranking is the realized results of audits. If known *a-priori*, this would result in the best possible ranking these data can give us. While practically not feasible, this ranking is useful as it provides a ceiling against which to compare other methods.

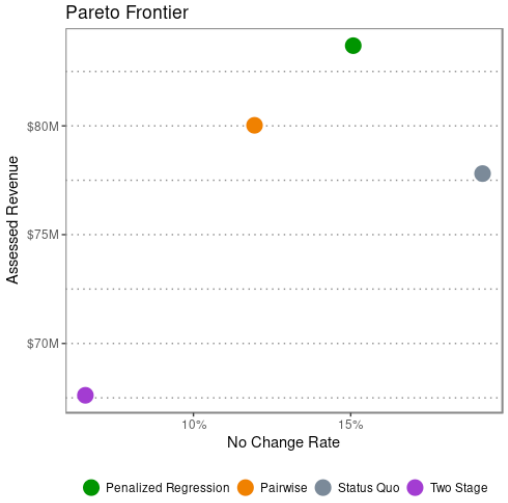
Pareto Frontier: A validation plot that visualizes the tradeoff between no-change rate and cumulative revenue at a given point of cumulative returns selected for multiple methods of ranking.

Lift Plot: A validation plot showing for one or more methods of ranking how cumulative revenue increases with the number of returns audited.

No-Change Progression Plot: A validation plot that compares for one or more methods of ranking the progression of cumulative no-change rate as a function of the number of returns audited.

Below is a Pareto frontier comparing cumulative no-change rate and cumulative assessed revenue of the first 50 percent of the returns in the validation dataset selected by each method. The top left of the plot is the ideal region for a method to perform, but because our overall objective is to maximize revenue, models that predict more revenue take precedence over minimizing the no-change rate. The Pareto frontier illustrates which methods can increase revenue *and* decrease the no-change rate. It shows that the two-stage model performs the best for reducing no-change rate but it achieves *less* revenue than status quo, while penalized regression conversely performs the best at maximizing revenue but performs the worst at reducing the no-change rate relative to the status quo. Pairwise performs well on both objectives but does not outperform two-stage in terms of no-change rate or penalized regression at maximizing revenue. Finally, pairwise and penalized regression perform better than status quo on *both* objectives while two-stage performs better only at minimizing no-change rate. These results suggest that penalized regression is the best of these alternative models, given its superiority at maximizing revenue while also decreasing the no-change rate compared to the status quo method.

FIGURE 3. The Pareto Frontier Validation Plot



Next, we present the lift plot and the no-change progression plot (Figures 4 and 5). These plots take the audit data for TY 2016, reorder the returns uniquely for each method of ranking, and calculate a metric as a function of the cumulative number of returns audited. Because each method is evaluated on the same data (and the same number of returns), they have the same beginning and ending calculated metrics (on the left and right ends of the x-axis, respectively).

As each method ranks returns differently, there will generally be a difference in the calculated metric for intermediate values, which is used to determine if one method is more promising than another. The advantage these plots have over the Pareto frontier is the ability to visualize how rankings progress through the pool of audits; this can inform decision-making.

FIGURE 4. The Lift Progression Validation Plot

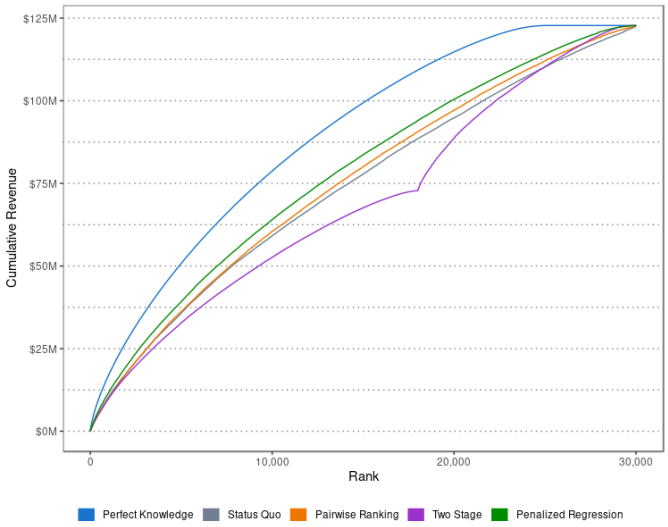
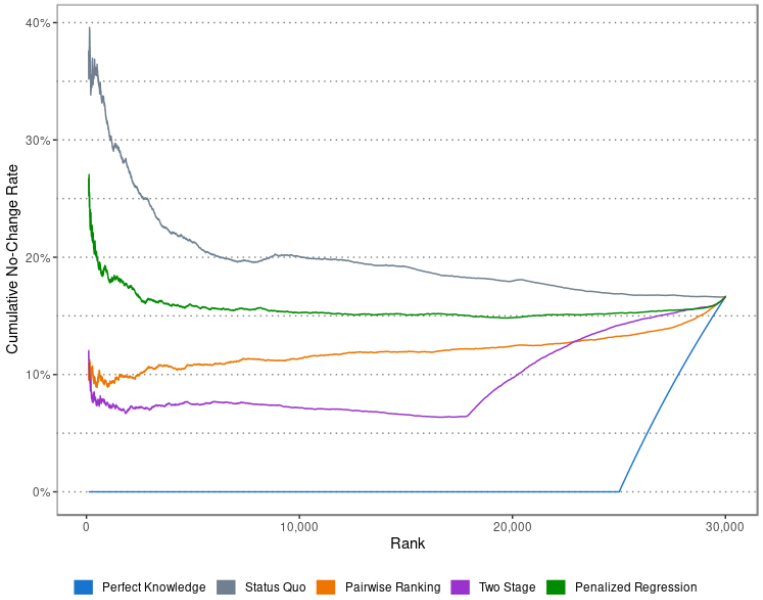


FIGURE 5. The No-Change Progression Validation Plot



For the lift plot, the calculated metric is cumulative revenue; methods that prioritize returns better will show *upward* vertical separation over inferior methods. While the takeaway from the Pareto frontier is consistent for the lift plot in the order in which the models perform on maximizing revenue, there are several aspects to note. Penalized Regression shows immediate improvement over status quo and at no point does it not show improvement. Pairwise seems to behave very similarly with status quo at early values of priority but starts to show visible improvement after 10,000 returns (~33 percent) are selected.

At no point does the two-stage ranking show better performance over status quo. An interesting note is the sharp edge that is observable in two stage's progression just before 20,000 returns (~66 percent) are selected. This is the point at which two-stage runs out of predicted changes and starts picking the returns that it classifies as a no-change. Another interesting note is that no-changes are visible on the perfect knowledge

ranking's right end, where it will exhaust all revenue-bearing audits and start selecting no-changes in the historical sample being used.

The no-change progression plot's calculated metric is the cumulative no-change rate at each point of selection. Unlike the lift plot, *downward* vertical separation between rankings shows improvement (toward a lower no-change rate). Perfect knowledge is a good illustration for this plot as it has a perfect change rate up until the point it is forced to pick no-changes, where it sharply approaches the overall no-change rate of the dataset.

Again, the takeaway from this plot is the same as the Pareto frontier. This plot shows that the two-stage model performs the best at minimizing no-change rate, followed by pairwise, then penalized regression. The point where the two-stage model starts picking classified no-changes is again observable as the progression quickly approaches the no-change rate of the entire sample. The assumption is that if the two-stage approach were allowed to pick data outside of this dataset, it would continue to perform at a no-change rate similar to before this inflection point. Likewise, all of the alternate methods would undoubtedly have selected some better returns from the pool of candidate returns beyond those selected by the status quo method (which are the only ones for which we have audit results). If any of these alternatives (including the perfect knowledge ideal) were able to select returns instead of those that actually *were* audited, their performance on both metrics would likely continue to be similar to their performance near the middle of the rankings shown in the figures.

Each plot emphasizes a different aspect, but no single plot tells the entire story. The Pareto frontier gives a snapshot comparing the models on both dimensions but does not illustrate the progression of either. The lift plot illustrates how each method progresses in terms of revenue, but not change, while the no-change progression plot is the opposite. The most important thing is that all plots are consistent together. All methods show improvement at minimizing no-change comparing to status quo, with two-stage projecting around a 10-percent decrease. The penalized regression model shows the best improvement over status quo at maximizing revenue. Both the penalized regression and pairwise methods improve on the status quo on *both* metrics, but pairwise does not achieve the revenue gains that seem possible with the penalized regression method.

Discussion

Correspondence audits at the IRS are limited in scope and traditionally rely on "rules of thumb" user-driven criteria (Rettig (2016)) to rank returns from a candidate population for audit prioritization. In this study, we investigated whether using machine learning techniques could improve audit outcomes for correspondence audits compared to the status quo ranking methods. Our results showed that when tested in operational experimentation, machine learning methods can yield substantially higher revenue for one type of correspondence audit and may decrease the occurrence of no-change audits for another type of correspondence audit. Further, we showed how refining models in this domain can benefit from prioritizing cases that have high audit value (revenue) and/or seeking to avoid cases likely to result in no change at all. In the end, the ranking that maximizes aggregate revenue at a given budget level is preferred.

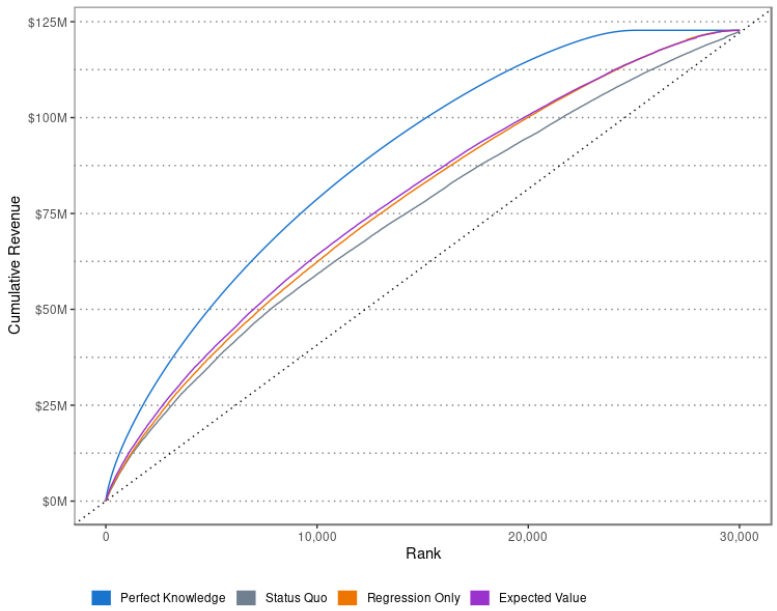
The appeal of machine learning for this use case is the method's ability to detect patterns in data that are predictive of a desired outcome but may not be intuitive according to theory or experience. In Part 1 of our study, we reported on pilot experiments conducted over several tax years to compare a machine learning approach for audit ranking to status quo methods for two audit categories. These proof-of-concept results revealed that it seems possible in principle to improve the operational ranking of candidate returns to increase revenue or to reduce the no-change rate without adversely affecting aggregate revenue. This was the motivation for the modeling approaches presented in Part 2 of this study, where we showed that various model specifications produce different outcomes.

In an ideal world with the ability to perfectly predict outcomes, there would be no need to consider audit no-change rates alongside revenue; indeed, in this ideal world, the ranking method that resulted in maximum revenue would also result in minimal no-change cases and minimal wasted cost, given that no-changes return no revenue. However, prediction is imperfect, and in this study, we discussed the need to ensure that ranking models do not unnecessarily increase no-change rates while pursuing the ultimate objective of maximizing revenue. Other researchers have observed in the tax domain that models can be tuned to account for competing or complementary objectives. For example, strike rate and efficiency may be in conflict, depending on how

models are constructed (Gupta and Nagadevara (2007)), and audit selection models may need to simultaneously account for revenue collection and cost savings in order to meet tax agency needs (Hsu *et al.* (2015)). Again, identifying the business needs of decision-makers and stakeholders is crucial to understanding how to apply data mining techniques to real life contexts (Kirkos and Manalopoulos (2004)).

Harmonizing potentially conflicting outcomes to maximize an ultimate objective requires iterative research and experimentation. To this point, as we have refined our methods for Audit Category 1 for correspondence audit, our most recent development of the expected value variant of the hurdle model shows similar performance to the penalized regression approach detailed in Part 2. For illustrative purposes, we introduce here this expected value variant of the hurdle model in which we use the same regression and binary classification models as in the hurdle model described above. In this refinement of the hurdle model presented in Part 2, it is encouraging that by accounting for no-change outcomes, the expected value model yields more overall revenue than its regression component does by itself. Figure 6 displays these results and shows that the expected value model shown in purple has slight visible separation from the regression-only model shown in orange. This further supports the assertion that better prediction results in both more revenue *and* lower no-change rates simultaneously.

FIGURE 6. Lift Plot Showing the Effect of Accounting for No-Changes in the Expected Value Model



Finally, when decision-makers in the tax domain or other contexts consider using data mining techniques in practical application, it is crucial to recognize what the technique can and cannot account for. A decision-maker will have many more considerations than we can address in this paper for this use case. For example, how does the cost of audit operations compare to the yielded revenue (U.S. Government Accountability Office (2012))? How much assessed revenue is collected? On this point, we saw an example from our experimentation with Audit Category 1, where the alternative machine learning ranking method resulted in fewer dollars assessed than the status quo method, but *more* dollars collected for TY 2016; this suggests that taxpayer behavior in remitting payments may be a confounding factor to consider in model training and validation. Further, what kind of returns can a tax agency's staff reasonably audit? Is machine learning the best tool to solve a given problem, and does an organization have the bandwidth to stand up infrastructure to test this approach—a nontrivial effort (Bots and Lohman (2003))? This paper's intent is not to provide answers to these questions, but to provide the tools and methodology to aid the people responsible for doing so.

Future research in this domain should expand the use case beyond two types of correspondence audit to consider other types of correspondence audit as well as other types of enforcement activities done by the IRS. Additional categories of correspondence audit represent different types of taxpayers as well as more variety in the types of line items considered in an audit. Further, other types of enforcement activities, such as field audits, may represent more complex tax returns that could present greater challenges for predicting audit outcomes. Additional research, wherever possible, should include operational experimentation as we have described here, in order to facilitate iterative learning and model improvement.

References

Asner, Lance S. (1993). "Neural Networks and Discriminant Function: Alternative Techniques of Selecting Tax Returns for Audit." *1993 IRS Research Bulletin*: 118–127.

Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. (1992). "A Training Algorithm for Optimal Margin Classifiers." In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152.

Bots, Pieter W.G., and Fred A.B. Lohman. (2003). "Estimating the Added Value of Data Mining: A Study for the Dutch Internal Revenue Service." *International Journal of Technology, Policy, and Management* 3(3-4): 380–395.

Brown, Robert E., and Mark J. Mazur. (2003). "IRS's Comprehensive Approach to Compliance Management." *National Tax Journal* (56)3: 689-700. Retrieved from https://www.ntanet.org/NTJ/56/3/ntj-v56n03p689-700-irs-comprehensive-approach-compliance.pdf?v=%CE%B1&r=48361853085826567.

Burges, Chris, Tai Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, Greg Hullender. (2005). "Learning to Rank Using Gradient Descent." ICML '05: *Proceedings of the 22nd International Conference on Machine Learning*, 89–96.

Chen, Tianqi, and Carlos Guestrin. (2016). "Xgboost: A Scalable Tree Boosting System." KDD '16: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, Wei, Liu Tie-Yan, Lan Yanyan, Ma Zhi-Ming, and Li Hang. (2009). "Ranking Measures and Loss Functions in Learning to Rank." In *Advances in Neural Information Processing Systems*, 315–323.

Cleary, Duncan. (2011). "Predictive Analytics in the Public Sector: Using Data Mining to Assist Better Target Selection for Audit." *Electronic Journal of e-Government* 9(2): 132–140.

Cragg, John G. (1971). "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." *Econometrica: Journal of the Econometric Society* 39(5): 829–844.

Cyr, Dennis, Thomas Eckhardt, Lou Ann Sandoval, and Marvin Halldorson. (2002). "Predictors of Unreported Income: Test of Unreported Income (UI) DIF Scores." *2002 IRS Research Conference*. Internal Revenue Service, Washington, DC. Retrieved from https://www.irs.gov/pub/irs-soi/puidif2.pdf.

Friedman, Jerome H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29(5): 1189–1232.

González, Pamela Castellón, and Juan D. Velásquez. (2013). "Characterization and Detection of Taxpayers with False Invoices Using Data Mining Techniques." *Expert Systems with Applications* 40(5): 1427–1436.

Gupta, M., and V. Nagadevara. (2007). "Audit Selection Strategy for Improving Tax Compliance—Application of Data Mining Techniques." In: Agarwal, A., and V. Ramana Venkata (eds.): *Foundations of E-Government*. Computer Society of India, Hyderabad.

Hashimzade, Nigar, Gareth D. Myles, and Matthew D. Rablen. (2016). "Predictive Analytics and the Targeting of Audits." *Journal of Economic Behavior and Organization* 124: 130–145.

Hsu, Kuo-Wei, Nishith Pathak, Jaideep Srivastava, Greg Tschida, and Eric Bjorklund. (2015). "Data Mining Based Tax Audit Selection: A Case Study of a Pilot Project at the Minnesota Department of Revenue." *Real World Data Mining Applications*. Switzerland: Springer International Publishing, 221–245.

Internal Revenue Service. (2019). *Federal Tax Compliance Research: Tax Gap Estimates for Tax Years 2011–2013*. Publication 1415 (Rev. 9-2019). Research, Applied Analytics & Statistics. Washington, D.C. Retrieved from https://www.irs.gov/pub/irs-pdf/p1415.pdf.

Kirkos, S., and Yannis Manolopoulos. (2004). "Data Mining in Finance and Accounting: A Review of Current Research Trends." In *Proceedings of the 1st International Conference on Enterprise Systems and Accounting* (ICESA), 63–78.

Luttati, Carol M .(2006). "Collection." *Journal of Tax Practice and Procedure* 8: 7.

Micci-Barreca, Daniele, and Satheesh Ramachandran. (2004). "Improving Tax Administration with Data Mining." White paper. *Elite Analytics LLC*.

Prechelt, Lutz. "Early Stopping—But When?" (1998). In *Neural Networks: Tricks of the Trade*. Germany: Springer, Berlin, Heidelberg, 55–69.

Rettig, Charles P. (2016). "IRS Audit Selection and Classification Processes." *Journal of Tax Practice and Procedure* 18: 17.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. (2014). "Dropout: A Simple Way to Prevent Neural Networks From Overfitting." *The Journal of Machine Learning Research* 15(1): 1929–1958.

U.S. Government Accountability Office. (2012). "IRS Could Significantly Increase Revenues by Better Targeting Enforcement Resources." Report GAO-13–151. Washington, DC.

Webley, P., M. Cole, and O.P. Eidjar, O.P. (2001). "The Prediction of Self-Reported and Hypothetical Tax-Evasion: Evidence From England, France and Norway. *Journal of Economic Psychology.* 22(2): 141–155.

Wedick, John L. (1983). "Looking for a Needle in a Haystack–How the IRS Selects Returns for Audit." *The Tax Advisor,* 675–675.

Yang,Y., E. Ge, and R. Barns, R. (2011). "Towards Effective and Efficient Identification of Potential Tax Agent Compliance Risk: A Stratified Random Sampling Approach." *e-Journal of Tax Research* 9(1): 116–137.

Zou, Hui, and Trevor Hastie. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2): 301–320.

Appendix 1: Other Validation Plots

In this section, we present two more plots that we have used for validating models.

Symmetric Difference Plot: For any two methods, invariably there are individual tax returns selected by both methods. This plot directly compares alternate ranking methods against the status quo ranking method by showing as a function of rank the difference between a cumulative metric for the alternative ranking method and that same metric for the status quo at that same rank. Examples of these metrics are the total number of unique returns selected and the difference in revenue.

Figure 7 is a symmetric difference plot for the Schedule C expense category of correspondence audits in which the cumulative metric is the number of unique returns selected by alternate methods compared with the status quo ranking method at any given rank. For example, we identified the 10,000 returns ranked the highest by the status quo ranking method and the 10,000 returns ranked the highest by the Penalized Regression ranking method and compared the two sets of returns. This identified around 3,000 returns in the Penalized Regression ranking that were not included among the top 10,000 returns ranked by the status quo method. Hence, the Penalized Regression curve goes through the 3,000 level at the rank of 10,000.

Analysis of unique returns selected gives an indication of how different an alternative method performs when selecting available returns to audit. An important detail to note is that this analysis can be done before the audit process begins, as it does not require post-audit information. Comparing unique returns between two imperfect rankings does not require perfect knowledge ranking, we use this method to validate our ranking before testing them in operation. If the method drastically deviates from status quo, it would give us cause for concern. In Figure 7, the status quo method of ranking serves as the baseline against which all other methods are compared, so it is on the x-axis throughout. The perfect knowledge ranking ranks returns according to the actual value of the metric being compared; in this case, it is tax revenue assessed as a direct outcome of the audits. The three alternative methods of ranking were presented in the analytic approaches section of this paper. Penalized Regression is the most similar to current practice in terms of returns selected (therefore that curve is closest to the x-axis). Pairwise and Perfect Knowledge are very similar in their deviation from current practice, each selecting around 4,000 unique returns among the highest-ranked 10,000 returns selected when compared to current practice. Two-stage deviates the most from current practice, selecting around 5,500 unique returns among the highest-ranked 10,000 returns.

When plotting the difference in revenue (Figure 8), the symmetric difference plot graphs the vertical separation between the two methods in the lift plot, making the distinction between methods more apparent. This plot shows that there is more than a $20-million gap between the perfect knowledge ranking and current practice over the course of the ranking process; this shows the maximum potential gain that can be made by selecting returns better. Penalized regression shows the most improvement, which is consistent with the other validation plots, showing a maximum potential gain of slightly over $5 million. Pairwise shows a maximum gain of about $2.5 million. Two-stage shows more than a $10.5-million loss when compared to current practice.

FIGURE 7. Symmetric Difference Plot Showing the Count of Unique Returns Selected Compared to Status Quo as a Function of Rank

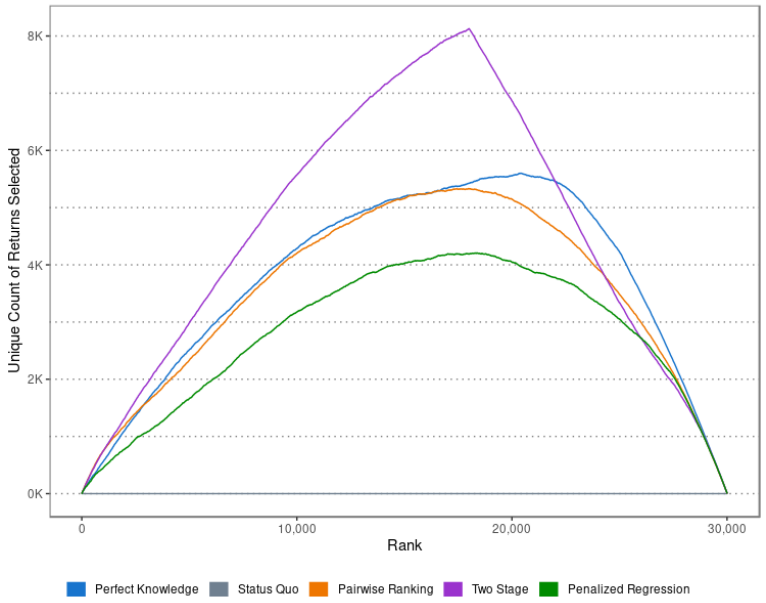
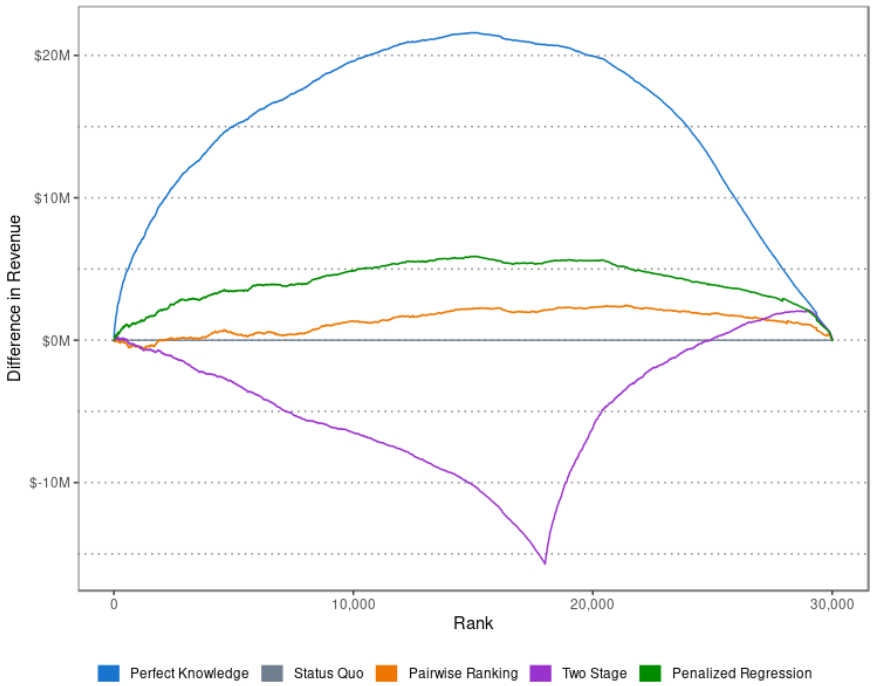


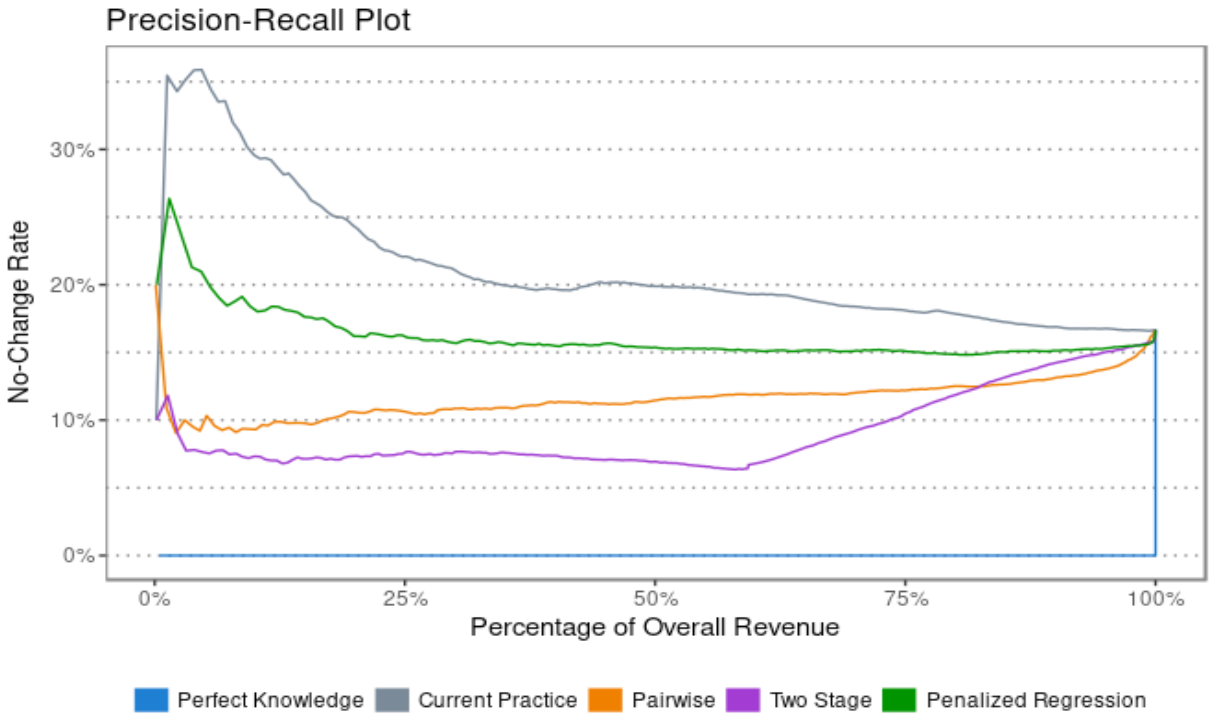
FIGURE 8. Symmetric Difference Plot Showing the Difference in Cumulative Revenue Between Alternate Methods and Status Quo as a Function of Rank



Precision-Recall Plot: Similar to the no-change progression plot, this validation plot gives a more sophisticated way to analyze how methods perform at avoiding no-changes. When labeling a return as a likely change or no-change, a 'cutoff' value in the model output (assuming continuous output) has to be determined where

anything to one side of the cutoff is labeled as a no-change and vice versa. For each possible cutoff, this plot calculates and graphs two metrics against each other, precision and recall. **Precision** is the number of true positives divided by the number of predicted positive cases and can be seen as analogous to the change rate, whereas the no-change rate is defined as (1 – precision). **Recall** is the number of true positives divided by the number of total positives in the data and can be interpreted as "available inventory of positive returns identified."

Figure 9. Precision-Recall Plot Showing No-Change Rate vs. Percentage of Overall Revenue



For Figure 9, we modify the recall metric to calculate the percentage of overall revenue identified at each potential cutoff point. While the takeaways are the same as the no-change progression plot, this plot scales the x-axis to the percentage of revenue captured. From this plot, we can say that two-stage maintains its low no-change rate but misses out on a large percentage of the overall revenue, while pairwise is able to capture a larger percentage of revenue while maintaining a relatively low no-change rate. Note that since the distribution of revenue tends to be skewed toward the low end, the percentage of revenue captured at a certain point is not necessarily equivalent to the percentage of revenue-bearing audits. While the choice for x-axis is dependent on preference, it may be helpful to generate plots with both types of x-axes.

Appendix 2. Other Analytic Approaches

In this section, we present two more approaches for balancing avoiding no-changes vs. finding cases with significant revenue.

Discounted Cumulative Gain (DCG)

DCG is a popular objective function coming from the learn-to-rank literature. It is an extension to Cumulative Gain (CG), which is the sum of relevance scores for a particular ranking of a dataset. DCG extends CG by logarithmically penalizing observations based on their ranking. Thus, highly relevant observations that appear low in the ranking will not contribute significantly to the sum. Given a set of observations with relevance scores, DCG seeks to produce a ranking that maximizes the function

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)},$$

where rel_i is a relevance score for the observation ranked at position i and p is a selected cutoff point in the ranking (Burges *et al.* (2005)). Note that the formulation of the numerator can vary—exponentiating relevance scores emphasizes bringing the most relevant observations to the top of the ranking. The choice of relevance score is up to the analyst. For example, in our use case it can be the revenue returned from an audit, the true rank in descending order, or a winsorized revenue to account for the audit distribution's wide tail.

DCG varies from pairwise ranking by incorporating the magnitude of the relevance score as well as the position in the ranking. While the optimal ranking for DCG will be the optimal ranking for pairwise (and vice versa), the objectives behave differently in cases where the data are not perfectly ordered. Consider a set of five observations with relevance scores {5, 2, 2, 1, 1}. We will compare two potential rankings of these data:

- Ranking A: {2, 5, 2, 1, 1}
- Ranking B: {5, 1, 1, 2, 2}

A pairwise objective would select Ranking A over Ranking B, since ranking A has only one misordered pair while Ranking B has two misordered pairs. A DCG objective, however, would select Ranking B over Ranking A. Using the DCG formulation above for the full set of observations, Ranking A sums to a DCG value of 24.9 while Ranking B sums to a DCG value of 35.6. It is evident that DCG's formulation favors rankings with the most relevant observations ranked highly, even at the expense of more mistakes elsewhere in the ranking.

While theoretically interesting, DCG doesn't work in practice for our use case as it places too much emphasis on correctly identifying highly relevant observations early on in the selection process that is motivated by the search engine use case. In order to be practical to audits, a smaller rate of convergence is required in the formulation's denominator.

FIGURE 10. Comparison of Denominators of DCG and Alternative Denominator

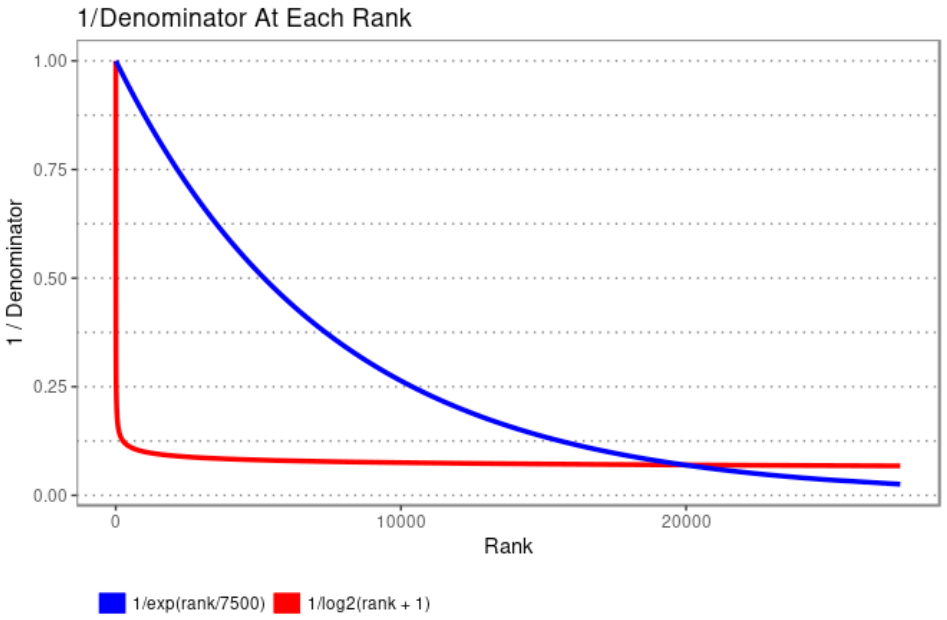


Figure 10 shows the standard denominator for DCG formulations in red, and an alternative denominator in blue, which shows a more desirable behavior of our use case. The standard denominator places disproportionate value on identifying relevant returns early on in the selection process and can lead to scenarios that select a final model during training, which correctly identifies a highly relevant return early on but does poorly in the remaining of the selection process. The alternative blue denominator gradually decreases over the course of the selection process and is more ideal for our use case. The blue denominator is for illustrative purposes only, would not scale the same to different audit sizes, and determining a denominator that is robust is an area for future research.

Exponential Mixture

The exponential mixture combines regression and classification models after training to produce a new ranking and can be viewed as a modified expected value. The effect of the exponential mixture is to prioritize predictions from the regression model early on in the selection process (even if the observations have a low change probability). Regression predictions are gradually penalized more heavily as the selection process progresses, which translates to more weight being placed on the probability of a return being a change. This allows the selection process to take risks on potentially high valued audits early on, but then account for the probability of no-change in the bulk of the selection pool.

The formulation for an exponential mixture, M(**x**) is as follows:

$$\mathrm{M}(\mathbf{x}) = f_c(\boldsymbol{x}) * (1 - \theta(r_x, N) \cdot g_{nc}(\boldsymbol{x})),$$

where

- $\boldsymbol{x}$ is the return
- $f_c(\boldsymbol{x})$ is the regression function assuming an audit on the return results in a change
- $\theta(r_x, N)$ is a penalty function ranging from [0,1] that decreases based on $\boldsymbol{x}$'s rank
- r_x is the rank of return x
- $\boldsymbol{N}$ is the max rank $\theta(r_x, N)$ will penalize $\boldsymbol{x}$.

Improving Taxpayer Response to Ineffective Audit Experiences: Service Messages as a Solution

Nina Collum (Louisiana Tech University), Susan Jurney (Oklahoma City University), and Mary Marshall (Louisiana Tech University)

Although the IRS continues to innovate by refining enforcement procedures, the consistently declining budget requires creative, minimal-cost solutions for improving detection results. By design, tax compliance analysis incorporates many elements of an increasingly complex system with many confounding factors. Thus, we leverage the experimental advantage with a simulated compliance setting that isolates the impact of decreased audit effectiveness, which we define as the amount of noncompliance detected during the investigation of a tax return, on subsequent compliance decisions of previously noncompliant taxpayers. This method enables us to disentangle the interconnected effects of varying audit rates, tax rates, and detection rates that are extremely difficult, if not impossible, to isolate in existing archival data. By isolating the effect of audit effectiveness, we not only establish a negative impact of decreased audit effectiveness, but we also identify and test the effect of a unique, low-cost solution.

Our study is primarily concerned with the effect of a taxpayer's audit in one year on the taxpayer's compliance in subsequent years. A number of prior studies have examined such indirect effects of audits with mixed results (Erard (1992); Bloomquist (2013); Mittone (2006); Maciejovsky *et al.* (2007); Alm *et al.* (2009); Kleven *et al.* (2011); Mittone *et al.* (2017); Hageman *et al.* (2020)). Consistent with prior findings that taxpayers incorporate information other than those included in Allingham and Sandmo's (1972) model into their compliance decisions across multiple periods, we test the unexplored effect of varied audit effectiveness on subsequent compliance.[1]

Prior research finds income tax audits increase compliance among noncompliant taxpayers by increasing the salience of the associated costs of that noncompliance (e.g., penalties, interest, and burden of experiencing the audit) (Boylan (2010); Kastlunger *et al.* (2011)). That is, research finds taxpayers who endure the costs of being audited are unlikely to risk experiencing an audit again in the subsequent year (Boylan (2010); Hageman *et al.* (2020)). However, we expect taxpayers will adjust their perceptions of being audited if they experience an audit that did not uncover all, or at least most, of their noncompliance. Thus, we predict and find taxpayers who experience a less (more) effective audit will decrease (increase) compliance, reducing tax collections and further compounding the continuing IRS budget cuts.

Although the simplest solution to decreased audit effectiveness is allocation of the enforcement resources necessary to support more effective audits, this is not a practical solution given the current budgetary environment. Instead, we examine another possible solution by assessing whether an increased focus on service-oriented messaging can offset the negative influence of declining audit effectiveness. A growing stream of academic literature suggests enforcement agencies should focus on balancing enforcement efforts with service-based efforts (e.g., Vossler and Gilpatric (2018); Hoffman *et al.* (2014); Alm *et al.* (2010)). Increased service efforts are not new for the IRS. The IRS provides tax assistance to taxpayers through its toll-free telephone helpline, its taxpayer assistance centers, and its website. The IRS also provides grants to IRS partner organizations for the Volunteer Income Tax Assistance (VITA) and Tax Counseling for the Elderly (TCE) programs.

[1] Allingham and Sandmo's (1972) model used audit rate, detection rate, and penalty rate to predict a taxpayer's expected utility for evasion.

Hoffman *et al*. (2014) suggest that including service efforts with enforcement efforts is more effective than either strategy by itself. Service efforts often require significant resources, which are unlikely to surface for new initiatives. Thus, we examine the effect of adding minimal cost service elements to an existing, but resource restricted, enforcement program. Further, research finds the mere presence of taxpayer service efforts can offset negative reactions to an audit, even if the taxpayer elects not to access the services. Thus, we propose a simple messaging effort to highlight the availability of services from the IRS.[2] As such, we test the effect of a service-minded "reminder" message that reinforces the IRS mission and the importance of tax collections for society. Encouragingly, results indicate taxpayers who experience a less effective audit are less likely to reduce compliance when they also view the service-oriented message, shown below:

> *THE MISSION OF THE INTERNAL REVENUE SERVICE*
>
> *Congress passes the tax laws and requires taxpayers to comply; however, the IRS is responsible for enforcing those laws. Thus, the IRS mission is to provide America's taxpayers top quality service by helping them understand and meet their tax responsibilities and enforce the law with integrity and fairness to all.*
>
> *The IRS wants to make it easier for you to make a complete and accurate return. We are here to give you advice and support if you need it.*

Our results also provide evidence of the psychological mechanism driving the effect of both audit effectiveness and service messaging on subsequent compliance. Specifically, we find that when audits are less effective, they can lead taxpayers to focus less on fulfilling their obligation to society (compared to when they experience audits that are more effective). In addition, we also find taxpayers view the IRS as more focused on service than on punishment when they *either* experience more audit effectiveness and/or view a service message following an audit.

Our experiment included a 2x2 between-subjects experiment distributed to U.S. taxpayers solicited from Amazon's Mechanical Turk, a research participant recruitment platform. In the experiment, participants performed an earnings task and decided how much income to report to a tax authority after each of the three experimental rounds. Audit effectiveness (100 percent of noncompliance is detected vs. 50 percent of noncompliance is detected) was manipulated after the second round and the service message (present vs. absent) was manipulated reporting second-round earnings and before completing the third-round reporting task. By manipulating two levels of audit effectiveness, our study draws conclusions about the effects of lower or higher effectiveness. The experimental method also benefits from the advantage of random assignment. Consequently, we were able to control for individual characteristics that may have been influenced if the experiment had higher stakes.[3]

Our findings contribute to theory in two primary ways. First, we extend the literature on the effect of audits on tax compliance. Prior literature finds audits increase tax compliance among those who were initially noncompliant (Boylan (2010)) and decrease tax compliance among those who were initially compliant (Hageman *et al.* (2020)). We add to this literature by examining varying audit effectiveness levels. Specifically, we highlight boundary conditions for prior findings in the current budgetary environment where audits should not be expected to reach full effectiveness.[4] Second, we add to a growing literature on the importance

[2] Importantly, we tested the effect of two service messages, one focused on reminding taxpayers of the availability of assistance and another focused on notifying the taxpayers of the "Taxpayer Bill of Rights." There were no significant differences between the two messages, suggesting that the exact form of the message is less important than the presence of it.

[3] Some argue results of experimental economics studies are limited to scenarios with similarly minimal compensation. However, the individual characteristics that might be influenced by higher stakes (e.g., risk preferences) are randomly assigned across conditions. Thus, standard practice assumes any effects of stakes would equally influence all conditions. If this influenced the results in any way, it would be to shift the means of all conditions. Any differences across conditions would remain.

[4] We acknowledge the idea of 100-percent efficiency is unlikely even with surplus resources; however, we focus on the differences between full efficiency, which may be expected by a taxpayer who has never been audited, and less efficiency, which is likely what taxpayers experience in practice.

of establishing both enforcement and service activities in a tax agency's operations. Although prior literature has examined the effects of both enforcement and service, we experimentally disentangle the interactive effects of the two items.

Our findings also inform policymakers, specifically those who are charged with increasing tax collections despite the increasingly scarce resources available to them for enforcement. By identifying a less costly option to increase taxpayer morale towards the agency, we provide a practical option for tax agencies to implement to offset the decreased enforcement resources. Although our study focuses on the effects of budget crises on the IRS's collections of Federal income tax, it is important to note that most State tax enforcement agencies face similar issues. Thus, our study informs tax enforcement at many levels.

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., T. Cherry, M. Jones, and M. McKee. 2010. Taxpayer information assistance services and tax compliance behavior. *Journal of Economic Psychology* 31(4): 577–586.

Alm, J., B. R. Jackson, and M. McKee. 2009. Getting the word out: Enforcement information dissemination and compliance behavior. *Journal of Public Economics* 93(3–4): 392–402.

Bloomquist, K. 2013. Incorporating indirect effects in audit case selection: An agent-based approach. Office of IRS Research.

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.

Erard, B. 1992. The influence of tax audits on reporting behavior. In J. Slemrod (Ed.) Why People Pay Taxes (pp. 95–115), Ann Arbor, MI: University of Michigan Press.

Hageman, A.M., E. G. LaMothe, and M. E. Marshall. 2020. The effect of audit burden on subsequent tax evasion. Working paper. Kansas State University.

Hoffman, E., K. Gangle, E. Kirchler, and J. Stark. 2014. Enhancing tax compliance through coercive and legitimate power of tax authorities by concurrently diminishing or facilitating trust in tax authorities. Law & Policy 36 (3): 290–313.

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.

Kleven, H. J., M. B. Knudsen, C. T. Kreiner, S. Pedersen, and E. Saez. 2011. Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark. *Econometrica* 79 (3): 651–692.

Maciejovsky, B., E. Kirchler, and H. Schwarzenberger. 2007. Misperception of chance and loss repair: On the dynamics of tax compliance. *Journal of Economic Psychology* 28 (6): 678–691.

Mittone, L., F. Panebianco, and A. Santoro. 2017. The bomb-crater effect of tax audits: Beyond the misperception of chance. *Journal of Economic Psychology* 61: 225–243.

Mittone, L. 2006. Dynamic behaviour in tax evasion: An experimental approach. *Journal of Socio-Economics* 35(5): 813–835.

National Taxpayer Advocate. 2019. 2019 Annual Report to Congress. https://www.taxpayeradvocate.irs.gov/reports/2019-annual-report-to-congress/.

Transactional Records Access Clearinghouse (TRAC). 2019. Millionaires and corporate giants escaped IRS audits in FY 2018. https://trac.syr.edu/tracirs/latest/549/.

Vossler, C.A. and Gilpatric, S.M. (2018). Endogenous audits, uncertainty, and taxpayer assistance services: Theory and experiments. *Journal of Public Economics* 165: 217–229.

Using the Internal Revenue Service Program Assessment Model Optimizer To Inform Resource Allocation Decisions

Rafael Dacal, Chris Lee, Deandra Reinhart, Sarah Shipley, Clay Swanson, and Ariel S. Wooten
(IRS, Small Business/Self-Employed Division)

As the Internal Revenue Service (IRS) continues to modernize, it is imperative that resource allocations are assigned in a logical and data-supported method. To that end, the IRS's Small Business/Self-Employed (SB/SE) Research and MITRE Corporation developed a Program Assessment Model Optimizer (PAM) to allocate new SB/SE staffing based on current staffing, downstream interactions, and enforcement tax revenue collected within critical IRS processes.

PAM is designed to allocate new resources to these IRS processes with a primary goal of maximizing revenue. As an example, suppose the IRS was given a specific number ("FTE_{max}") of new full-time employees[1] to allocate across all compliance programs within the SB/SE Division. There are many possible allocations of the new FTEs among SB/SE compliance programs. The programs interact with each other, so it is important to be aware that adding FTEs to one program may impact the need for FTEs in other programs. PAM accounts for these variables and interactions and calculates the optimal FTE allocations to maximize potential enforcement tax revenue collected and assist SB/SE decision-makers.

PAM uses linear programming, a tool for solving complex optimization problems. There are four basic steps for setting up a linear programming problem:[2]

1. Identify and label the decision variables at each process step;
2. Determine the objective and use the decision variables to write an expression for the objective function as a linear function of the decision variables;
3. Determine the explicit constraints and write a functional expression for each of them as either a linear equation or a linear inequality in the decision variables; and
4. Determine the implicit constraints and write each as either a linear equation or a linear inequality in the decision variables.

The goal of PAM is to allocate FTEs across all SB/SE enforcement programs at an optimal level. With this knowledge, we can define an example decision variable as FTE_p, where p represents an enforcement program.

To set up the objective function, we need to decide what specific measure to optimize. PAM can optimize on any measure that has available and comparable data for all Examination and Collection functions. For the purposes of this example, we will allocate FTEs to maximize enforcement tax revenue collected. If we know the enforcement tax revenue collected per FTE within each compliance program ($revenue_p$), we can define the objective function as:

$$\max \sum_{p} revenue_p FTE_p$$

[1] Full-time employees will be expressed as full-time equivalents (FTEs), which represent the workload for one full-time employee in one calendar year.

[2] James Burke, *Linear Optimization*, available at https://sites.math.washington.edu/~burke/crs/407/notes/section1-18.pdf.

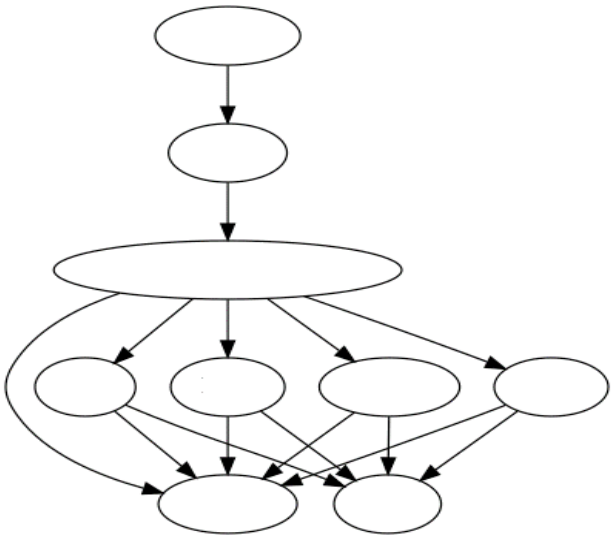
As FTE_p increases, the objective function grows larger and the IRS can collect an arbitrarily large amount of enforcement tax revenue by choosing very large numbers of FTEs. However, the values of FTE_p are limited by various constraints. An obvious constraint is that the sum of all newly allocated FTEs must be less than or equal to the maximum number of FTEs available for allocation:

$$\sum_p FTE_p \leq FTE_{max}$$

Enforcement programs are connected within the overall IRS enforcement system, creating additional constraints for PAM. Tax return modules[3] move between programs, so the same module can expend FTEs and produce tax revenue within multiple programs. Adding FTEs to work tax modules in one program may result in increasing the work for one or more other programs as the modules flow through the compliance system. Therefore, each program affected by the increased workload may require additional FTEs.

PAM models this interaction by expressing the compliance system as a network of nodes and arcs. In the example network flow shown in Figure 1, the nodes (ovals) represent "steps" and the arcs (connecting lines) show the flow[4] of "commodities" between steps.

FIGURE 1. Example Network Flow



In PAM, a "step" is defined as a discrete point where work is performed, revenue is realized, and/or work is routed to other steps. Most steps represent the IRS enforcement programs eligible to receive new FTEs, but a few steps are created solely for routing purposes.

A "commodity" is a type of work that usually represents a particular group of tax modules (for example, employment tax returns). FTEs within a step can work multiple types of commodities, so it is important to use the step-commodity combination to distinguish the type of tax modules worked within each step.

Each step-commodity combination has its own set of characteristics calculated from historical data, including:

- Enforcement tax revenue collected per tax module;
- Work rate (tax modules worked per FTE);

[3] A tax return module is a filed tax return.

[4] Flows carry commodities that are either "required" inventory or "discretionary" inventory. Constraint 4 in Table 1 states that all required inventory must be executed. Constraint 5 states that discretionary inventory is either worked or abandoned.

- Transition rate at which tax modules move to other steps;
- Current FTEs (before new FTEs are allocated); and
- Available tax module inventory.

PAM does not model individual tax modules; everything is treated as continuous flows. As a result, all work within a commodity is considered to be homogenous modules and all calculated characteristics are multiyear averages with standard deviations. The tax module characteristics are constant, regardless of the volume of work completed. The first module worked always yields the same revenue as the final module worked. This simplification is a key assumption made by PAM.

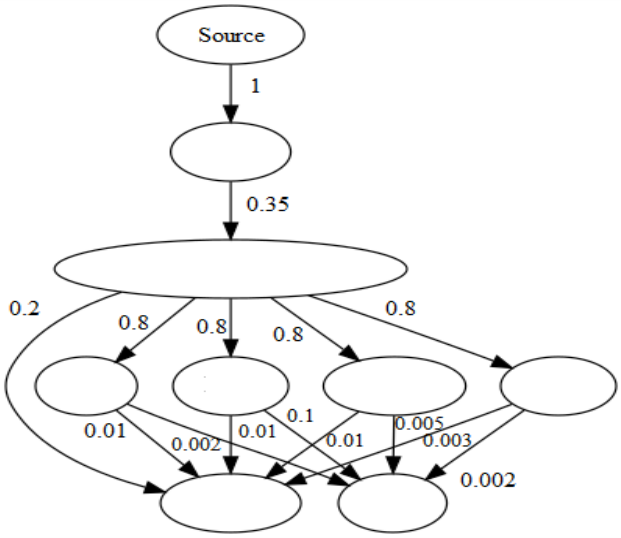
In reality this assumption does not always hold, so commodity splitting is necessary if there is reason to believe that the group being represented is not actually homogenous. There are many possible causes for non-homogeneity, including:

- *Decreasing marginal returns*: the enforcement tax revenue generated by one additional module worked is less than the average revenue of modules currently worked;
- *Different work rates*: an identifiable subset of modules can be worked faster or slower than the rest; and
- *Different routing*: an identifiable subset of modules is routed differently than the rest.

For the larger commodities, we split the original commodity into multiple new commodities or "subcommodities." All modules within a subcommodity are assumed homogenous.

Consider the possible paths of a particular commodity through the sample compliance system in Figure 2. Each tax module of this commodity starts at the Source node and travels the network using a path defined by historical rates. The number next to each arc represents the "transition rate" or percentage of modules that travel to the next step. Not shown (but also present) are the enforcement tax revenue collected and FTE time used by each module at each step.

FIGURE 2. Example Network Flow with Transition Rates



Given this information, we can calculate the impact of adding one new FTE to a specific step to work modules of this particular commodity. The impact is measured by the following outcomes:

- Number of new tax modules worked at this step;
- Amount of tax revenue collected from these tax modules at this step;
- Number of tax modules flowing to future steps;

- Number of FTEs needed to work these modules at future steps, and
- Amount of revenue collected from these tax modules at future steps.

PAM runs similar calculations within a linear programming framework and a multicommodity network flow model to determine the FTE allocation that will maximize the enforcement tax revenue collected. The following four stages are necessary for running PAM, given we are trying to allocate x new FTEs.

1. **Optimization**

   Step-commodity characteristic[5] averages and standard deviations calculated using historical data may not represent current characteristics. To mitigate this uncertainty, PAM randomly generates characteristics using the historically calculated averages and standard deviations. Given this set of characteristics, PAM calculates the optimal allocation of x FTEs. We generally repeat this stage 1,000 times[6] to create 1,000 distinct optimal solutions using randomly generated step-commodity characteristics for each iteration.

2. **Cluster Analysis**

   Assume we generated 1,000 optimal FTE allocations in Stage 1. Some of the allocations may be very similar. Rather than evaluate all of them, Stage 2 creates clusters of similar allocations and averages the FTE allocations in each cluster. This gives us a smaller set of representative solutions to evaluate.

3. **Reoptimization**

   Assume we created five clusters of FTE allocations in Stage 2. Once again, PAM randomly generates step-commodity characteristics and under these conditions, calculates the expected enforcement tax revenue collected for each cluster's FTE allocation. This process allows the potential solutions to "compete" against each other to identify the FTE allocation producing the most revenue. We repeat this stage multiple times to allow for competitions under different step-commodity characteristics. For example, we may run 100 iterations, generating expected revenue for each cluster 100 times and logging the number of "wins" for each cluster. By running PAM repeatedly with randomized inputs, we can identify which solutions are robust.

4. **Results Assessment**

   Assume we ran 100 cluster competitions for our five clusters in Stage 3. The average revenue per cluster and the number of cluster "wins" are presented in tables and graphs.[7] Some FTE allocations may be very consistent, while others may have high variability, producing high revenue under some conditions but low revenue under others. These estimates allow the decision-makers to assess the benefits of the five FTE allocations.

Table 1 summarizes the PAM objective function, and the explicit and implicit constraints.

[5] Revenue, work rate, transition rate, existing FTEs, available inventory.

[6] Each iteration runs in less than one second.

[7] In general, the calculations in PAM are not rounded but the results are rounded to the nearest integer.

TABLE 1. PAM Objective Function and Constraints

Objective Function: Maximize Revenue	
Explicit constraints	
1	Executed work is within capacity of workforce.
2	Total new hires are less than some user-specified upper bound.
3	Program utilization must be greater than some user-specified rate.
4	All required work must be executed.
5	Discretionary inventory is either worked or abandoned.
6	Total work is the sum of required and discretionary work executed.
7	Outbound required work is a fixed fraction of all cases worked at a step.
8	If a required commodity can be worked in more than one place, it is worked in only one of them.
9	Outbound discretionary work can be picked up in at most one next step.
10	Execution cost constraint
11	New hire cost constraint
Implicit constraints	
	All variables assumed to be continuous and non-negative.

Limitations

As noted earlier, a key assumption made by PAM is that all modules within a subcommodity are considered to be homogenous, so the first module worked yields the same revenue as the final module worked. In reality, most programs prioritize modules within their inventory according to the likely extent to which they will yield enforcement revenue. This means that the impact of increasing or decreasing FTEs in those programs is governed by the marginal relationship between revenue and FTEs—not the average relationship. PAM creates subcommodities of modules (using expected enforcement revenue) to help approximate the marginal relationship. For example, all modules in Subcommodity A are homogeneous, but the average enforcement revenue for modules in Subcommodity A is higher than the average enforcement revenue for modules in Subcommodity B.

Currently, PAM calculates FTE allocations without taking into account the cost of those FTEs. In reality, FTEs in one program (e.g., experienced revenue agents) are much more costly (due to job series, pay grade level, and experience) than in another program (e.g., tax examiners who conduct correspondence audits). A potential enhancement could consider the varying cost of the FTE resources for each subcommodity.

The ultimate objective of the IRS is to maximize overall tax revenue for any given budget—not just enforcement revenue. To the extent that enforcement actions change future voluntary payment of tax revenue, PAM could ideally take into account the impact enforcement actions have on future voluntary revenue as well.

5

Appendix

Conference Program

10th Annual IRS-TPC Joint Research Conference on Tax Administration
Held Virtually
June 18, 2020

Program

9:30–9:40 Opening

Eric Toder (Codirector, Urban-Brookings Tax Policy Center) and

Barry Johnson (Director, Statistics of Income, IRS, RAAS)

9:40-11:10 Behavioral Responses to Audits

Moderator: Robert McClelland (Urban-Brookings Tax Policy Center)

- The Specific Deterrence Implications of Increased Reliance on Correspondence Audits
  Brian Erard (B. Erard & Associates), Erich Kirchler, and Jerome Olsen (University of Vienna)
- The Specific Indirect Effect of Correspondence Audits: Moving from Research to Operational Application
  Leigh Nicholl, Maxwell McGill, Lucia Lykke (MITRE Corporation), and Alan Plumley (IRS, RAAS)
- The Effect of Audit Risk and Detection Risk on Tax Compliance
  James Alm and Matthias Kasper (Tulane University)

Discussant: *Janet Holtzblatt (Urban-Brookings Tax Policy Center)*

11:10–12:20 New Insights on Taxpayer Behavior

Moderator: Brett Collins (IRS, RAAS)

- Size, Characteristics, and Distributional Effects of Income Tax Evasion in Italy
  Martina Bazzoli (IRVAPP), Paolo Di Caro and Marco Manzo (Italian Dept. of Finance), Francesco Figari (Univ. of Insubria, University of Essex), and Carlo Fiorio (Univ. of Milan)
- Taxpayer Responses to Third-party Income Reporting: Evidence from Spatial Variation Across the U.S.
  Bibek Adhikari and Timothy F. Harris (Illinois State University), and James Alm (Tulane University)
- The Effects of an Employment Tax Enforcement Regime on U.S. Small Business and Proprietor Payment Compliance
  Rafael Dacal (IRS, SB/SE)

Discussant: *Jamie McGuire (Joint Committee on Taxation)*

12:20–12:30 Break

12:30–1:15 Keynote Speaker

Charles Rossotti (former IRS Commissioner)

1:15–2:45 Advances in Taxpayer Service

Moderator: Fran Cappelletti (IRS, TAS)

- Free Assisted Tax Preparation Outreach Experiments
  Rizwan Javaid and Brenda Schafer (IRS, RAAS), Jacob Goldin (Stanford University), Tatiana Homonoff (New York University), and Adam Isen (Department of the Treasury)
- Enforcement Versus Outreach - Impacts on Taxpayer Burden
  Anne Herlache, Stacy Orlett, Ishani Roy and Alex Turk (IRS, RAAS)
- Perspectives on New Forms of Remote Identity Proofing and Authentication for IRS Online Services
  Rebecca Scollan and Ronna Ten Brink (MITRE Corporation)

Discussant: *Mary-Helen Risler (IRS, RAAS)*

2:45–4:15 Doing More With Less

Moderator: *Tom Hertz (IRS, RAAS)*

- Can Machine Learning Improve Correspondence Audit Case Selection? Considerations for Algorithm Selection, Validation, and Experimentation
  Ben Howard and Lucia Lykke (MITRE), and Alan Plumley (IRS, RAAS)
- Audit Productivity, Taxpayer Service, and Compliance: Can a Service Mindset Overcome a Dwindling Enforcement Budget?
  Nina Collum and Mary Marshall (Louisiana Tech Univiversity), and Susan Jurney (Oklahoma City University)
- Using the Internal Revenue Service Program Assessment Model Optimizer To Inform Resource Allocation Decisions
  Deandra Reinhart and Clay Swanson (IRS, SB/SE)

Discussant: *Arnie Greenland (A. G. Analytics, LLC)*

4:15–4:20 Wrap-up

Eric Toder (Codirector, Urban-Brookings Tax Policy Center)