Analysis Plan

Project Name: Reducing filing errors via outreach to tax preparers and clients

Project Code: 2504

Date Finalized: 4/15/2025

Project description

In 2023, the Internal Revenue Service (IRS) estimated over $24 billion dollars in overpayments from three high-priority refundable tax credit programs: the Additional Child Tax Credit, American Opportunity Tax Credit, and Earned Income Tax Credit. The majority of returns that claim these benefits are prepared by paid tax preparers. This evaluation builds evidence that will help the IRS make data-informed decisions to continuously improve their education and outreach efforts under the Return Integrity and Compliance Services (RICS) Return Preparer Strategy to enforce tax compliance among tax return preparers.

The goals of this evaluation are threefold. Our first goal is to quantify the effects of two different types of outreach: outreach to tax preparers (in the form of a phone call and a letter) versus coupling this tax preparer outreach with letters sent directly to clients of tax preparers. A second aim is to understand how the effects of client outreach differ when most clients of a tax preparer are sent letters (i.e., the high-saturation group) or when only a few clients of a preparer are sent letters (i.e., the low-saturation group). The third aim of this evaluation is to understand the extent to which the distribution method for letter-based outreach induces different effects.

Evaluation design

There are two primary interventions to be evaluated in this project. The first intervention is a client letter. In one evaluation design focused only on the client letter, we evaluate how the saturation of client letters within a preparer's client pool and the client letter distribution method influences tax compliance. The sample consists of clients of tax preparers designated by the IRS as "Group 1 and Group 2" tax preparers.[1] There are three main treatment arms in this evaluation design:

1. **Control group**: Tax preparers randomized into this treatment arm do not receive any outreach nor do their clients.

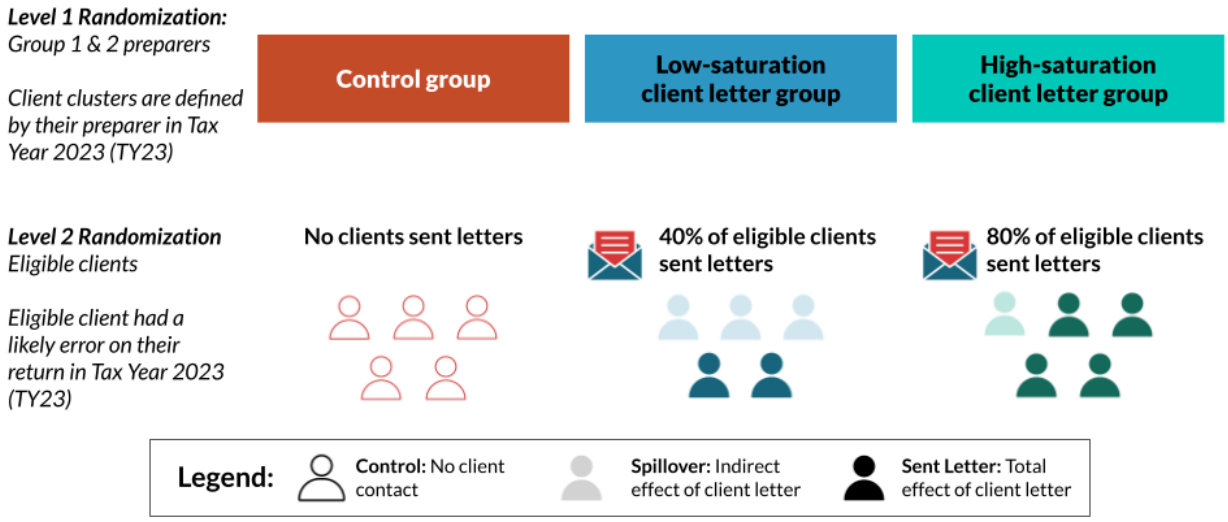
[1] IRS groups preparers into different groups depending on the types of outreach they have received in the past (which affects their eligibility for outreach in this tax season). We use the same language for convenience, but the group numbers are not meaningful for the evaluation.

2. **High-saturation client letter group:** Tax preparers randomized into this treatment arm had 80% of their eligible clients (client who submitted a TY 2023 return with at least one likely error when claiming certain refundable tax credits) sent the client letter. Tax preparers are not sent any outreach. Within this client letter group, there are three sub-groups of clients:
    a. **Client sent letter via the National Distribution Center (NDC) ("NDC high-saturation client letter group")**: Clients are sent a client letter as a mailed letter via NDC. These include clients who do not have an online account and those who have an online account and were randomized to be sent a letter via NDC.
    b. **Client sent letter via Online Accounts (OLA) ("OLA high-saturation client letter group")**: Clients are sent a client letter as a mailed notice via OLA. These include clients who have an online account and were randomized to be sent a letter via OLA.
    c. **Spillover ("high-saturation spillover group"):** Clients are not sent a client letter but their preparer was randomized to the client letter group.[2]
3. **Low-saturation group:** Tax preparers randomized into this treatment arm had 40% of their eligible clients (client submitted a TY 2023 return with at least one likely error when claiming certain refundable tax credits) sent a client letter. Tax preparers are not sent any outreach. Within this client letter group, there are three sub-groups of clients:
    a. **Client sent letter via NDC ("low-saturation client letter NDC group")**: Clients are sent a client letter as a mailed letter via NDC. These include clients who do not have an online account and those who have an online account and were randomized to be sent a letter via NDC.
    b. **Client sent letter via OLA ("low-saturation client letter OLA group")**: Clients are sent a client letter as a mailed notice via OLA. These only include clients who have an online account and were randomized to be sent a letter via OLA.
    c. **Spillover ("low-saturation spillover group")**: Clients are not sent a client letter but their preparer was randomized to the client letter group and therefore has other clients who did receive it.

[2] Note that this group *in theory* includes both clients who were eligible to be sent a client letter due to errors in their own returns, and clients of the same preparer who were not eligible to be sent the letter. However, we only observe clients who were eligible to be sent the letter. Randomization occurred only within this subpopulation of a preparer's clients, and when we say that 80% (or 40%) of the preparer's clients were randomized to be sent a letter, those percentages are out of the eligible population of the preparer's clients, not out of all clients of a given preparer. When we refer to the spillover group, we are referring only to eligible clients who were not sent letters.

Figure 1 provides a graphical representation of this evaluation design.

Figure 1. Evaluation design for preparers in Groups 1 and 2 and their Filing Season (FS) 2023 clients



The second intervention entails tax preparer outreach in the form of a preparer letter pre-announcing a phone call followed by a phone call. In a second evaluation design (shown in Figure 2), we evaluate how the combination of the client letter and preparer outreach affects tax compliance. Again, we also evaluate how the client letter distribution method influences tax compliance (shown in Figure 3). The sample consists of Group 5 tax preparers (grouping is defined by the IRS). There are three main treatment arms in this evaluation design:

1. **Control group:** Tax preparers randomized into this treatment arm do not receive any outreach nor do their clients.
2. **Preparer call group:** Tax preparers randomized into this treatment arm are sent the pre-call preparer letter and are called. Clients of these tax preparers do not receive any outreach directly from the IRS.
3. **Preparer call + letter group:** Tax preparers randomized into this treatment arm are sent the pre-call preparer letter and called. Additionally, 40% of the eligible clients (clients who submitted a TY 2023 return with at least one likely error when claiming certain refundable tax credits) of these tax preparers are sent a client letter.
    a. **Client sent letter via NDC ("NDC client letter group"):** Clients are sent a client letter as a mailed letter via NDC. These include clients who do not have an online account and those who have an online account and were randomized to be sent a letter via NDC.
    b. **Client sent letter via OLA ("OLA client letter group"):** Clients are sent a client letter as a mailed notice via OLA. These include clients who have an online account and were randomized to be sent a letter via OLA.

c. **Spillover ("spillover group"):)** Clients are not sent a client letter but their preparer was randomized to the client letter group and therefore has other clients who did receive it.

Figure 2: Evaluation design for preparers in Group 5 and their filing season (FS) 2024 clients

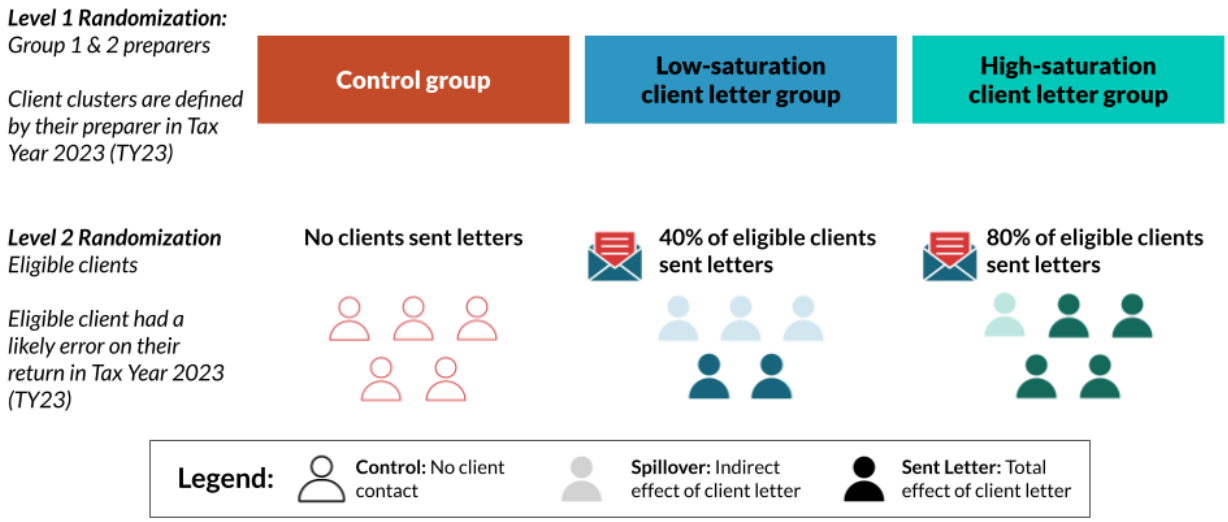
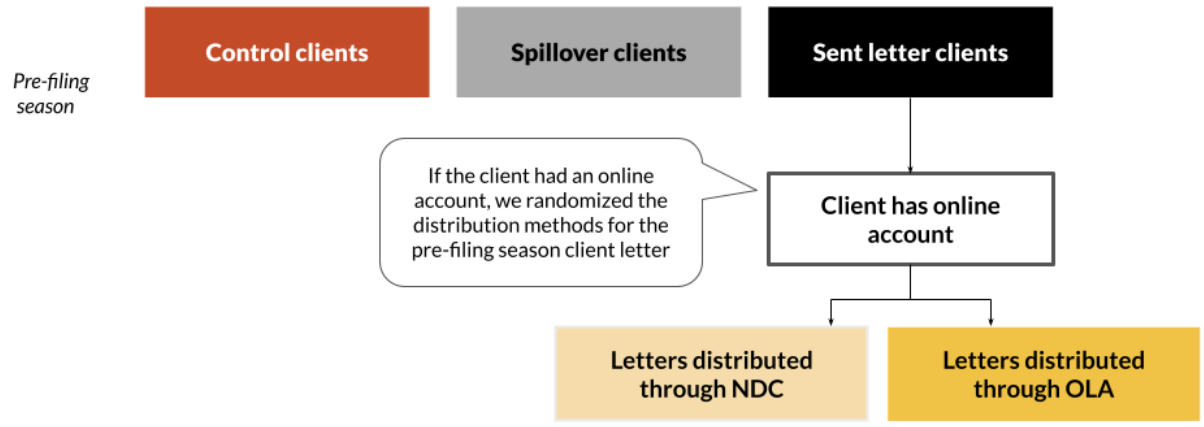


Figure 3: Evaluation design of impacts of distribution method for client letters



Preregistration details

This Analysis Plan will be posted on the OES website at oes.gsa.gov before outcome data are analyzed.

Hypotheses

Clients of Group 1 and 2 Preparers

Research question 1 (primary): Does being in a client letter group affect tax compliance?

Hypothesis 1A: Compared to the control group, clients assigned to a client letter group (whether they are sent the letter themselves or are spillover clients) have different tax compliance.

Hypothesis 1B: Compared to the control group, clients assigned to the high-saturation client letter group have different tax compliance.

Hypothesis 1C: Compared to the control group, clients assigned to the low-saturation client letter group have different tax compliance.

Research question 2 (primary): Does being sent a letter affect tax compliance?

Hypothesis 2A: Compared to the control group, clients sent a letter have different tax compliance.

Hypothesis 2B: Compared to the control group, clients sent a letter in the high-saturation client letter group have different tax compliance.

Hypothesis 2C: Compared to the control group, clients sent a letter in the low-saturation client letter group have different tax compliance.

Research question 3 (primary): What is the spillover effect of being in a client letter group, but not being sent a letter, on tax compliance?

Hypothesis 3A: Compared to the control group, clients assigned to a spillover group have different tax compliance.

Hypothesis 3B: Compared to the control group, clients assigned to the high-saturation spillover group have different tax compliance.

Hypothesis 3C: Compared to the control group, clients assigned to the low-saturation spillover group have different tax compliance.

Research question 4 (primary): How do the effects on tax compliance differ when most clients are sent letters versus fewer clients are sent letters?

Hypothesis 4A: Compared to the effect on tax compliance in the low-saturation client letter group, the effect on tax compliance is different among clients assigned to the high-saturation client letter group.

Hypothesis 4B: Compared to the effect on tax compliance among clients sent a letter in the low-saturation client letter group, the effect on tax compliance is different among clients sent a letter in the high-saturation client letter group.

Hypothesis 4C: Compared to the effect on tax compliance in the low-saturation spillover group, the effect on tax compliance is different among clients assigned to the high-saturation spillover group.

<u>Clients of group 5 Preparers</u>

Research question 5 (primary): Does tax preparer outreach affect tax compliance?

Hypothesis 5A: Compared to the control group, clients of preparers assigned to the preparer call or call + letter groups have different tax compliance.

Hypothesis 5B: Compared to the control group, clients of preparers assigned to the preparer call group have different tax compliance.

Hypothesis 5C: Compared to the control group, clients of preparers assigned to the call + letter group have different tax compliance.

Hypothesis 5D: Compared to the effect on tax compliance in the preparer call group, the effect on tax compliance is different among clients of preparers assigned to the call + letter group.

Clients of groups 1, 2, and 5 Preparers

Research question 6 (primary): Are there differences in tax compliance by client outreach distribution method?

Hypothesis 6A: Among clients who have an online account, those who were randomized to receive client outreach via OLA will have different tax compliance than those who were randomized to receive client outreach via NDC.

Data and data structure

This section describes variables that will be analyzed, as well as changes that will be made to the raw data with respect to data structure and variables.

Data source(s):

Our primary data source will be processed, return-level data that Taxpayer Services (TS) pulls for return preparers and clients (i.e., returns) at the end of the 2025 filing season (by the end of June 2025). The primary time periods will be TY 2023 returns filed during the 2024 filing season (for pre-treatment covariates and blocking) and TY 2024 returns filed during the 2025 filing season (for outcomes).

Outcomes to be analyzed:

We plan to measure outcomes at the client level. Client-level outcomes will be measured among clients who used a high-risk preparer (IRS Group 1, Group 2, and Group 5) and submitted a TY 2023 return with at least one likely error when claiming certain refundable tax credits.[3] We will track outcomes for these clients regardless of the preparer they use to file their TY 2024 return. Since there are multiple approaches to aggregate outcomes from the client/return level to the tax preparer level, we explore preparer-level aggregations in the secondary/exploratory analysis.

Primary outcomes:

We have two primary outcomes. The outcomes defined below are calculated at the client-level.

1. **Refund amount:** a continuous numeric variable reflecting the return-level refund amount. For the purposes of this study, this measure will be used as an estimate of protected revenue. Refund amount is imputed to be $0 if the client does not file TY 2024 in the 2025 filing season.

[3] In the secondary/exploratory analysis, we propose expanding the set of clients to include both eligible and ineligible clients of high-risk tax preparers (IRS Group 1, Group 2, and Group 5).

2. **Sum of erroneous dollars:** A continuous numeric variable that equals the benefit amount for credits and benefits claimed with likely errors. Note that HOH filing status error is not included in this summation since filing status does not correspond to a credit/benefit amount. Furthermore, this summation does not distinguish how much of a particular tax credit is associated with an erroneous amount and non-erroneous amount. Sum of likely erroneous dollars will be imputed as $0 if the client does not file TY 2024 in the 2025 filing season or does not claim benefits on their TY 2024 return.

Secondary outcomes:

We are also interested in examining the following secondary outcomes:

1. **Tax benefit error:** a binary variable that is equal to one if the client files a return that contains one or more likely errors in claiming certain benefits, which for the purposes of this study include: the earned income tax credit (EITC), child tax credit/additional tax credit/credit for other dependents (CTC/ACTC/ODC), American opportunity tax credit (AOTC) and head of household (HOH) filing status.[4] This variable is equal to zero otherwise, including if the client does not claim these benefits or does not file a tax return.
2. **Earned income tax credit (EITC) error**: a binary variable that is equal to one if the return contains a likely error when claiming EITC and is equal to zero otherwise. This variable will be imputed as zero if the client does not file TY 2024 during the 2025 tax filing season or does not claim the EITC.
3. **American opportunity tax credit (AOTC) error**: a binary variable that is equal to one if the return contains a likely error when claiming AOTC and is equal to zero otherwise. This variable will be imputed as zero if the client does not file TY 2024 during the 2025 tax filing season or does not claim the AOTC.
4. **Combined child tax credit error:** a binary variable that is equal to one if the return contains a likely error when claiming the ACTC/CTC/ODC and is equal to zero otherwise. This variable will be imputed as zero if the client does not file TY 2024 during the 2025 tax filing season or does not claim the ACTC/CTC/ODC.
5. **Head of household error**: a binary variable that is equal to one if the return contains a likely error when claiming a head of household filing status (HOH) and is equal to zero otherwise. This variable will be imputed as zero if the client does not file TY 2024 return during the 2025 tax filing season or does not claim head of household filing status.
6. **Change in filing method ("any method change"):** a binary variable equal to one if a client changed their tax return filing method. This includes if a client used a different return preparer than the return preparer they used during the 2024 filing, self-filed, or did not file. This variable is equal to zero otherwise.

[4] https://www.eitc.irs.gov/tax-preparer-toolkit/preparer-compliance-focused-and-tiered/compliance

7. **Self-file**: a binary variable that is equal to one if the client filed their own tax return during the 2025 filing season and is equal to zero otherwise. This variable will be imputed as zero if the client does not file TY 2024 during the 2025 tax filing season.
8. **Did not file:** a binary variable that is equal to one if the client did not file TY 2024 during the 2025 filing season and is equal to zero otherwise.
9. **Change in preparer:** a binary variable that is equal to one if a client filed using a different return preparer than the return preparer they used during the 2024 filing season and is equal to zero otherwise. This variable will be imputed as zero if the client does not file TY 2024 during the 2025 tax filing season.
10. **Preparer FY26 eligible (defined only at the preparer-level)**: a binary measure of whether the preparer's proportion of likely errors would qualify for a return preparer program intervention next tax year. Because the FY 2026 eligibility criteria may not be set by the time we analyze the outcome variables, we plan to use the FY 2025 eligibility criteria.

These outcomes are calculated at the client/return level using TY 2024 return data from the 2025 filing season. In addition, the secondary outcomes described below are calculated using TY 2024 return data from the 2025 filing season (endline year) and using TY 2023 return data from the 2024 filing season (baseline year).

11. **Change in erroneous EITC dollars**: this measure is calculated as the EITC dollars for EITC credits claimed erroneously during the endline year minus the EITC dollars for EITC credits claimed erroneously during the baseline year.[5] [6]

12. **Change in erroneous ACTC dollars:** this measure is calculated as the ACTC dollars for ACTC credits claimed erroneously during the endline year minus the ACTC dollars for ACTC credits claimed erroneously during the baseline year.

13. **Change in AOTC dollars:** this measure is calculated as the AOTC dollars (regardless of errors present) during the endline year minus the AOTC dollars (regardless of errors present) during the baseline year.

14. **Change in refund dollars:** this measure is calculated as the refund dollars (regardless of errors present) during the endline year minus the refund dollars (regardless of errors present) during the baseline year.

Imported variables:

N/A

[5] Our analysis with the "change variables" as outcomes will not include lagged baseline variables as covariates.
[6] Note that the data does not distinguish between portions of a refundable tax credit which are flagged with likely errors and portions which are not flagged. As such, the entirety of the refundable tax credit (e.g., EITC or ACTC) is treated as erroneous dollars.

Transformations of variables:

N/A

Transformations of data structure:

Thus far, we have referred to clients as the unit of randomization and thus also as the unit of analysis, since it is easier to think of treatment effects acting on people. However, it is more precise to refer to the unit of analysis as the tax return, as individuals may file joint returns, or may claim others as dependents on their return, so a "client" in this case may in fact refer to two or more individual people. We conducted randomization by returns, and the outcomes data will be provided to us at this level as well.

In most cases, we expect to be able to follow the same people from the 2024 filing season (TY 2023) to the 2025 filing season (TY 2024). However, in cases where clients begin to file jointly or filed separate returns while filing jointly for TY 2023, we will make the following changes:

> **Individuals who filed separately in TY 2023 but jointly in TY 2024:** It is possible that the two members of the couple were randomly assigned to different treatment groups. In this case, we will associate the jointly filed return from TY 2024 with each individual. In other words, the return will be included in the regression twice; once associated with person X and once associated with person Y.
>
> **Individuals who filed jointly in TY 2023 but separately in TY 2024:** In this case, we will focus on the primary filer, and follow outcomes only for that person.

Since we have no reason to expect differential creation or dissolution of couples across assignment groups, we do not anticipate that these changes to the data will impact our results.

Data exclusion:

For the purposes of this study, we will exclude client outliers using IRS's typical criteria. Note that our analysis is limited to TY 2024 returns filed during the 2025 tax filing season, so we exclude amended returns that are filed during the 2025 tax filing season for previous years, and we exclude TY 2024 returns that are filed (or amended) after the 2025 tax filing season.

Note that those who die or file late (i.e., after we receive outcomes data) will be treated as if they did not file, but will still be included in the analysis. Similarly, note that any amended returns will not be accounted for in our data, as our analysis will be based on the return submitted as of the end of the 2025 filing season.

Treatment of missing data:

We do not anticipate substantial missing data since our data capture the full sample of taxpayers. Not submitting a FY 2024 return in filing season 2025 is an outcome of interest, and thus missing observations are re-coded as zero. We describe our imputation method for clients who do not file returns for each outcome variable above.

Our analysis will rely on data received and processed by the end of June 2025. Until then, there may be individuals who have filed their returns, but their return has yet to be processed fully. In this case, outcomes data for some measures will be missing until their return is fully processed.

Since we measure client-level outcomes regardless of their filing method during filing season 2025, we do not anticipate any missing data once return processing has been completed.

Descriptive statistics, tables, and graphs

We plan to produce bar charts for the two primary outcomes for the following groups:

- **Figure A - Answering RQ1** - Based on the Groups 1 and 2 pool, a bar chart with four bars that shows the mean for the control group, mean for the client letter group (pooling together low-saturation and high-saturation groups), mean for the high-saturation group (regardless of client level treatment status), and mean for the low-saturation group (regardless of client-level treatment status)
- **Figure B - Answering RQ2 and RQ3** - Based on the Groups 1 and 2 pool, a bar chart with three bars that shows the mean for the control group, mean for the sent letter group (pooling together low-saturation and high-saturation groups), and mean for the spillover letter group (pooling together low-saturation and high-saturation groups)
- **Figure C - Answering RQ4** - Based on the Groups 1 and 2 pool, a **b**ar chart with five bars that shows the mean for the control group, mean for the high-saturation sent letter group, mean for the low-saturation sent letter group, mean for the spillover high-saturation group, and mean for the spillover low-saturation group
- **Figure D - Answering RQ5** - Based on the Group 5 pool, a bar chart with four bars that shows the mean for control group, mean for the preparer outreach groups (pooling together preparer call and call + letter groups), mean for the preparer call group, mean for the call + letter group
- **Figure E - Answering RQ6** - Based on both pools, a bar chart with three bars that shows the mean for the control group who have online accounts, mean for the NDC letter group for clients who have online accounts, and mean for the OLA letter group for clients who have online accounts

Statistical models and hypothesis tests

This section describes the statistical models and hypothesis tests that will make up the analysis — including any follow-ups on effects in the main statistical model and any exploratory analyses that can be anticipated prior to analysis.

Statistical models:

We rely on the following regression specifications. All analyses examine the intent-to-treat (ITT) effect of being randomized to the condition, regardless if the client or tax preparer receives outreach.

Group 1 and 2 Preparers

Research question 1 - Preparer-level assignment

In Research Question 1, we are interested in the impacts of preparer-level assignment on client outcomes, regardless of the client-level assignment (i.e., to be sent a letter or to the spillover group).

For *Hypothesis 1A (H1A)*, we rely on a regression that pools preparer-level assignment (high- or low-saturation letter group) to identify the impact of any exposure to the client letter via preparer-level random assignment. The unit of analysis is at the client level.

Specification 1:

$$Y_{ijt} = \beta_0 + \beta_1 ClientLetterGroup_{ij} + \beta_2 Y_{ijt-1} + \gamma Z'_{jt-1} + \varepsilon_{ijt}$$

where i indexes baseline client using return preparer j in tax return year t and:

- Y_{ijt} is our primary or secondary outcome of interest, as defined above;
- $ClientLetterGroup_{ij}$ is one if tax preparer j was randomized to the client letter group (either high-saturation or low-saturation treatment group); zero otherwise.
- Y_{ijt-1} is the lagged outcome measure from the 2024 filing season;
- Z_{jt-1} are the categorical variables used to generate preparer blocks based on measures from the 2024 filing season; and
- ε_{ijt} is a client level error term.

We test the null hypothesis $\beta_1 = 0$ to answer whether compared to the control group, clients assigned to a client letter group (regardless of their individual treatment assignment) have different tax compliance (*Research Question 1*, *Hypothesis 1A*).

Additionally, we are interested in decomposing these pooled effects by the two treatment arms assigned at preparer level (i.e., assignment to either low- or high-saturation group). We model the individual treatment effects using *Specification 2* that includes an indicator for assignment to the low-saturation client letter group and separate indicator for assignment to the high-saturation client letter group.

Specification 2:

$$Y_{ijt} = \beta_0 + \beta_1 LowSaturationGroup_j + \beta_2 HighSaturationGroup_j + \beta_3 Y_{ijt-1} + \gamma Z'_{jt-1} + \varepsilon_{ijt}$$

where i indexes baseline client using return preparer j in tax return year t and:

- Y_{ijt} is our primary or secondary outcome of interest, as defined above;

- $LowSaturationGroup_j$ is one if tax preparer j for client i was randomized to the low-saturation client letter group;
- $HighSaturationGroup_j$ is one if tax preparer j for client i was randomized to the high-saturation client letter group;
- Y_{ijt-1} is the lagged outcome measure from the 2024 filing season;
- Z_{jt-1} are the categorical variables used to generate preparer blocks based on measures from the 2024 filing season; and
- ε_{ijt} is a client-level error term.

We test the null hypothesis $\beta_1 = 0$ to answer whether compared to the control group, clients assigned to a low-saturation client letter group (regardless of client treatment assignment) have different tax compliance (*Research Question 1*, *Hypothesis 1C*). Similarly, we test the null hypothesis $\beta_2 = 0$ to answer the same question for the high-saturation client letter group (*Research Question 1*, *Hypothesis 1B*). We also use *Specification 2 to* test the null hypothesis $\beta_1 - \beta_2 = 0$, which tests for the equality of the effects between the treatment arms *(Research Question 4*, *Hypothesis 4A)*.

Research questions 2 and 3 - client-level assignment

Since random assignment occurs among preparers and then among clients, we are interested in measuring the treatment effects of client-level assignment to outreach (*Research Question 2* and *Research Question 3*). We are interested in measuring the impacts of being sent a letter in *Research Question 2* and in being in a spillover group in *Research Question 3*.

In *Specification 3* (as with *Specification 1*), we pool across preparer-level treatment arms (assignment to high- or low-saturation letter group) and compare these composite treatment groups to the control group.

Specification 3:

$$Y_{ijt} = \beta_0 + \beta_1 SentLetter_{ij} + \beta_2 Spillover_{ij} + \beta_3 Y_{ijt-1} + \gamma Z'_{jt-1} + \varepsilon_{ijt}$$

where i indexes baseline client using return preparer j in tax return year t and:

- Y_{ijt} is our primary or secondary outcome of interest, as defined above;
- $SentLetter_{ij}$ is one if client i was assigned to be sent letter and their tax preparer j was randomized to either client letter group (high-saturation or low-saturation client letter group) and 0 otherwise;

- $Spillover_{ij}$ is one if client i was assigned to the spillover group and their tax preparer j was randomized to either client letter group (high-saturation or low-saturation client letter group) and 0 otherwise;
- $Y_{ij,t-1}$ is the lagged outcome measure from the 2024 filing season;
- Z_{jt} are the categorical variables used to generate the blocks; and
- ε_{ijt} is a client-level error term.

We test the null hypothesis $\beta_1 = 0$ to answer *Research Question 2*, *Hypothesis 2A* and test the null hypothesis $\beta_2 = 0$ to answer *Research Question 3*, *Hypothesis 3A*.

Finally, in *Specification 4*, we decompose the pooled model across both levels of assignment to measure the effects of each combination of preparer-level and client-level treatments arms: high-saturation client letter group and sent letter, high-saturation client letter group and spillover group, low-saturation client group and sent letter, and low-saturation client letter group and spillover group.

Specification 4:

$$Y_{ijt} = \beta_0 + \beta_1 SentLetterHS_{ij} + \beta_2 SpilloverHS_{ij} + \beta_3 SentLetterLS_{ij} + \beta_4 SpilloverLS_{ij} + \beta_5 Y_{ijt-1} + \gamma Z'_{ijt-1} + \varepsilon_{ijt}$$

where i indexes baseline client using return preparer j in tax return year t and:

- Y_{ijt} is our primary or secondary outcome of interest, as defined above;
- $SentLetterHS_{ij}$ is one if tax preparer j was randomized to the high-saturation client letter treatment and client i was randomized to be sent a client outreach letter; 0 otherwise.
- $SpilloverHS_{ijt}$ is one if tax preparer j was randomized to the high-saturation client letter treatment and client i was in the high-saturation spillover group; 0 otherwise.
- $SentLetterLS_{ijt}$ is one if tax preparer j was randomized to the low-saturation client letter treatment and client i was sent a client outreach letter; 0 otherwise.
- $SpilloverLS_{ijt}$ is one if tax preparer j was randomized to the low-saturation client letter treatment and client i was in the low-saturation spillover group; 0 otherwise.
- $Y_{ij,t-1}$ is the lagged outcome measure from the 2024 filing season;
- Z_{it} are the categorical variables used to generate the blocks; and
- ε_{ijt} is an error term.

We test the null hypotheses $\beta_1 = 0$ and $\beta_3 = 0$ to answer *Hypothesis 2B* and *Hypothesis 2C* in *Research question 2 (*the effects of being sent a letter compared to the control group). We test the null hypotheses $\beta_2 = 0$ and $\beta_4 = 0$ to answer *Hypothesis 3B* and *Hypothesis 3C* in *Research Question 3* (the spillover effect of being in a letter group but not sent a letter).

Finally, we also use *Specification 4 to* test the null hypothesis $\beta_1 - \beta_3 = 0$, which tests the equality of effects of being sent a letter in the high-saturation client letter group vs. low-saturation client letter group and the null hypothesis $\beta_2 - \beta_4 = 0$, which tests the equality of effects among spillover clients in the high-saturation client letter group vs. low-saturation client letter group (*Research Question 4*, *Hypothesis 4B and Hypothesis 4C*).

Group 5 Preparers

Research question 5

In *Research Question 5*, we are interested in the impacts of the preparer call and call + letter treatments on the outcomes for clients of Group 5 preparers. We modify *Specifications 1 and 2* to this context to answer this research question.

As in *Specification 1*, we pool across preparer-level treatment arms in *Specification 5* to measure the impact of exposure to either treatment compared to the control group.

Specification 5:

$$Y_{ijt} = \beta_0 + \beta_1 PreparerOutreachGroup_{ij} + \beta_2 Y_{ijt-1} + \gamma Z'_{jt-1} + \varepsilon_{ijt}$$

where i indexes baseline client using return preparer j in tax return year t and:

- Y_{ijt} is our primary or secondary outcome of interest, as defined above;
- $PreparerOutreachGroup_{ij}$ is one if tax preparer j for client i was randomized to the preparer call or call + letter groups; zero otherwise.
- Y_{ijt-1} is a vector of lagged primary outcome measures from the 2024 filing season;
- Z_{jt-1} are the categorical variables used to generate preparer blocks based on measures from the 2024 filing season; and
- ε_{ijt} is a client level error term.

We test the null hypothesis $\beta_1 = 0$ to answer whether compared to the control group, clients who used a preparer assigned to either the preparer call or call + letter group (regardless of the client treatment assignment) have different tax compliance (*Research Question 5*, *Hypothesis 5A*).

Next, as in *Specification 2*, we decompose the pooled effect across treatment arms in *Specification 6* to measure the individual effects of assignment to the preparer call group or assignment to the call + letter group.

Specification 6:

$$Y_{ijt} = \beta_0 + \beta_1 PreparerCallGroup_{ij} + \beta_2 CallClientLetterGroup_{ij} + \beta_3 Y_{ijt-1} + \gamma Z'_{jt-1} + \varepsilon_{ijt}$$

where i indexes baseline client using return preparer j in tax return year t and:

- Y_{ijt} is our primary or secondary outcome of interest, as defined above;
- $PreparerCallGroup_{ij}$ is one if tax preparer j for client i was randomized to the preparer call group and 0 otherwise;
- $CallClientLetterGroup_{ij}$ is one if tax preparer j for client i was randomized to the call + letter group and 0 otherwise;
- Y_{ijt-1} is the lagged outcome measure from the 2023 tax year;
- Z_{jt-1} are the categorical variables used to generate preparer blocks based on measures from the 2024 filing season; and
- ε_{ijt} is a client-level error term.

We test the null hypothesis $\beta_1 = 0$ to answer whether compared to the control group, clients of preparers assigned to the preparer call group have different tax compliance (*Research Question 5*, *Hypothesis 5B*). Similarly, we test the null hypothesis $\beta_2 = 0$ to answer the same question for clients of preparers assigned to the call + letter group (regardless of their individual treatment assignment) (*Research Question 5*, *Hypothesis 5C*). We also use *Specification 6 to* test the null hypothesis $(\beta_1 - \beta_2 = 0)$, which tests for the equality of effects between the treatment arms (*Research Question 5*, *Hypothesis 5D*).

<u>Research question 6</u>

Specification 7 (client-level):

$$Y_{ijt} = \beta_0 + \beta_1 OLA_{ijt} + \beta_2 Y_{ij,t-1} + \gamma Z'_{ijt} + \alpha_j + \varepsilon_{ijt}$$

where i indexes baseline client using return preparer j in tax return year t and:

- Y_{ijt} is our primary or secondary outcome of interest, as defined above;
- OLA_{it} is one if client i was randomized to receive the client letter via OLA and is zero otherwise;

- $Y_{ij,t-1}$ is the lagged outcome measure from the 2024 filing season;
- Z_{ijt} are the categorical variables used to generate the blocks;
- α_j is tax return preparer fixed-effects; and
- ε_{ijt} is an error term.

Specification 7 is defined at the baseline client level and conditions on clients who were sent a client letter and have an online account. This specification will pool across the two evaluations and thus encompasses clients of Group 1, Group 2, and Group 5 tax preparers. We will test the null hypothesis $\beta_1 = 0$ to answer whether clients randomized to receive outreach via OLA have different tax compliance than clients randomized to receive outreach via NDC (*Research Question 6*, *Hypothesis 6A)*.

We will run all models using OLS with linearly-adjusted covariates, and we will use heteroskedastic robust standard errors (HC1) clustered by the client's baseline preparer. We use OLS for the binary outcomes for better interpretability of the treatment effect estimates.

Confirmatory analyses:

We will treat the following tests as confirmatory, also specifying the family of tests for the purpose of adjusting for multiple comparisons within a family.

Table 1. Family of tests

Outcome	Test (H_1)	Family
RQ1: Does being in a client letter group affect tax compliance?		
H1A: Control group vs. client letter group (specification 1)	$\beta_1 \neq 0$	1
H1B: Control group vs. high-saturation client letter group (specification 2)	$\beta_2 \neq 0$	1
H1C: Control group vs. low-saturation client letter group (specification 2)	$\beta_1 \neq 0$	1
RQ2: Does being sent a letter affect tax compliance?		
H2A: Control group vs. sent a letter (pooling low- and high-saturation groups) (specification 3)	$\beta_1 \neq 0$	2

H2B: Control group vs. high-saturation sent a letter (specification 4)	$\beta_1 \neq 0$	2
H2C: Control vs. low-saturation sent a letter (specification 4)	$\beta_3 \neq 0$	2
RQ3: What is the spillover effect of being in a client letter group, but not being sent a letter, on tax compliance?		
H3A: Control group vs. spillover group (pooling low- and high-saturation groups) (specification 3)	$\beta_2 \neq 0$	3
H3B: Control group vs. high-saturation spillover group (specification 4)	$\beta_2 \neq 0$	3
H3C: Control group vs. low-saturation spillover group (specification 4)	$\beta_4 \neq 0$	3
RQ4: How do the effects on tax compliance differ when most clients are sent letters versus fewer clients are sent letters?		
H4A**:** Low-saturation client letter group vs. high-saturation client letter group (specification 2)	$\beta_1 \neq \beta_2$	4
H4B: Low-saturation sent letter vs. high-saturation sent letter group (specification 4)	$\beta_1 \neq \beta_3$	4
H4C: Low-saturation spillover vs. high-saturation spillover group (specification 4)	$\beta_2 \neq \beta_4$	4
RQ5: Does tax preparer outreach affect tax compliance?		
H5A: Control group vs. preparer call or call + letter group (specification 5)	$\beta_1 \neq 0$	5
H5B: Control group vs. preparer call group (specification 6)	$\beta_1 \neq 0$	5
H5C: Control group vs. call + letter group (specification 6)	$\beta_2 \neq 0$	5
H5D: Preparer call group vs. call + letter group (specification 6)	$\beta_1 \neq \beta_2$	5
RQ6: Are there differences in tax compliance by client outreach distribution method?		

H6A: NDA vs. OLA (specification 7)	$\beta_1 \neq 0$	6

Exploratory analyses:

We will conduct exploratory analysis in three categories: 1) preparer-level outcomes; 2) alternative outcome calculations; 3) additional Group 5 hypotheses tests; and 4) heterogeneous effects.

Preparer-level outcomes:

In addition to client-level outcomes, we will explore alternative specifications where outcomes are aggregated to the preparer-level. These outcomes will be measured among preparers identified as high-risk based upon TY 2023 returns filed (IRS Group 1, Group 2, and Group 5). Within this category of outcomes, there are three sub-categories:

- **Preparer-level baseline client outcomes** will be measured by aggregating client-level TY 2024 outcomes for the preparer's eligible clients from TY 2023 (submitted a TY 2023 return with at least one likely error when claiming certain refundable tax credits), regardless of the preparer they use to file their TY 2024 return.
- **Preparer-level endline client outcomes** will be measured by aggregating client-level TY 2024 outcomes for the preparer's eligible clients (clients who submitted a TY 2023 return with at least one likely error when claiming certain refundable tax credits) from TY 2024, regardless of the preparer they used to file their TY 2023 return.
- *(if data are easily accessible)* **Preparer-level all baseline client outcomes** will be measured by aggregating client-level TY 2024 outcomes for the all preparer's clients from TY 2023 (both those eligible and ineligible to receive a letter), regardless of the preparer they use to file their TY 2024 return.

Table 5 below defines how each preparer-level outcome variable will be calculated. Note that if a tax preparer does not file any TY 2024 returns during the 2025 filing season, we will re-code their outcomes as zero (i.e., zero refund amount, zero erroneous dollars).

Table 2: Outcomes and aggregation

	Client-level	**Preparer-level**
Primary outcomes		
Refund amount	Continuous	Sum
Sum of erroneous dollars	Continuous	Sum
Secondary outcomes		

Likely tax benefit error	Binary	Proportion
Earned income tax credit (EITC) likely error	Binary	Proportion
American opportunity tax credit (AOTC) likely error	Binary	Proportion
Combined child tax credit likely error (ACTC/CTC/ODC)	Binary	Proportion
Head of household (HOH) likely error	Binary	Proportion
Change in filing method ("any method change")	Binary	Proportion
Self-file	Binary	Proportion
Did not file	Binary	Proportion
Change in filing method to different method	Binary	Proportion
Preparer FY26 eligible		Binary

To analyze preparer-level outcomes, we will conduct all of the regression specifications provided in the confirmatory analysis section but redefine the unit of analysis to be at the tax preparer-level. Specifically, we will compute the following calculations to arrive at the preparer-level outcomes. Preparer-level baseline clients' continuous outcome measures (refund amount and sum of erroneous dollars) will be defined as the following:

$$Y_j = \sum_{i=1}^{n} Y_{ji}$$

where i indexes clients (total of n_j eligible clients) of tax preparer j in tax year $t - 1$ (regardless if client i and tax preparer j form a pair in tax year t). For the proportional outcomes (i.e., tax benefit error), the preparer-level baseline client outcome will be calculated as the following:

$$Y_j = \frac{\sum_{i=1}^{n} Y_{ij}}{n}$$

For preparer-level endline client outcomes, continuous outcome measures will be defined as the following:

$$Y_j = \sum_{k=1}^{m} Y_{jk}$$

where k indexes clients (total of m_j eligible clients) of tax preparer j in tax year t. For the tax benefit error outcome measure, the preparer-level baseline client outcome will be calculated as the following:

$$Y_j = \frac{\sum_{k=1}^{m} Y_{jk}}{m}$$

Alternative outcome calculations:

We will conduct exploratory analysis on the following alternative transformations of the outcome variables of interest:

- Explore distribution of total refund amounts (defined at the client- and at the preparer-level). If most observations are strictly positive, explore the possibility of using log transformation of outcome variables.
- Transform outcome variables into ranks, akin to the method discussed in Lei (2024).[7]

Additional Group 5 hypotheses tests:

We will conduct multi-level analysis for clients of preparers in the Group 5 pool that includes the following exploratory research questions and hypotheses:

- **Exploratory research question 1:** Does being sent a client letter and being in the preparer call + letter group affect tax compliance?
  - **Exploratory hypothesis 1A:** Compared to the control group, clients sent letters and whose preparers are assigned to the call + letter group have different tax compliance.
  - **Exploratory hypothesis 1B:** Compared to the preparer call group, clients sent letters and whose preparers are assigned to the call + letter group have different tax compliance.
- **Exploratory research question 2:** Does being in the spillover group and being in the call + and letter group affect tax compliance?
  - **Exploratory hypothesis 2A:** Compared to the control group, clients in the spillover group whose preparers are assigned to the call + letter group have different tax compliance.
  - **Exploratory hypothesis 2B:** Compared to the preparer call group, clients in the spillover group whose preparers are assigned to the call + letter group have different tax compliance.

Heterogeneous effects:

We will explore how effects vary across the following sub-groups of clients:

- Clients of preparers at or above the 80th percentile for number of baseline clients (regardless of their eligibility for the letter) vs. clients of preparers at or below the 20th percentile for number of baseline clients (regardless of their eligibility for the letter)

[7] Lei, Lihua (2024). "Causal Interpretation of Regressions with Ranks." https://arxiv.org/pdf/2406.05548.

- Clients of preparers at or above the 80th percentile for the proportion of clients eligible for the client letter vs. clients of preparers at or below the 20th percentile for the proportion of clients eligible for the client letter
- Clients with a baseline refund amount one standard deviation or more above the mean refund amount vs. clients with a baseline refund amount one standard deviation or less below the mean refund amount

Inference criteria, including any adjustments for multiple comparisons:

We will apply multiple comparisons corrections within each set ("family") of hypotheses associated with a given research question (6 questions), where each hypothesis has two primary outcomes (refund amount and sum of erroneous dollars). Because some of the outcomes within a family may be highly correlated, we will run simulations to control the family-wise error rate, in line with #7 in Coppock (2015).[8] We will use a cutoff of $p = 0.05$ to determine statistical significance. All tests will be two-tailed.

Limitations:

One limitation of our evaluation relates to the sample selection for the intervention. As previously stated, the interventions were only implemented on high-risk tax preparers (IRS Group 1, Group 2, and Group 5). We therefore will not be able to generalize our findings to an out-of-sample tax preparer population (i.e., "low-risk" tax preparers with higher initial tax compliance).

Another limitation is the inability to determine tax preparer types. For example, we cannot identify whether the tax preparer is an independent tax preparer or is associated with a larger office of tax preparers. It is possible that clients coded as switching tax preparers will have still utilized the same tax preparer office but just filed using a different tax preparer. We will assess the prevalence of this issue and its potential impact on the robustness of the findings using the IRS' "network ID" tax preparer variable.

Finally, we note that one primary outcome and several secondary outcomes rely on a probabilistic model to measure likely tax return errors, which will mean our estimates of effects are less precise than they would be if we could observe errors directly, which would require information from audits.

[8] Coppock, Alex. 2015. "10 Things to Know About Multiple Comparisons." *Evidence in Government and Politics.* https://egap.org/resource/10-things-to-know-about-multiple-comparisons.

Appendix A - Statistical Power

Research questions 1-4 (Group 1 and 2 Preparers)

Tables 2 and 3 show the minimum detectable effect (MDE) sizes in terms of refund amount (in dollars) for our primary hypothesis for our evaluation that includes clients of Groups 1 and 2 preparers. This analysis uses the GUI power calculator provided by Baird et al. (2018) as a companion to their paper, which gives the power to detect several different potential effects of interest in a saturation design.[9] We supplement this with power calculations using Stata's -power twomeans- command (for RQ1).

As a benchmark for these effect sizes, we use findings from a 2021 evaluation of sending letters to clients of tax preparers. In that randomized evaluation, we found that being sent a letter reduced refund amount by $426 (s.e. = 67.45) off a control mean of $4,525. Being in the spillover group reduced refund amount by $201 (s.e.=66.87). In that study, the percentage of clients per preparer sent letters was not held fixed, but on average, 40% of a preparer's eligible clients were sent letters. As shown below, our estimated minimum detectable effects are in line with these effect sizes.

We use the following assumptions in our power analysis:

- α= 0.05
- Power = 80%
- Standard deviation of refund amount = $1,741[10]
- Intra-cluster correlation among clients who used the same preparer = 0.11[11]
- Number of clients per preparer = 10. In practice there is heterogeneity in the number of clients per preparer. The heterogeneity in the size of the client pools between different preparers will increase the MDEs above what we estimate below; however, this also gives an underestimate of the size of the overall client population, which may offset the loss in power, depending on how much of the total variance is explained by between-cluster versus individual-level variance.

Finally, note that this power analysis does not adjust for multiple comparisons that would reduce power or for the inclusion of covariates that could improve precision and increase power.

Appendix Table 1 shows the MDE effect sizes for *Research Questions 1-3*, where each treatment group is compared to the control group. The corresponding research hypothesis is in parentheses.

Appendix Table 1: Minimal Detectable Effects (MDEs) for *Research Questions 1-3*

Client-level random assignment	Preparer-level random assignment		
	Low- and high-saturation	Low-saturation	High-saturation
RQ1: Effect for client letter group	$174 (H1A)	$192 (H1B)	$192 (H1C)

[9] https://pdel.ucsd.edu/about/tools/index.html.
[10] This is the standard deviation for the average refund amount across clients in our sample during the baseline year.
[11] This is derived from the 2021 evaluation.

RQ2: Effect for sent letter group	$192 (H2A)	$209 (H2B)	$209 (H2C)
RQ3: Effect for spillover group	$192 (H3A)	$261 (H3B)	$261 (H3C)

Appendix Table 2 shows the MDE effect sizes for *Research Question 4*, where the effects for the low-saturation group are compared to the effects for the high-saturation group. The corresponding research hypothesis is in parentheses.

Appendix Table 2: Minimal Detectable Effects (MDEs) for *Research Question 4*

Client-level random assignment	Preparer-level random assignment
	Low-saturation vs. high-saturation group
Difference in effects for client letter group	$157 (H4A)
Difference in effects for sent letter group	$487 (H4B)
Difference in effects for spillover group	$313 (H4C)

Research question 5 (Group 5 Preparers)

Appendix Table 3 shows the MDEs for the evaluation that includes clients of Group 5 preparers. We use the same assumptions listed above for Groups 1 and 2 preparers. These calculations were done using Stata's -power twomeans- command.

Appendix Table 3: Minimal Detectable Effects (MDEs) for *Research Question 5*

Preparer-level random assignment	MDE
Call only + Call plus letter vs. Control	$139 (H5A)
Call only vs. Control	$157 (H5B)
Call plus letter vs. Control	$157 (H5C)
Call only vs. Call plus letter	$174 (H5D)

Research question 6 (Clients with online accounts who were sent letters across Group 1, 2, and 5 preparers)

We also calculate the MDE for the impact of the distribution method of letters to clients with online accounts on average refund amount. The sample includes clients who had online accounts and were sent letters across preparers in Groups 1, 2, and 5. We compare means for clients with online accounts sent letters via NDC versus OLA. We calculate being able to detect an effect of $135.73 or greater when clustering by preparer (an estimated 3.1 percent reduction in average refund amount). This analysis does not account for client-level covariates or preparer fixed effects.

We calculate this MDE using the following assumptions and parameters:

- α= 0.05
- Power= 80%
- Average refund amount for the NDC sent letter group = $4328
- Standard deviation in refund amount for the NDC sent letter group = $1750
- The total number of clusters, in other words the sample size for preparers who had clients sent letters = 1050.
- Sample size for clients sent letters with online accounts =9700
    - Sent via OLA (the treatment group) = 7150
    - Sent via NDC (the control group) = 2550
- Average number of observations per cluster = 9
- Intra-cluster coefficient = 0.11.