POLICY BRIEF

AUGUST 2020

CALIFORNIA POLICY LAB

2018 Tax Filing Season Honesty and Accuracy Nudge

PATRICK KENNEDY, ELIZABETH LINOS, APARNA RAMESH, AND JESSE ROTHSTEIN

EXECUTIVE SUMMARY

The California Policy Lab (CPL) and the California Franchise Tax Board (FTB) found that a simple honesty and accuracy nudge targeted at tax filers did not produce detectible changes in income reporting or taxes paid during the 2018 tax filing season. There is some evidence that the nudge produced a modest decrease in a common tax error related to filer-reported withholdings.

CPL and the FTB, in partnership with a tax preparation software provider, conducted a randomized controlled trial (RCT) to test the impact on tax filer behavior of presenting an honesty and accuracy nudge at the beginning of the tax preparation process. The nudge reminded taxpayers of their responsibility to complete the return accurately and honestly, and was intended to keep this responsibility in the forefront of their minds while completing the return.

Early evidence from experiments conducted in lab settings suggests these types of nudges can be effective. This experiment attempted to replicate these types of nudges in the tax filing context to measure their potential impacts. Our findings suggest that **this particular nudge did not have detectible effects on income reported or taxes paid.** Our sample size is large enough to be able to rule out even small effects. A possible explanation is that the nudge may not have caught users' attention, as very few clicked the accompanying "what's this?" dialogue box. However, our findings are consistent with replication studies, published after this trial was begun, that also found honesty nudges to be ineffective.

We find suggestive evidence that the nudge caused a modest decrease in tax filing errors related to filer-reported withholding. FTB flags discrepancies between taxpayer reported withholding amounts and those reported by employers to the Employment Development Department (EDD), and automatically revises the returns to reflect the value reported by EDD. Although this discrepancy is the most common tax error captured by FTB's reviews, it occurs in less than one percent of tax returns in our sample. This makes it difficult to measure the impact precisely or to be confident that the apparent effect is not due to chance.

MOTIVATION

A major component of FTB's role is to help taxpayers report their incomes and their tax obligations accurately. A sizable tax gap (taxes owed to California, but not reported) exists in California partly because some taxpayers may misrepresent self-reported elements on their return, such as their income and deductions to reduce their tax liability. Further, because some taxpayers do not report their elements such as income and deductions accurately, the FTB spends time and resources detecting and correcting errors on taxpayer returns. Small changes in reporting across millions of California taxpayers' returns can have a profound impact on California's tax revenue. Early evidence from a study conducted in a lab setting suggested that a small nudge — a signed declaration that the return is accurate, at the top of the form — could induce more honest reporting. This project attempted to replicate this nudge in the tax filing context to measure its effectiveness.

RESEARCH QUESTIONS

Does giving an honesty & accuracy nudge to taxpayers before they report their income, deduction, credit, exemption and adjustment information cause them to be:

1) **more honest** in completing their tax return, potentially increasing California's revenue?
2) **more accurate** in completing their tax return, reducing California's expense of correcting errors?

INTERVENTION

FTB sought to design a nudge modeled after a similar nudge found to be effective in a lab setting.[1] Typically, taxpayers sign a penalty of perjury statement at the end of the return filing process. FTB partnered with a commercial off-the-shelf electronic tax preparer to repeat this language at the beginning of the filing process for a randomly selected group of filers. This nudge reminded filers, prior to any prompts to report income, deduction, credit, exemption, and adjustment information, of their responsibility to complete their returns accurately and honestly, and asked them to check a box acknowledging that fact (though checking the box was not required). In contrast to the Shu et al. (2012) intervention, no signature was required, as the process was completed online. See Figure 1 for more details. Randomly chosen filers were presented with the nudge; those who did not receive the nudge completed their tax returns as usual, with the penalty of perjury statement at the end of the filing process.

METHODOLOGY

This study employed a randomized controlled trial (RCT) to test the effectiveness of the intervention. By randomly assigning filers to either receive the nudge (treatment group) or to business as usual (control group), the experiment generated two groups that were statistically similar except for one important difference — whether they received the nudge. Any difference in reporting outcomes across the two groups could then be attributed to the nudge itself. If the nudge was effective at inducing more honest reporting and claiming, we should observe that, on average, filers in the treatment group reported higher incomes and claimed fewer deductions, adjustments, and credits. If the nudge was effective at inducing more accurate behavior, we should observe fewer corrections issued by the FTB to the treatment group than the control group.

The RCT was implemented among filers who used a commercial off-the-shelf electronic tax preparation software package. The sample size was approximately 11,500. To increase the power to detect small effects, most analyses controlled for taxpayer age, filing status, the month when the return was submitted, and the value of the outcome (income, deduction, adjustment, or credit claimed) in the previous year. We conducted a series of robustness checks to ensure that results were not changed by excluding these controls, to reduce the influence of outliers, and to check for differences in impacts between new and returning users of the software.

ACKNOWLEDGEMENTS

We are very grateful to the commitment of the California Franchise Tax Board to informing policy through rigorous research. We thank Allen Prohofsky, Julie Moreno, Sean McDaniel, Monica Trefz, Chad Angaretis, and Xudong Chen. We also thank Alex Kauffman for his excellent research assistance. All errors should be attributed to the authors.

FIGURE 1: Screen with treatment

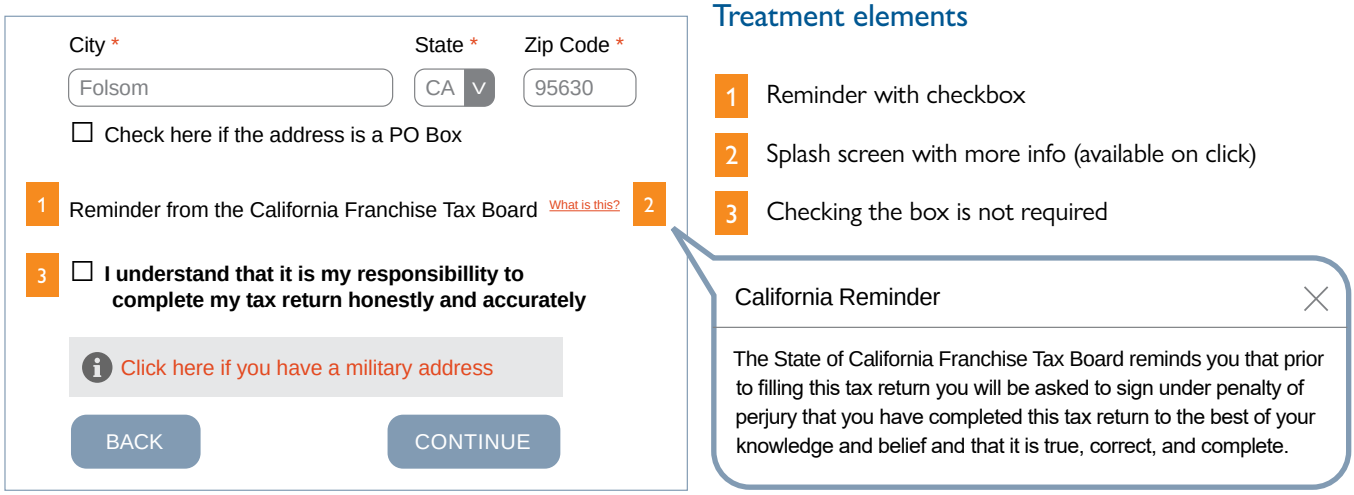
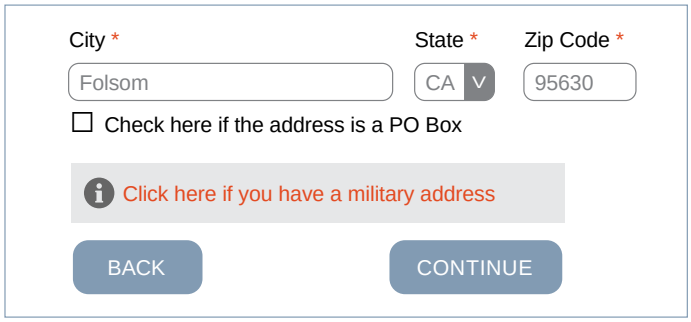


FIGURE 2: Screen without treatment



IMPLEMENTATION

The intervention was conducted by a third-party software provider. The intervention targeted returning users of the software in 2017 — that is, users who had also filed their returns using the same provider in 2016 — who listed addresses in California. Half of such users would receive the honesty and accuracy nudge, and half would not. Randomization was based on the seconds digit of the timestamp when the taxpayer submitted their name and address information; the software was configured so that users who saved their partially completed returns and continued later retained their initial timestamps. Timestamps with seconds digits between 00 and 29 were assigned to treatment and those with seconds digits between 30 and 59 were assigned to control.

Due to an error in implementation, both the population and assignment share rules were misapplied for returns initiated before February 26, 2018. These returns constituted about 20% of the eventual analysis sample. First, the software selected only new users of the software, not returning users (the selection was intended for all users). Second, only one-third of filers were assigned to treatment (based on seconds digits 00 to 19). Both errors were fixed on February 26, and this error is accounted for in our analysis.

The ultimate analysis sample consists of 11,557 tax return filers. Of these, 2,158 (18.7%) initiated a return before the errors were corrected. Overall, 47% of filers were assigned to treatment and 53% to control.

Figure 1 portrays the nudge that treated filers received. All treated users saw a message after entering their address information reminding them of their responsibility to complete their tax return honestly and accurately. Next to the main text of the nudge, filers also had an option to click the "what's this?" dialogue box to receive more information about the honesty and accuracy statement. Note that less than half of one percent of filers in the treatment group clicked the "what's this?" dialogue box — raising possible concerns about the strength of the nudge.

DATA

We employ two different datasets: 1) filing data and 2) paragraph code data. Both pertain to filers in 2016 and 2017 who used a particular commercial tax preparation package. Among users of this package, the data set includes returning California users who initiated returns before February 26, 2018, and new California users who initiated returns after that date.

Filing:

This analysis employs filing data from 2016 and 2017 tax returns. Key measures are drawn from:

- California Resident Income Tax Return (CA 540)
- California Alternative Minimum Tax and Credit Limitations (CA Schedule P)
- California Adjustments (Schedule CA)
- Federal Income Tax Return (1040)
- Federal Itemized Deductions (1040 Schedule A)

Our 2017 data includes a sample of 11,557 filers. Of these, 10,965 (95%) had filed 2016 returns.

Summary statistics for variables of interest are displayed in the first column of Table 1. The average federal adjusted gross income for filers using this software in 2017 was around $42,500. This is comparable to the overall population of California income tax return filers.

Paragraph codes:

Once tax returns are filed, the FTB performs a series of checks on submissions to detect, adjudicate, and resolve any reported inaccuracies. As part of this process, the FTB issues letters (with explanatory codes), to filers on errors it detects, revises the refund or balance due, and lists any next steps the filer should take to resolve the issue. The type of error is known within the FTB as a "paragraph code". Analyzing differences between the treatment and control filers in the number of errors caught by FTB (i.e., the number of paragraph codes issued) provides one indication of whether the nudge impacted the accuracy of reporting.

The error rate among 2017 filers was around 2–4% a year, ranging over 80 different types of errors. The total number of paragraph code observations in our sample is around 1,100. The most common error was the GC paragraph, which flags discrepancies between the withholding amounts reported on a tax return and those obtained from EDD data. As Table 1 indicates, 0.74% of taxpayers in our sample had the GC paragraph code in 2017.

Balance tests:

Using 2016 filing data, we find the sample is balanced across treatment and control. Balance tests are used to verify that randomization was conducted correctly. If so, we expect that all 2016 characteristics should have similar distributions in the treatment and control groups. Balance tests were conducted using 2016 exemptions, filing status, total tax, amount owed, capital gains, losses, earned income tax credit amounts, total taxes paid, and wages, salaries, and tips reported. Once controlled for month of tax filing, these tests (Table 2) indicate a balanced sample. While there is a statistically significant difference in the 2016 federal deduction claimed, once adjusted for outliers, this difference disappears. This holds across filers who initiated their returns before and after the randomization error in February.

ANALYSIS

To measure the effect of the nudge, we examined the differences between treatment and control in the filing data and paragraph code data for 2017. We analyzed a total of 28 different outcomes from tax returns, as well as the overall error rate from paragraph code data and several different paragraph code error rates. Table 3 lists results across 15 major outcomes of interest. These tables display results from four specifications:

- **No controls.** Simple comparison of averages between treatment and control groups, without controls.
- **Controlling for month of tax filing.** This is needed to account for differences in treatment probabilities over time.
- **Controlling for month of tax filing plus age and filing status.** The additional controls may improve precision by accounting for variation in filer demographics that may not have been perfectly balanced through randomization.
- **Controlling for month of tax filing, age, filing status, and the 2016 value of the outcome.** There is substantial heterogeneity in the population in outcomes we examine. Controlling for the prior year's outcome removes much of this heterogeneity and

allows for more precise estimates of the treatment's impact. Note that when 2016 returns were not available, we assigned taxpayers the mean value; the specification also includes a control for returns that were imputed in this way.

RESULTS

We do not find statistically significant effects of the nudge on exemptions taken, income reported, taxes owed, credits claimed, or deductions claimed. However, we do find a small reduction in withholding errors, known as paragraph code "GC".

Filing:

We examined the differences between treatment and control among 13 different outcomes related to filing data, shown in Table 3.

We do not find a statistically significant effect of the nudge on any of the elements taxpayers report on their tax forms. While in some cases estimates move in the anticipated direction — for example, recipients of the nudge filed returns showing slightly larger total tax obligations, on average, than the control group — in almost all cases we cannot rule out the hypothesis that they are due to chance alone. Moreover, for several outcomes, differences move in the opposite direction of what was originally expected. For example, the treatment group claimed more adjustments (a way to lower tax liability) than the control group (Table 3), a difference that is statistically significant. This is the opposite of our expectation, which was that the nudge would reduce adjustments. The treatment group was also more likely to take the charitable contribution deduction and less likely to claim the standard deduction — although these differences are not significantly different from zero in the most robust specification.

Paragraph codes:

Using the same specifications outlined above, we analyzed the impact of the nudge on the total number of paragraph codes received by a filer. The estimated impact was very small and not statistically significant.

We also analyzed the impact of the nudge on the most common paragraph codes. We found a reduction in the rate of the "GC" error, which is the most common paragraph code:

> *"The withholding claimed on your return has been revised to the amount that we could verify with the Employment Development Department (EDD). We also considered any W-2's and forms 1099 reporting withholding that you provided with your tax return. If you have documentation supporting the original amount of withholding claimed, please contact us."*

The honesty and accuracy nudge appears to reduce the frequency of the GC error and issuance of a notice by FTB by approximately .38 percentage points, or 44%. This result holds across all specifications.

Although this discrepancy occurs in only 0.87% of tax returns in our sample, it is the most common tax error committed by filers in our sample. However, using filing data, we do not observe a statistically significant change in reported withholding itself — which we would have expected to see if the nudge had successfully induced different patterns of reporting in withholding on the return itself.

We advise caution in interpreting these results. Given the large number of outcomes we considered, we would expect to find some statistically significant differences due to chance alone. This risk is exacerbated when we examine very rare outcomes, for which just a few observations can yield sizable percentage differences in frequency. We thus see this result as suggestive but far from definitive.

ROBUSTNESS CHECKS

Outliers:

Our baseline estimates are less precise than we anticipated, for two reasons. First, complications in the randomization procedure reduced the sample size. Second, we observed several outliers with very high incomes, deductions, and taxes paid. These outliers increased the variability of our outcome measures and reduced the precision of our estimates. Thus, the baseline analyses in Table 3 are not precise enough to allow us to rule out the possibility that the intervention had an effect of the expected direction and magnitude.

To mitigate the influence of large outliers, we censor each outcome measure at the 90th percentile. If the nudge did have an effect, the censored specification should allow us to detect it, so long as the nudge's impact was not confined to the highest-income taxpayers.

The censored specification gives similar results to the baseline, but allows for much more precision. However, even with this specification, we do not see a demonstrable impact of the nudge on taxes owed. We can rule out the possibility that the nudge increased the tax bill of the average treated

taxpayer by $15 or more. Similarly, we can rule out an effect on the effective tax rate (California taxes owed as a share of total income) larger than +0.03 percentage point.

Other approaches to reducing the influence of outliers, such as taking the square root of each outcome variable or scaling each variable as a share of total reported income, yielded results consistent with our finding from the censored specification: namely, that the nudge did not have an effect.

Heterogeneity:

We also explored differences in the impact of the nudge across different subgroups. Again, there was little sign of effects on any group. The dimensions we explored were:

- **Timing of tax filing:** Given the randomization correction, the study population differed for those who began filing their returns before and after February 26. We estimated the effects of the treatment separately for each group.
- **New v. returning filer:** Similarly, we examined whether the effect differed between those who had used the commercial tax preparer's (CTP's) service in 2016 and those who had not.
- **New tax filers:** This analysis limits the sample to those who did not file 2016 returns. (These specifications exclude the control for the 2016 value of the outcome variable.)

These specifications did not yield evidence of differential effects, either for filing or paragraph code outcomes.

In only one case was there was an indication of differential effects: among filers who began their returns after February 26, the treatment reduced the share who claimed the standard deduction by 1.8 percentage point, and is statistically significant. However, in other specifications, the estimated effect on early filers is of the opposite sign and not statistically significant. Given that we only detect an effect once among many different specifications, we believe this result is a fluke. When estimating a large number of comparisons, as we did here, some will turn out to be statistically significant due to chance alone. We discourage readers from interpreting this as a true effect.

POWER AND PRECISION

It is possible that, in reality, effects of this nudge are positive but very small and could be detected in a much larger experiment. In projecting the outcomes of this experiment, we were over-optimistic about the ability to detect effects in our baseline specification, in large part because we under-estimated the variability in outcomes in the study sample. We were able to recover the anticipated power only when we implemented ad hoc adjustments to reduce the influence of the taxpayers with the highest incomes and obligations. These adjustments were not pre-specified. An analysis with a much larger sample size could potentially detect effects concentrated among these taxpayers.

Alternatively, it is possible that the intervention has very small effects in the broader population, but that we were unable to detect them. The effects would have to be quite small, given the power of our adjusted specifications. However, because the intervention is essentially free to implement, even very small effects may justify the intervention. A much larger sample would be needed to detect effects of only a few dollars on the annual tax obligation.

CONCLUSION

The nudge did not have detectable effects on the amounts filers reported on their returns — including income, adjustments, deductions, credits, and exemptions. This may indicate that the nudge was ineffective. Consistent with this finding, a replication study of the 2012 nudge study, conducted after the FTB piloted the honesty nudge, also found that the honesty nudge was ineffective.[2] As outlined above, it may be that the effects are positive but too small for us to detect without a much larger sample. The nudge itself may simply not have been enough to capture taxpayers' attention.

There is suggestive but not definitive evidence that the nudge may have impacted errors in a subgroup of taxpayers. We find negative effects on the frequency of a common tax filing error related to filer-reported withholding. The honesty and accuracy nudge reduces the frequency of errors in reported withholding by approximately 0.38 percentage points, or 44%. However, because this outcome is so rare, we cannot be confident in this effect.

TABLE 1: Summary Statistics

VARIABLE	MEAN	SD
Exemptions (amount)	10	263
Exemptions (count)	0.31	0.75
Taxable Wages (CA)	38,923	90,417
Deduction (CA amount)	6,811	6,228
Taxable Income (CA)	34,994	34,610
Credits (CA)	18	44
Total Tax (CA)	1,135	2,318
EITC	16.87	123.99
Amount Owed (CA)	89.65	546.39
Exemptions (count, federal)	1.41	1.00
Wages (federal)	37,066	37,038
Total Income (federal)	42,906	36,991
Adjustments (federal)	347	1,344
Adjusted Gross Income (federal)	42,559	36,836
Deduction (federal)	9,561	7,094
Truncated deduction amount (federal)	8,966	4,053
Taxable Income (federal)	2,855	31,877
Medical / Dental Deduction	316	2,552
Taxes Paid Deduction	1,127	3,661
Business Income / Loss	661	6,930
Charitable Contribution Deduction	231	1,361
Standard Deduction Indicator	0.69	0.46
Medical Deduction Indicator	0.04	0.20
Business Deduction Indicator	0.10	0.30
Charity Deduction Indicator	0.09	0.29
Tax as share of wage/salary income (CA)	0.02	0.03
Tax as share of total income (CA)	0.02	0.02
CA Withholding	1,416	2,544
Total # of Tax Errors	0.0283	0.1910
Error: Withholding Misspecified (0/1)	0.0074	n.a.

N = 11,557

TABLE 2: Balance Tests

Average difference between treatment and control groups

2016 OUTCOMES	(1) NO CONTROLS		(2) CONTROLS FOR MONTH	
	DIFFERENCE	S.E.	DIFFERENCE	S.E.
No 2016 Return	0.000501	0.0041	0.00386	0.0041
Exemptions Amount	-11.33	5.106	-5.070	5.096
Exemptions Count	-0.0343	0.0148	-0.0161	0.0148
Taxable Wages	-1102.6	1492.5	-1281.8	1501.8
Deductions (CA)	20.71	314.8	-6.434	316.8
Taxable Income (CA)	-337.9	1200.01	-518.1	1207.5
Credits (CA)	-1.290	1.082	-0.520	1.086
Total Tax (CA)	-87.44	133.1	-103.0	133.9
EITC	-2.535	1.839	-1.146	1.846
Amount Owed (CA)	24.95	13.55	21.35	13.6
Exemptions (Fed)	-0.0147	0.0167	-0.00156	0.0167
Wages (Fed)	1238.7	875.0	686.2	878.6
Total Income (Fed)	1502.5	908.6	816.4	911.5
Adjustments (Fed)	14.81	24.85	8.580	25.0
AGI (Fed)	1487.6	905.5	807.8	908.3
Deduction Amount (Fed)	301.4	133.2	317.5	133.9
Truncated deduction amount (Fed)	-198.3	135.1	52.72	133.9
Taxable Income (Fed)	82.80	723.3	136.3	727.8
Medical Dental Deductions	25.61	44.8	14.77	45.1
Taxes Paid Deduction	62.02	90.36	47.23	90.9
Business Income Deduction	-72.48	96.81	-105.8	97.4
Charity Deduction	25.11	21.59	19.94	21.7
Standard Deduction (Indicator)	0.0085	0.0096	0.002	0.010
Med Deduction (Indicator)	0.00181	0.0035	0.00102	0.0035
Business Deduction (Indicator)	0.00463	0.0056	0.000966	0.006
Charity Deduction (Indicator)	0.0104	0.0055	0.00854	0.006
CA Tax wages (scaled)	-0.000161	0.0005	-0.00035	0.0005
CA Tax income (scaled)	-0.000212	0.0003	-0.00035	0.0003
CA Withholding	-111.8	127.9	-117.0	128.7
Total # of Tax Errors	-0.00236	0.004	-0.00206	0.004
Error: Withholding Misspecified	-0.00204	0.0015	-0.00167	0.0015

Note: Bold coefficients are significantly different from zero at the 5% level.

TABLE 3: Primary Results

OUTCOME	(1) MEAN AND SD	(2) BASELINE (NO CONTROLS)	(3) + MONTH CONTROLS	(4) + AGE & FILING STATUS	(5) + PRIOR YEAR OUTCOME
Total Tax (CA)	1,135.31 [2,318.5]	33.43 (43.21)	8.193 (43.43)	10.60 (43.41)	22.07 (41.04)
EITC	16.87 [124.0]	-2.341 (2.311)	-1.169 (2.323)	-1.345 (2.304)	-0.781 (2.146)
Wages (federal)	37,065.95 [37,038.0]	463.6 (690.3)	378.3 (694.3)	363.2 (673.4)	73.97 (518.8)
Total Income (federal)	42,906.16 [36,990.9]	575.3 (689.4)	379.4 (693.1)	202.6 (674.5)	-53.94 (542.0)
Adjustments (federal)	346.92 [1,344.3]	**54.84** **(25.05)**	**50.51** **(25.21)**	**50.16** **(25.19)**	**46.77** **(21.14)**
Deduction (federal)	9,561.45 [7,094.2]	257.9 (132.2)	246.6 (133.0)	168.8 (120.6)	11.10 (92.98)
Truncated deduction amount (federal)	8,965.88 [4,052.8]	120.6 (75.52)	124.1 (75.95)	57.52 (58.98)	57.89 (57.00)
Taxable Income (federal)	28,524.5 [31,877.2]	449.1 (594.1)	163.9 (597.1)	101.7 (594.1)	-2.714 (478.2)
Standard Deduction Indicator	0.69 [0.5]	**-0.0262** **(0.00858)**	**-0.0196** **(0.00861)**	-0.0140 (0.00743)	-0.0144 (0.00742)
Tax as share of wage/salary income (CA) (in %)	1.95 [3.20]	-0.04 (0.07)	-0.08 (0.07)	-0.09 (0.06)	-0.04 (0.05)
Tax as share of total income (CA)	1.55 [1.77]	-0.02 (0.03)	-0.04 (0.03)	-0.03 (0.03)	-0.02 (0.02)
Charity Deduction Indicator	0.09 [0.3]	**0.0136** **(0.00536)**	**0.0116** **(0.00539)**	0.0103 (0.00533)	0.00618 (0.00414)
CA Withholding	1416.03 [2543.70]	-6.442 (47.41)	-16.21 (47.70)	-15.93 (47.63)	-1.289 (44.07)
Total # of Tax Errors	0.0283 [0.1910]	-0.0013 (0.0035)	-0.0004 (0.0036)	-0.0005 (0.0036)	-0.0003 (0.0036)
Error: Withholding Misspecified (0/1)	0.0087 —	**-0.0045** **(0.0017)**	**-0.0039** **(0.0017)**	**-0.0039** **(0.0017)**	**-0.0038** **(0.0017)**

Note: Bold coefficients are significantly different from zero at the 5% level.

The California Policy Lab builds better lives through data-driven policy. We are a project of the University of California, with sites at the Berkeley and Los Angeles campuses.

This research publication reflects the views of the authors and not necessarily the views of our funders, our staff, our advisory board, the Regents of the University of California, or the California Franchise Tax Board.

Endnotes

1 See: **Shu, L. L., et al. (2012)**. Participants were asked to self-report results from a puzzle and fill out a reimbursement form for the study. The self-reporting document was designed to look like IRS form 1040, but participants were not told that filling out the reimbursement form was also part of the experiment. Placing a signature affirming honesty at the start of the document reduced the self-reported miles driven on the reimbursement form. After our study was implemented, a replication study was published that found that the results could not be replicated (**Kristal, et al., 2020**). We were not aware of this non-replication finding when we began our study.

2 See **Kristal, et al 2020** for more details.