



Tulane Economics Working Paper Series

On the External Validity of Laboratory Tax Compliance Experiments

James Alm
Department of Economics
Tulane University
jalme@tulane.edu

Kim M. Bloomquist
Office of Research
U.S. Internal Revenue Service
kim.bloomquist@irs.gov

Michael McKee
Department of Economics
Appalachian State University
mckeejm@appstate.edu

Working Paper 1512
March 2015

Abstract

An essential issue for laboratory experiments to inform policy debates is the “external validity” of the experimental results; that is, does behavior in the laboratory apply to behavior that occurs in the naturally occurring world? We examine this issue of external validity in the specific context of laboratory experiments on tax compliance, using two different types of evidence. We find that the behavioral patterns of subjects in the laboratory conform to that of individuals making a similar decision in naturally occurring settings. We also find that the behavioral responses of students are largely the same as non-students in identical experiments.

Keywords: experimental methods, external validity, tax compliance
JEL codes: H2, H26, C9

ON THE EXTERNAL VALIDITY OF LABORATORY TAX COMPLIANCE EXPERIMENTS

James Alm, Kim M. Bloomquist, and Michael McKee*

ABSTRACT

An essential issue for laboratory experiments to inform policy debates is the “external validity” of the experimental results; that is, does behavior in the laboratory apply to behavior that occurs in the naturally occurring world? We examine this issue of external validity in the specific context of laboratory experiments on tax compliance, using two different types of evidence. We find that the behavioral patterns of subjects in the laboratory conform to that of individuals making a similar decision in naturally occurring settings. We also find that the behavioral responses of students are largely the same as non-students in identical experiments. (JEL C9, H0, H3)

* Portions of this research were funded by the US IRS (TIRNO – 07 – P – 00683). The views expressed here are those of the authors and should not be interpreted as those of the U.S. Internal Revenue Service. Previous versions of this paper have been presented at the November 2010 National Tax Association Annual Conference in Chicago, IL, at the June 2011 Internal Revenue Service – Tax Policy Center Research Conference in Washington, D.C., and a seminar at Virginia Tech (April 2013). We are grateful to Charles Christian, John Deskins, Brian Erard, Elaine Maag, Rosemary Marcus, Alan Plumley, Joel Slemrod, Nic Tideman, three anonymous referees, and the Editor for helpful comments and discussions. Please address all correspondence to James Alm.

Alm: Department of Economics, Tulane University, 6823 St. Charles Avenue, 208 Tilton Hall, New Orleans, LA 70118 (phone 504 862 8344; fax 504 865 5869; email jalm@tulane.edu).

Bloomquist: National Headquarters, Office of Research, U.S. Internal Revenue Service, 1111 Constitution Avenue NW, Washington, D.C. 20224 (phone 202 874 0171; fax 202 874 0660; email kim.bloomquist@irs.gov).

Michael McKee: Department of Economics, Walker College of Business, Appalachian State University, Boone, NC 28608 (phone 828 262 6080; fax 828 262 6105; email mckeemj@appstate.edu).

I. INTRODUCTION

Laboratory methods are now widely accepted as a methodological approach in economics and, increasingly, they have been used to examine specific public policy issues. There is much to be gained from careful laboratory experiments. They offer a low cost means of testing (and replicating) policy innovations, and they generate precise data on individual behavior, thereby allowing estimation of behavioral responses. Importantly, they allow many policy innovations to be introduced singly and exogenously in a controlled environment, and as a result laboratory experiments are typically seen as having a high degree of “internal validity” (Campbell and Stanley, 1966; Brewer, 2000; Shadish, Cook, and Campbell, 2002) because the causal relation between variables can be properly demonstrated. However, as emphasized by Plott (1987), using laboratory experiments can allow more general inferences regarding human behavior only when the setting implemented in the laboratory parallels what is observed in the naturally occurring world. Even beyond this requirement of “parallelism” is the need for generalizing behavioral observations from the lab to the field, or “external validity”. Internal validity can be demonstrated through the evaluation of the design. However, external validity can only be verified empirically and only with respect to the specific setting being investigated. This paper investigates the external validity of laboratory experiments, and does so in the context of experiments on tax compliance behavior.

Tax evasion is central to many important policy questions. Current estimates report the “tax gap” (or difference between taxes owed and taxes paid) in the United States to be \$450 billion annually (Internal Revenue Service, 2012). Beyond these massive revenue losses, evasion creates major misallocations in resource use when individuals alter their behavior to cheat on their taxes. Its presence requires that government expend resources to detect noncompliance, to measure its magnitude, and to penalize tax evasion. Evasion alters the distribution of income in arbitrary,

unpredictable, and unfair ways, and it may contribute to feelings of unjust treatment and disrespect for the law. More broadly, it is not possible to understand the true impact of taxation without recognizing the existence and the effects of tax evasion.

Laboratory methods have been used to examine a wide range of policies that may affect the compliance decision, policies that have not always proven amenable to either theoretical analyses or empirical analyses with field data. However, laboratory studies of compliance are sometimes viewed with skepticism. The most common criticism is that the student subjects typically used in experiments may not be representative of taxpayers. Undergraduates may have little experience with filing tax returns, and their economic and demographic backgrounds may differ from that of taxpayers. Another criticism is that the context of laboratory compliance experiments does not closely enough resemble the context in which actual compliance decisions are made. As a result, there is a concern that experimental results on policy innovations that rely upon student subjects in lab compliance experiments cannot generalize to the population.

It is this issue that we examine here. Building on previous research, we present several types of evidence on the external validity of experiments on individual compliance decisions. A first question examines whether behavior by laboratory participants is replicated by behavior of individuals making a similar decision in the naturally occurring world; that is, *do participants in laboratory experiments exhibit different patterns of behavior than individuals in a similar naturally occurring setting?* To answer this question, we utilize a special data set from the U.S. Internal Revenue Service (IRS) assembled as part of its National Research Program (NRP). These data allow us to compare actual taxpayer behavior with data generated by laboratory subjects, where everyone is engaged in a similar tax reporting decision. A second question examines a different aspect of external validity: that is, *do students behave differently than non-students in identical*

laboratory experiments? We are able to answer this question with further analysis of previously reported data from laboratory experiments that compare the decisions of a population of adults with those of undergraduate students, both of whom participate in the identical laboratory experiment.

Together, we are therefore able to examine whether the “moments of the data” (e.g., the mean reporting compliance rate and its distribution) are similar when estimated in naturally occurring versus laboratory settings, and also whether the “treatment effects” of policy innovations are similar when estimated with different subjects (e.g., students and non-students).

Our analysis indicates that there is an overall similarity between the behavior of individual taxpayers in the field and of student subjects making comparable decisions in the laboratory, so that data from the laboratory closely aligns with data from the field. Our analysis also indicates that student and non-student subjects exhibit broadly similar behavior in the laboratory, even though there are some small differences in their responses to individual policy treatments. These results confirm that compliance behavior in the laboratory generalizes beyond the laboratory.

II. THE PROMISE AND THE PITFALLS OF LABORATORY EXPERIMENTS

As a science, economics is based on the development of theory and on the ability of that theory to explain observed behavior. However, unlike some other sciences, economics faces difficulties in empirically testing the predictive power of its theories using field data from the naturally occurring world. Even where field data are readily available, it is almost impossible to ensure the independence required to conduct econometric research using field data (Manski, 2000).¹ Controlled field experiments can achieve this independence, and they often use participants who are representative of the larger population of interest; however, field experiments require simplified

¹ See also the alternative perspectives of Leamer (1983), Heckman and Smith (1995), and Harrison and List (2004).

procedures, they are costly to implement, and they may raise ethical issues.² Overall, despite many significant methodological advances in recent years, there are few instances in which identification using field data, whether naturally occurring or from controlled field experiments, is uncontroversial and easily achieved.³

The use of laboratory experiments is a different response to these difficulties. Experimental methods involve the creation of a real microeconomic system in the laboratory, one that parallels the naturally occurring world that is the subject of investigation and one in which subjects (usually students) make decisions that yield individual financial payoffs whose magnitude depends on their decisions.⁴ The essence of this system is control over the environment, the institutions, the incentives, and the preferences that subjects face. Control over preferences is particularly crucial, and is achieved via the method of “induced values”. As described by Smith (1976), “[s]uch control can be achieved by using a reward structure to induce prescribed monetary value on actions.”

Tax compliance seems especially amenable to laboratory investigation. Theoretical models yield ambiguous results when asked to incorporate many of the factors deemed relevant to the individual compliance decision, and many empirical studies of tax compliance using field data are plagued by the absence of reliable information on individual compliance decisions. It is difficult to measure – and to measure accurately – something that by its very nature people want to conceal. Even when data are available and not subject to confidentiality restrictions, it is also difficult to control in econometric work for the resulting errors in variables and the many unobservable factors

² There is an extensive literature on the use of field trials in economic policy. The 1960s was a period of wide use of field trials in a variety of policy endeavors, including the provision of education services and income support programs. For many reasons, especially their costs and their potential for irreversible damages, the use of field trials has largely been abandoned (Ferber and Hirsch, 1982; Gramlich, 1997). More recently field experiments have been increasingly used to test hypotheses derived from basic theory (Harrison and List, 2004; List, 2006). In general, the intent of these studies is less to establish external validity of an experimental design but to provide a substitute for the laboratory by introducing social settings to the decision tasks.

³ For an especially provocative perspective on the difficulties of achieving identification, written for a non-technical audience, see Scheiber (2007).

⁴ For comprehensive surveys of experimental methods, see Davis and Holt (1993) and Kagel and Roth (1995).

that affect the compliance decision. Even aside from cost concerns, controlled field experiments face many of these same problems stemming largely from a loss of control over the decision setting. Laboratory methods allow many factors suggested by theory to be introduced orthogonally. Experiments also generate precise data on individual compliance decisions, which allow econometric estimation of individual responses in ways that are simply not possible with field data. Indeed, laboratory methods have been used to examine a wide range of factors in the compliance decision, factors that have not proven amenable to either theoretical or empirical analyses with field data (Alm and Jacobson, 2007).

Of course, there are some obvious limitations of laboratory experiments, especially if the intention is to use the results for informing public policy. Perhaps the most compelling critique comes from Levitt and List (2007), who caution researchers about making the “parallelism” assumption necessary in using laboratory experiments to make general statements about behavior outside the laboratory. As we have argued earlier, parallelism is an internal validity issue addressed by the design. However, the deeper essence of criticisms such as Levitt and List (2007) is the external validity of the results. This issue can only be addressed empirically. If laboratory results comport with field observations where such are available and comparable, then one has greater confidence in applying the lab results in cases where field data are not available.⁵

Of perhaps most relevance to the external validity of compliance experiments are subject pool effects. It is typically the case that laboratory subjects for various tax compliance experiments are drawn from student populations. Levitt and List (2007) suggest that student responses are unlikely to be the same as non-student responses in large part because students are younger, better educated, less representative, and less experienced in the decisions being examined than non-

⁵ The general critique of Levitt and List (2007) has itself been the subject of energetic critiques. See Falk and Heckman (2009), Camerer (2011), Armentier and Boly (2012), Kagel (forthcoming), Harrison, Lau, and Rutström (forthcoming), and Frechette (forthcoming).

students. If valid, these concerns are especially germane for tax compliance experiments where a common comment on experimental analysis of tax compliance is that “undergraduate volunteers differ from the taxpayer population in very important ways”, and so cannot “tell us something” about typical taxpayer behavior (Gravelle, 2009).

Subject pool effects can be examined by comparing the responses of student subjects with non-student subjects in (more-or-less) identical laboratory experiments. There are relatively few such studies, but the available evidence is that the experimental responses of students are often largely the same as the responses of other subject pools in similar laboratory experiments (Plott, 1987; Ball and Cech, 1996; Güth, Schmidt, and Sutter, 2007; Charness and Villeval, 2009; Güth and Kirchkamp, 2012).⁶ Plott (1987) reports comparisons of behavior of student subjects with those of corporate executives in the same policy decision setting, and he observes similar decisions among the student subjects and the executives. Dyer, Kagel, and Levin (1989) study bidding behavior in auctions using experienced traders and students as subjects, and find similar results; Shogren, Fox, Hayes, and Roosen (1999) also find comparable responses between student and non-student subjects in a study of food safety choices in retail, survey, and experimental settings.

Also of importance for the external validity of compliance experiments are context effects. “Context” relates to the complex combination of individuals’ perceptions and past experiences that influence how individuals respond in a laboratory setting designed to mimic the naturally occurring setting; that is, does the context in the laboratory decision resemble the context in the field for the same decision? The contextual setting effect can be examined by comparing student and non-student responses in laboratory experiments to the responses of participants in similarly constructed controlled field experiments, in which the same basic choice is examined in both settings.

⁶ For some contrary evidence that reports some differences between students and non-students, see Cappelen et al. (2010) and Bigoni, Camera, and Casari (2013).

Brookshire, Coursey, and Schulze (1987) compare prices obtained by buyers of strawberries in a laboratory setting versus those in a field setting. The field setting in their study mimicked the laboratory market institution, but implemented it with non-students making purchase decisions in their homes rather than in the lab. They find equivalent bidding behavior in both settings. More recently there are investigations of behavior of fishermen (Carpenter and Seki, 2011) and of water markets (Chermak et al., 2013). A range of other studies is summarized by Camerer (2011), in which student responses in laboratory experiments are compared to responses of participants in controlled field experiments in such areas as sports card trading, open-air flea markets, donations to student funds, soccer, communal fishing ponds, proofreading/exam grading, and restaurant spending. In most – although not in all – cases, these comparisons have shown no significant differences in behavior.

This previous literature has considered subject pool and context effects, but rarely have both been examined in the pursuit of a common decision setting and, to our knowledge, there are no studies that have looked at tax compliance.⁷ Our focus is on tax reporting behavior, and we ask whether behavior observed in the lab is likely to be similar to behavior observed in the naturally occurring environment. Specifically, *do participants in laboratory experiments behave differently than individuals in a similar but naturally occurring setting?* Further, *do student subjects in tax compliance laboratory experiments behave differently than non-student subjects in identical laboratory experiments?* The next sections present our results.

⁷ Note that tax compliance has been the subject of several controlled field experiments. In a typical experiment, a treatment group of individuals is randomly selected to receive a letter from the tax authority suggesting that they will be under special scrutiny, while a control group of individuals does not receive the letter. Comparison of the treated group with the control group then gives a measure of the effectiveness of increased enforcement. One of the first of these field experiments was by Slemrod, Blumenthal, and Christian (2001); more recent examples include Iyer, Reckers, and Sanders (2010), Kleven et al. (2011), Fellner, Sausgruber, and Traxler (2011), Pomeranz (2013), and Castro and Scartascini (2013), and Bott et al. (2013).

III. TEST (1): EXPERIMENTAL RESULTS VERSUS NON-EXPERIMENTAL RESULTS

A first type of evidence compares experimental behavior with behavior in similar but naturally occurring settings (i.e., the field), in order to determine whether patterns of behavior in the field match patterns in the laboratory. For this evaluation of context effects, we compare the behavior of student subjects in experiments with that of individuals making similar decisions in the field who are subjected to random taxpayer audits conducted under the National Research Program (NRP) of the Internal Revenue Service (IRS). We discuss the data, and we then present comparisons of tax reporting compliance by students in laboratory experiments versus actual taxpayers.

A. Data: Taxpayer Sample versus Experimental Sample

The comparisons here involve two separate data sets: taxpayer (field) data and experimental (laboratory) data. The “Taxpayer Sample” is a sub-sample of NRP data for tax year (TY) 2001 (Bennett, 2005). In that year the NRP audited tax returns of 44,768 taxpayers selected using stratified random sampling, which can be weighted to represent the population of 125.8 million taxpayers who filed timely tax returns. Our sub-sample consists of taxpayers whose sole source of income (pre- and post-audit) is from a Schedule C sole proprietorship. Filers with Schedule C income were selected since this source of income has no third-party information reporting (e.g., Form W-2 for wage income), and this mimics our laboratory setting in which there is no matched information on earnings. This reduced the NRP data to 1,673 NRP audit cases weighted to represent 1.1 million taxpayers. Our sub-sample was further narrowed to make the tax reporting task as similar as possible to the situation faced by laboratory subjects. Only those Schedule C filers having positive taxable income as determined by the examiner were selected. Again, excluding taxpayers with zero taxable income was done to ensure that taxpayers selected for comparison share circumstances similar to those faced by experimental subjects who decide to report none, some, or

all of a positive amount of income. The resulting sample of taxpayers contains 1,101 NRP audit cases weighted to represent the tax returns of 559,555 individuals. Finally, within this data set there were 29 cases where reported taxable income exceeded the amount of taxable income following examination. These cases (representing 13,131 taxpayers) were assumed to have 100 percent reporting compliance.⁸

Table 1 displays summary statistics for the Taxpayer Sample. The figures in the two rightmost columns refer to the mean of the individual reporting compliance rates and the overall mean reporting compliance rate, defined as the amount of taxable income reported divided by the amount of taxable income per exam.⁹ The range of taxable income per exam for this sample spans five orders of magnitude from less than \$40 to more than \$4 million. The probability of audit for individual taxpayers as a whole in calendar year 2002 was 0.57 percent and 1.72 percent for all Schedule C filers (IRS, 2002).

The experimental data (“Experimental Sample”) were collected from college-age subjects using a basic experimental design similar to the design discussed in more detail later (Alm and McKee, 2006; McKee, Alm, Cherry, and Jones, 2008; Alm, Jackson, and McKee, 2009; Alm, Deskins, and McKee, 2009; Alm, Cherry, Jones, and McKee, 2010, 2012).¹⁰ Participants earned income, chose whether to file a tax return, and (conditional upon filing) self-reported tax liability to the tax authority at an announced tax rate. Audits occurred with an announced probability, and any underreporting was discovered by the audit. If the participant had not paid the appropriate tax, then

⁸ For example, if the taxpayer reported \$110 in Schedule C net profits and the NRP examiner determined the correct amount should have been \$100, then the calculated reporting compliance ratio of 1.1 was recoded to 1.0. Recoding these 29 cases ensures that all 1,101 observations of the reporting compliance ratio fall in the range between 0 and 1, inclusive.

⁹ The term “income per exam” refers to the income that should have been reported based on the judgment of NRP examiners, and reflects population weights, as appropriate.

¹⁰ Note that this Experimental Sample overlaps partially with the experiment data in Alm, Cherry, Jones, and McKee (2010, 2012), which we use later and in which both students and non-students were subjects. For our analysis in this section, we used only student responses from Alm, Cherry, Jones, and McKee (2010, 2012). All of our other experimental studies included only student participants.

both unpaid taxes and penalties were collected. This process was repeated over multiple rounds, and subjects were paid their after-tax earnings at the end of the experiment.

The “Full Sample” of these experimental data consists of 16,560 observations from 1,072 individual subjects, and contains observations for base case (or no treatment) scenarios and several treatment scenarios, including the existence of a public good, unofficial communication among participants and official communication from the tax authority. In our comparisons, we use a “Selected Sample”, or data from only the base case scenarios.¹¹ In these base case sessions, participants were informed of the number of audits performed (including zero if no audits were performed) following each round. This is similar to IRS policy that makes publicly available the number of audits performed each year. The Selected Sample subset has 3,780 observations from 252 individuals. Descriptive statistics for both samples are shown in Table 2, for the five different audit rates in the experiments.

B. Mean Reporting Compliance Rates

A comparison of Tables 1 and 2 shows that mean reporting compliance rates (computed as the average of individual compliance rates) for the lowest two audit rate categories in the Selected Sample of the Experimental Sample are comparable to the unweighted mean compliance rate for individuals in the Taxpayer Sample.¹² The mean reporting compliance rate in the Experimental Sample is 0.286 when the audit rate is zero and 0.368 when the audit rate is 0.05. Assuming the TY 2001 audit rate of 1.72 percent for Schedule C filers, we interpolate a reporting compliance rate of

¹¹ We use the base case simulations since these observations exclude behavioral influences potentially induced by the specific treatments explored in the non-base case simulations, influences that we believe are likely absent in actual taxpayer behavior. However, as we report below, our basic findings hold using either the Full Sample or the Selected Sample of experimental data.

¹² An individual’s compliance rate is computed for each subject after each round in the Experimental Sample and for each individual in the NRP-based Taxpayer Sample. We believe using the unweighted mean for comparison is appropriate here since it is not possible to construct weights for the Experimental Sample that would equate this group with the NRP stratified sample weights.

0.314 for the Experimental Sample for this audit rate. This rate is essentially identical to the unweighted mean reporting compliance rate for individuals of 0.313 for the Taxpayer Sample.¹³

We also conducted a simple test to determine if the observed difference is or is not statistically significant. For this test we constructed four additional taxpayer samples using NRP data for TY 2006 to TY 2009 using the same criteria applied to the TY 2001 NRP data. Using these data we calculated mean unweighted reporting compliance rates of 0.341, 0.321, 0.324, and 0.327, respectively, for these additional tax years, all with a largely unchanged audit rate. The average mean reporting compliance for all five observations is 0.325, and the 95 percent confidence interval is 0.013 based on a standard deviation of 0.010 using these five observations. This implies that the population mean is between 0.312 and 0.338, which encompasses our interpolated value of 0.314 for the Experimental Sample.¹⁴ This finding could be further strengthened by having more experimental observations on individuals' reporting behavior for audit probabilities that more closely reflect conditions in the naturally occurring world (i.e., audit probabilities between 0.00 and 0.05), as well as by additional years of NRP data.

C. Distribution of Reporting Compliance Rates

Another way to externally validate the experimental data is to compare the distribution of subjects' reporting compliance rates to those of actual taxpayers. Figure 1 displays the distribution of reporting compliance rates for the Taxpayer Sample (unweighted and weighted), and Figure 2 shows the distribution of individual reporting compliance rates for the Experimental Sample for

¹³ The virtually identical values of 0.314 for the base case Experimental Sample and 0.313 for the TY 2001 Taxpayer Sample are apparently coincidental, although we find that the point estimate for the Experimental Sample falls within the 95 percent confidence interval using five independent Taxpayer Sample observations. See the discussion in the text.

¹⁴ A comparison using the full Experimental Sample gives similar results. Using the data for the full Experimental Sample found in Table 2, the interpolated compliance rate is 0.331. This value also falls within the 95 percent confidence interval calculated using the five years of NRP data.

different audit rates. (We omit in Figure 2 the observations from the Experimental Sample where the audit probability is 0.40 for brevity.)

Visual inspection of these plots reveals that both the Taxpayer Sample and the Experimental Sample have a bi-modal distribution and an apparently random distribution of observations between these two modes. It is also evident from these plots that both samples exhibit a small and similarly sized group of individuals who exhibit 100 percent compliance even though the rational choice (from a purely economic standpoint) is to underreport income. Once again, laboratory experiments can reliably replicate the behavior in the naturally occurring world.¹⁵

IV. TEST (2): STUDENTS VERSUS NON-STUDENTS IN IDENTICAL EXPERIMENTS

The comparison of experimental behavior with behavior in similar but naturally occurring settings (i.e., the field) addresses one aspect of external validity (context effects). A second type of evidence compares student and non-student subjects in the same experimental setting in order to address subject pool effects.

These comparisons are based on further analysis of data derived from laboratory experiments conducted by Alm, Cherry, Jones, and McKee (2010, 2012).¹⁶ In both studies the subject pool consisted of students and non-students, but the focus in those papers was on the policy instrument performance rather than subject pool effects. Here we utilize these data to address the issue of subject pool effects by testing whether behavior is statistically different across the student versus non-student pools. By using two studies involving different subject pools and run at different

¹⁵ An additional comparison between behavior in laboratory experiments and field data would compare the behavioral elasticities estimated with laboratory data to the elasticities estimated with field data. The field data that we have here does not allow us to make these estimations and comparisons. However, Alm, Jackson, and McKee (1992) compared behavioral responses to audit and tax rates estimated with laboratory data to responses estimated with field data, and found very similar elasticities.

¹⁶ These studies are hereafter referred to as ACJM 2010 and ACJM 2012, respectively. Since no subject participated in more than one session, we have a total of 347 subjects for our analysis here, as discussed later.

times, we broaden the base for analyzing the effects of alternate subject pools. We first discuss the experimental designs, and we then present the comparison of student versus non-student responses.

A. The Experimental Designs

The basic experimental setting was common to both papers, and implemented the fundamental elements of the voluntary filing and reporting system of the individual income tax in most countries. The setting was “context rich”, in that tax language was used throughout. Participants earned income by performing a task, chose whether to file a tax return, and (conditional upon filing) self-reported tax liability to the tax authority at an announced tax rate. At the time of filing and reporting decisions, only the individual knew his or her true (or expected) level of tax liability, and could choose to file and then to report any amount from zero on up. Audits occurred with an announced probability, any underreporting was discovered by the audit, and the participant was required to pay unpaid taxes and penalties if he or she had not paid the appropriate taxes. This process was repeated over a number of rounds each representing a “tax year”. Participants were informed that they would be paid their after-tax earnings at the end of the experiment, converted from lab dollars to U.S. dollars at a fixed and announced conversion rate. Sessions lasted 20 rounds; this was not announced to the subjects.

Participants were told, with certainty, of the audit probability, the penalty rate, and the tax rate. The tax rate was set at 35 percent for all sessions; the penalty rate was also fixed for all sessions at 150 percent (i.e., unpaid taxes plus a penalty of 50 percent of unpaid taxes if audited). The audit probability for filed tax returns was varied once within the session. Participants were also told that there was a zero probability of audit if no tax form was filed.¹⁷ There was no public good

¹⁷ The probability of audit if the individual does not file was set at zero to reflect the fact in most countries that an individual who does not file faces no effective chance of detection. The actual audit probability for non-filers in the field may not strictly be zero. However, there is substantial evidence that this non-filing audit rate is effectively very close to zero. For example, in the U.S. the IRS conducts audits of non-filers based on tips, on “lifestyle audits” in which

financed by the tax payments in order to focus subject attention entirely on the filing and tax reporting tasks rather than fiscal exchange.

Into this setting, various policy innovations were introduced. A first set of experiments (ACJM 2010) investigated the effects of *information services* on compliance decisions. Here the basic tax reporting decision was “complicated” in different treatments through the introduction of uncertainty regarding the true tax liability, and then information services were provided by the “tax administration” that partially or fully resolved the uncertainty, thereby allowing subjects to compute more easily their tax liabilities. Also contributing to complicating the decisions were a tax deduction (comparable to an itemized deduction) and a tax credit (comparable to the Earned Income Tax Credit), each of which was conditional upon filing. The tax deduction was set at 15 percent of income, and the tax credit began at a given level and declined at a stated rate as income increases. As a treatment, the exact levels of the deduction and credit were uncertain to the taxpayer at the time of filing. Uncertainty was implemented via mean-preserving spreads (with a uniform distribution) in each, where the participants were informed of the means and the ranges of the allowed credit and deduction. As an additional treatment, information services were provided that resolved the uncertainty. The information was complete, accurate, and costless to the participant.

A more direct set of *positive inducements* was also investigated in a second set of experiments (ACJM 2012). In one treatment income tax credits were introduced that were available to participants but only to those who filed a tax return. In a second treatment a “social safety net” (e.g., unemployment replacement income) was present in which individuals faced some probability of unemployment but replacement income could be provided, with the benefits conditional upon

visible expenditures are a flag for an audit, or through passive income sources such as deposit interest. However, the frequency of non-filing audit is very low even in the U.S., and in many countries it is essentially zero (Erard and Ho 2001). Accordingly, we elected to implement for simplicity a zero audit probability in the lab setting. Note that this framework required only that the probability of audit for non-filers be less than the probability of audit for filers.

past filing behavior. There was a known probability of unemployment, and, if the individual became unemployed and earned no income, then he or she was unemployed for two periods. Unemployment replacement income was received only if the individual had filed a tax return in each of the two previous periods, the level of which was based on reported income.

These various treatments are summarized in Table 3A and 3B, with Table 3A showing the information services design of ACJM 2010 and Table 3B showing the positive inducements design of ACJM 2012.¹⁸ In Table 3A, treatment T1 provides a baseline setting that entails no uncertainty and no tax authority information. The second treatment (T2) introduces tax liability uncertainty, in which participants face uncertainty regarding their allowed deduction and tax credit. The third treatment (T3) entails the same uncertainty as in the second treatment, but introduces the option of resolving the uncertainty by receiving information from the tax authority; that is, participants in this treatment were able to click on a button to reveal the true levels of the deduction and the tax credit. In Table 3B, treatment T4 establishes a baseline with no positive inducements, a tax credit is introduced in T5, and an unemployment benefit is introduced in T6. The parameters used for the different treatments are reported in Table 4. The Appendix shows a representative screen.¹⁹

As noted above, the experimental interface and instructions made intensive use of tax language. Participants also decided whether or not to file a tax return. They disclosed tax liability in the same manner as on the typical tax form (e.g., entering income, deductions, and credits on a tax form). There was a time limit on the filing of income, comparable to a filing deadline, and the individual was automatically audited if he or she failed to file on time. A timer was shown on the screen; when 15 seconds remained, the timer changed color to red, and the clock began to flash as a

¹⁸ The main intent of ACJM 2010 and ACJM 2012 was to investigate policies to induce filing when non-filing is possible. Both policies were found to be effective.

¹⁹ T1 and T4 were separate “baselines” for each respective series of experiments, and did not present the same environment.

reminder that the filing period was about to end.

The dedicated laboratory consisted of 25 networked computers, a server, and software designed for these experiments. Sessions were conducted at a major state university, using both students and staff as participants.²⁰ Recruiting was conducted using the Online Recruiting System for Experimental Economics (ORSEE) developed by Greiner (2004). The participant database was built using announcements sent via email to all students and staff. Participants were invited to a session via email, and were permitted to participate in only one tax experiment. The experiments followed procedures that implemented a single- and double-blind setting (e.g., no subject communication, use of computer screens to convey information, no individual identification, complete privacy in subject payment). Methods adhered to all guidelines concerning the ethical treatment of human subjects.

Of most importance for the purposes of this analysis, participants included both students and non-students, thereby allowing one aspect of the external validity of experiments to be examined: *do students behave differently than non-students in identical experiments?* A given session consisted of *either* student *or* non-student participants, not both. The experimental design was identical for students and non-students, with only the compensation varied for students and non-students by means of the exchange rate. Sessions lasted approximately one hour. For student participants the conversion rate was 80 lab dollars to 1 U.S. dollar while staff participants received a higher exchange rate to reflect their higher outside earnings, with a conversion rate of 50 lab dollars to 1 U.S. dollar. Earnings averaged \$18 for student subjects. The average payoff for staff was \$28.

B. Laboratory Experimental Results

²⁰ The student portion of the subject pool covered a very broad range of year in studies and major, and no single major exceeded 8 percent of the pool. The staff pool was similarly diverse, covering all levels of support staff and non-academic professional staff.

Three hundred forty seven individuals participated in a session in one of the two series of experiments. We present the distribution of subjects and some basic demographic data by treatment in Table 5. In the sessions designed to investigate the role of tax information services (T1 through T3), there were 131 subjects, 54 percent of whom were students. In the sessions designed to investigate the effects of positive inducements (T4 through T6), there were 216 subjects with 55 percent students. Note that, while the design of the experiments was balanced in terms of treatments, it was not completely balanced in terms of equal numbers of students and non-students for each treatment. Since the design was not strictly balanced in terms of numbers of students/non-students, a simple comparison by treatment of results for students versus non-students (e.g., average compliance rates) may be misleading. Instead, we focus on our econometric results since this method includes control variables that address subject characteristics.²¹

In order to control for various factors, for each series we estimate the conditional effects of design parameters on reporting behavior, while holding other factors constant. We estimate these responses separately for the two experimental designs, using the basic specifications of:

Information Services

$$Y_{i,t} = \beta_0 + \beta_1 \text{Income}_{i,t} + \beta_2 \text{Wealth}_{i,t} + \beta_3 \text{AuditProbability}_{i,t} + \beta_4 \text{TaxLiabilityUncertainty}_i + \beta_5 \text{TaxAgencyInformation}_t + \beta_6 X_i + \psi_t + u_i + \varepsilon_{i,t},$$

Positive Inducements

$$Y_{i,t} = \beta_0 + \beta_1 \text{Income}_{i,t} + \beta_2 \text{Wealth}_{i,t} + \beta_3 \text{AuditProbability}_{i,t} + \beta_4 \text{TaxCredit}_t + \beta_5 \text{UnemploymentBenefit}_t + \beta_6 X_i + \psi_t + u_i + \varepsilon_{i,t},$$

where the dependent variable $Y_{i,t}$ denotes subject i 's decision to report income in period t ; $\text{Income}_{i,t}$ is subject i 's earned income in period t ; $\text{Wealth}_{i,t}$ is subject i 's accumulated earnings (or *Wealth*) in period t ; $\text{AuditProbability}_{i,t}$ is the audit rate for subject i in period t ; $\text{TaxLiabilityUncertainty}_i$ is an indicator variable that signifies the presence of uncertainty about tax features in period t ;

²¹ In fact, we found some differences in the *levels* of compliance between the two subject pools across treatments. However, the *changes* in compliance in response to the treatment effects were quite similar for both subject pools, and it is this result that we emphasize in our discussion.

$TaxAgencyInformation_t$ is an indicator variable that signifies the presence of agency-provided information in period t ; $TaxCredit_i$ is an indicator variable that signifies the presence of a tax credit that the subject can claim on filing a tax report; $UnemploymentBenefit_t$ is an indicator variable that signifies the presence of a safety net that (partially) makes up for income lost due to unemployment; X_i denotes dummy variable denoting whether the participant is a student (coded 1) or staff (coded 0) as well as possible interaction terms,²² ψ_t is a set of $T-1$ dummies that capture potential non-linear period effects (T denotes the number of time periods); u_i are random effects that control for unobservable individual characteristics; $\varepsilon_{i,t}$ is the contemporaneous additive error term; and β_k is the coefficient for variable k . We also include a dummy variable for whether the individual is audited in the previous period ($LagAudit_{it}$). We report results for a (subject) random effects generalized least squares estimation of the panel data with standard errors corrected for clustering at the subject level. The dependent variable is defined as the reporting compliance rate of individual i in period t , where $Y_{i,t}$ equals reported tax paid divided by true tax owed of individual i in period t .²³

We first estimate the models for the pooled sample (students and non-students), reported as Model 1 in Table 6A for the information services experiments (T1, T2, and T3) and in Table 6B for the positive inducements experiments (T4, T5, and T6). The various coefficient estimates are consistent with expectations.

In Model 2 we introduce a dummy variable (*Student*) denoting whether the subject is in the student or the non-student pool, equal to 1 for students and 0 for non-students. The coefficient on *Student* is not statistically different from zero for both series of experiments. Also, the remaining

²² In ACJM 2010 and ACJM 2012, a vector of demographic variables (e.g., gender, subject age, subject own preparation of tax returns, subject claimed as a dependent on parental tax returns) was included. However, these variables are highly correlated with participant membership in the subject pool, and so are not included separately in the current analysis.

²³ Note that *Income* and *Wealth* are exogenous variables, which justifies their inclusion as explanatory variables. *Income* is earned each period prior to the tax reporting decision, and performance on the task (sorting nine numbers into the correct order) is uncorrelated with the tax reporting decision. Similarly, *Wealth* is accumulated over time, and, between the income earning task and the random nature of the audits, this variable is not correlated with past decisions.

coefficients are virtually unchanged across Model 1 and Model 2 for the different subject pools, a result that demonstrates that the pooled analysis (e.g., students/non-students) in ACJM 2010 and AJCM 2012 is appropriate.

A more critical test of subject pool effects involves testing for differences in behavioral responses of students versus non-students to the policy initiatives. To do this, we interact the *Student* dummy variable with the policy treatment variables associated with each session. These results are reported in Model 3 in Tables 6A and 6B. For the information services setting (Table 6A), tax liability uncertainty actually increases compliance for the student pool (Model 3) while the overall effect of uncertainty is negative (Models 1 and 2), perhaps because the relative lack of experience in filing and reporting by the student pool leads them to overestimate the costs of reporting errors. However, the coefficient on the interaction effect when the information service is offered is not different from zero; that is, we cannot reject the null hypothesis that students and non-students respond in the same way. Also, the coefficient for the information services variable in Model 3 is largely the same as in Model 1 or Model 2.

When positive inducements are examined (Table 6B), the coefficients on the interaction terms are never significant, and the coefficients on the treatment variables themselves are not different than when these interaction terms are or are not included. Here students respond to the presence of the safety net (unemployment benefits) exactly as do non-students. When the refundable credit is interacted with student status, the coefficient is again not significant, also indicating that the subject pools respond in the same manner.²⁴

Another policy variable modified during these experiments is the change in the probability of being selected for an audit. Only two audit rates were implemented in these experiments, and it is

²⁴ The purposes of ACJM 2010 and ACJM 2012 were narrowly defined to study possible policy actions of the tax agency in the areas of information services and positive inducements, and so they did not fully explore the effects of audit policies.

not overly surprising that the coefficient on *Audit Probability* is never significant. In addition, we interact *Audit Probability* and *Student* in Model 4 in Tables 6A and 6B. The coefficient on this interaction term is not significant either for the information services series (Table 6A) or for the positive inducement series (Table 6B).

In sum, the coefficient on the subject-type dummy variable by itself is never statistically significant for either set of experiments, and the coefficients on the student-treatment interaction terms are in most all cases also insignificant, with the only exception arising over uncertain tax liabilities. Overall, students behave largely the same as non-students in identical experiments in their reporting decisions, especially in their *changes* in compliance behavior in response to the policy variables (if not necessarily in the *levels* of their compliance behavior).²⁵

V. CONCLUSIONS

Our analysis suggests two main conclusions regarding the external validity of tax compliance experiments. Both conclusions are consistent with the result that students and non-students behave largely the same. Even so, both conclusions also suggest areas where care must be taken in transferring the results of laboratory experiments to field settings.

First, experimental data versus non-experimental (NRP) data indicate very similar patterns. The comparison of the Taxpayer and Experimental Samples finds that the experimental data can reliably replicate known features of taxpayer compliance behavior for similar decisions in the naturally occurring world, including a bi-modal distribution of reporting compliance rates, a largely random distribution of individuals between the extremes, and the existence of a small group of “pathologically honest” individuals who report 100 percent of income. The Taxpayer and Experimental Samples also appear quite similar with respect to a point estimate of the levels of

²⁵ Note that we find similar results for filing behavior.

reporting compliance, with the caveats noted above. The interpolated reporting compliance rate for the Experimental Sample is indistinguishable from the mean reporting compliance rate for individuals for the Taxpayer Sample, and is within the 95 percent confidence interval based on five independent NRP sample data sets for TY 2001 and TY 2006 to TY 2009.

Second, the experimental responses of students are largely similar to the experimental responses of non-student subject pools when faced with policy treatments. When student status is interacted with the policy changes being implemented, the resulting coefficients are not generally significant. However, there is at least one exception to this result, and this gives rise to a caveat: care must be taken when the policy treatment may incorporate a substantial level of external experience. We find that students respond differently to the presence of tax liability uncertainty, and our conjecture is that this may be the result of non-student subjects having more experience with this specific phenomenon in the field. Regardless, however, we still find that the *changes* in compliance behavior in response to institutional changes (treatments) of these pools (if not always their *levels*) largely parallel each other.

In sum, our results are consistent with studies showing that laboratory behaviors largely parallel real-world behaviors in settings that compare similar types of decisions in similar types of settings. Our results are also consistent with studies that demonstrate that student and non-student subjects behave and especially respond similarly. Concerns with the external validity of experimental results, at least in the context of tax compliance and in the comparison of changes in behavior, seem largely unwarranted.

Even so, we recognize that one must use the results from laboratory experiments with some care. However, such use depends largely upon the purpose of the experiment. According to Roth (1987), experiments can be classified into three broad categories that depend upon the dialogue in

which they are meant to participate. “Speaking to Theorists” includes those experiments designed to test well-articulated theories. “Searching for Facts” involves experiments that examine the effects of variables about which existing theory has little to say. “Whispering in the Ears of Princes” identifies those experiments motivated by policy issues. To date, most experiments in behavioral public economics have fallen into the first two categories. However, this is now changing, and experiments are being increasingly used to illuminate policy debates, especially in the area of tax compliance.

In sum, we believe that the reported results demonstrate that laboratory experiments in the area of tax compliance behavior meet the key conditions for external validity. This is an important result, especially since empirical analyses of compliance behavior with naturally occurring field data is limited and field experiments of compliance are costly to implement. We do not argue that laboratory experiments can be used to calibrate field results (e.g., provide point estimates). The stakes are obviously smaller in the lab, and the decision settings are necessarily less rich. Thus, the magnitudes of the responses to the external stimuli will be different in the two environments. However, as Kessler and Vesterlund (2014) argue, “... for most laboratory studies it is only relevant to ask whether the qualitative results are externally valid” (e.g., the direction of response), and not whether an exact quantitative result (e.g., the magnitude of response) is found in laboratory versus field data. They contend that “...there is much less (and possibly no) disagreement on the extent to which the qualitative results of a laboratory study are externally valid”. Indeed, our results in this paper are largely consistent with their position: we have shown that the behavioral patterns are sufficiently similar that we can safely predict the effects that would arise in the field from a policy based on the results observed in the laboratory.

We find the result of our investigations both comforting and plausible. We believe that these results suggest that the burden should now be on skeptics to prove that results from laboratory compliance experiments differ in meaningful ways from the behavior we observe in the field.

REFERENCES

- Alm, James, Todd Cherry, Michael Jones, Michael McKee. "Taxpayer Information Assistance Services and Tax Reporting Behavior." *Journal of Economic Psychology*, 31 (4), 2010, 577-586.
- Alm, James, Todd Cherry, Michael Jones, Michael McKee. "Social Programs as Positive Inducements for Tax Participation." *Journal of Economic Behavior & Organization*, 84 (1), 2012, 85-96.
- Alm, James, John Deskins, and Michael McKee. "Do Individuals Comply on Income Not Reported by Their Employer?" *Public Finance Review*, 37 (2), 2009, 120-141.
- Alm, James, Betty R. Jackson, and Michael McKee. "Estimating the Determinants of Taxpayer Compliance with Experimental Data." *National Tax Journal*, 45 (1), 1992, 107-114.
- Alm, James, Betty R. Jackson, and Michael McKee. "Getting the Word Out: Increased Enforcement, Audit Information Dissemination, and Compliance Behavior." *Journal of Public Economics*, 93 (3-4), 2009, 392-402.
- Alm, James, and Sarah Jacobson. "Using Laboratory Experiments in Public Economics." *National Tax Journal*, 60 (1), 2007, 129-152.
- Alm, James, and Michael McKee. "Tax Compliance as a Coordination Game." *Journal of Economic Behavior & Organization*, 54 (3), 2004, 297-312.
- Alm, James, and Michael McKee. "Audit Certainty, Audit Productivity, and Taxpayer Compliance." *National Tax Journal*, 59 (4), 2006, 801-816.
- Armentier, Olivier, and Amadou Boly. "On the External Validity of Experiments in Corruption," in *New Advances in Experimental Research on Corruption (Research in Experimental Economics, Volume 15)*, edited by Danila Serra and Leonard Wantchekon. Bingley, UK: Emerald Group Publishing Limited, 2012, 117-144
- Ball, Sheryl B., and Paula-Ann Cech. "Subject Pool Choice and Treatment Effects in Economic Laboratory Research." *Research in Experimental Economics* 60, 1996, 239-292.
- Bennett, Charles. "Preliminary Results of the National Research Program's Reporting Compliance Study of Tax Year 2001 Individual Returns." Paper presented at the Annual IRS Research Conference, 2005. Washington, D.C.
- Bigoni, Maria, Gabriel Camera, and Marco Casari. "Strategies of Cooperation and Punishment among Students and Clerical Workers." *Journal of Economic Behavior & Organization*, 94, 2013, 172-182.
- Bott, Kristina, Alexander W. Cappelen, Erik Ø. Sørensen, and Bertil Tungodden. "You've Got Mail: A Randomised Field Experiment on Tax Evasion." Norwegian School of Economics and Business Administration Discussion Paper. Oslo, Norway, 2013.

- Brewer, Marilyn B. "Research Design and Issues of Validity," in *Handbook of Research Methods in Social and Personality Psychology*, edited by Harry T. Reis and Charles M. Judd. Cambridge, UK: Cambridge University Press, 2000, 3-16.
- Brookshire, David, Donald Coursey, and William D. Schulze. "The External Validity of Experimental Economics Techniques: Analysis of Demand Behavior." *Economic Inquiry*, 25 (2), 1987, 239-50.
- Camerer, Colin F. "The Promise and Success of Lab-field Generalizability in Experimental Economics: A Critical Reply to Levitt and List." California Institute of Technology, Division of the Humanities and Social Sciences Working Paper. Los Angeles, CA, 2011.
- Campbell, Donald T., and Julian C. Stanley. *Experimental and Quasi-experimental Designs for Research*. Chicago, IL: Rand McNally College Publishing, Co., 1966.
- Cappelen, Alexander W., Knut Nygaard, Erik Ø. Sørensen, and Bertil Tungodden. "Efficiency, Equality and Reciprocity in Social Preferences: A Comparison of Students and a Representative Population." Norwegian School of Economics and Business Administration Discussion Paper. Oslo, Norway, 2010.
- Carpenter, Jeffrey, and Erika Seki. "Do Social Preferences Increase Productivity? Field Experimental Evidence from Fisherman in Toyama Bay." *Economic Inquiry*, 49 (2), 2011, 612-630.
- Castro, Lucia, and Carlos Scartascini. "Tax Compliance and Enforcement in the Pampas: Evidence from a Field Experiment." IDB Working Paper Series No. IDB-WP-472. Washington, D.C.: Inter-American Development Bank, 2013.
- Charness, Gary, and Marie-Claire Villeval. "Cooperation and Competition in Intergenerational Experiments in the Field and the Laboratory." *The American Economic Review*, 99 (3), 2009, 956-978.
- Chermak, Janie M., Kate Krause, David S. Brookshire, and H. Stu Burness. "Moving Forward by Looking Back: Comparing Laboratory Results with Ex Ante Market Data." *Economic Inquiry*, 51 (1), 2013, 1035-1049.
- Davis, Douglas D., and Charles A. Holt. *Experimental Economics*. Princeton, NJ: Princeton University Press, 1993.
- Dyer, Douglas, John H. Kagel, and Daniel Levin. "A Comparison of Naïve and Experienced Bidders in Common Value Offer Auctions: A Laboratory Analysis." *Economic Journal*, 99 (1), 1989, 108-115.
- Erard, Brian, and Chih-Chin Ho, C.-C. "Searching for Ghosts: Who Are the Non-filers and How Much Tax Do They Owe?" *Journal of Public Economics*, 81 (1), 2001, 25-50.
- Falk, Armin, and James J. Heckman. "Lab Experiments Are a Major Source of Knowledge in the Social Sciences." *Science*, 326 (5952), 2009, 535-538.
- Fellner, Gerlinde, Rupert Sausgruber, and Christian Traxler. "Testing Enforcement Strategies in the Field: Threat, Moral Appeal and Social Information." *Journal of the European Economic Association*, 11 (3), 2011, 634-660.
- Ferber, Robert, and Werner Z. Hirsch. *Social Experimentation and Public Policy*. Cambridge, MA: Cambridge University Press, 1982.
- Frechette, Guillaume R. "Laboratory Experiments: Professionals versus Students," in *The Methods of Modern Experimental Economics*, edited by Guillaume R. Frechette and Andrew Schotter. New York, NY: Oxford University Press, forthcoming.
- Gramlich, Edward M. "Reflections of a Policy Economist." *The American Economist*, 41 (1), 1997, 22-30.

- Gravelle, Jane. "Comments on Innovative Approaches to Improving Tax Compliance." *The IRS Research Bulletin, Recent Research on Tax Administration and Compliance, Selected Papers Given at the 2008 IRS Research Conference*. Washington, D.C., 2009, 59-60.
- Greiner, Ben. "The Online Recruitment System ORSEE 2.0 - A Guide for the Organization of Experiments in Economics." Working Paper Series in Economics 10, Department of Economics, University of Cologne. Cologne, Germany, 2004.
- Güth, Werner, and Oliver Kirchkamp. "Will You Accept Without Knowing What? The Yes-No Game in the Newspaper and in the Lab." *Experimental Economics*, 15 (4), 2012, 656-666.
- Güth, Werner, Carsten Schmidt, and Matthias Sutter. "Bargaining Outside the Lab – A Newspaper Experiment of a Three-person Ultimatum Game." *Economic Journal*, 117 (518), 2007, 449-469.
- Harrison, Glenn W., Morten Lau, and E. Elisabet Rutström. "Theory, Experimental Design and Econometrics Are Complementary (And So Are Lab and Field Experiments)," in *The Methods of Modern Experimental Economics*, edited by Guillaume R. Frechette and Andrew Schotter. New York, NY: Oxford University Press, forthcoming.
- Harrison, Glenn W., and John A. List. "Field Experiments." *Journal of Economic Literature*, 42 (2), 2004, 1009-1055.
- Heckman, James J., and Jeffrey A. Smith. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*, 9 (2), 1995, 85-110.
- Internal Revenue Service. *IRS Data Book 2002*, Publication 55B. Available online at <http://www.irs.gov/pub/irs-soi/02databk.pdf>. Washington, D.C., 2002.
- Internal Revenue Service. "Tax Gap for Tax Year 2006." Available online at http://www.irs.gov/pub/neListwsroom/overview_tax_gap_2006.pdf. Washington, D.C., 2012.
- Iyer, Govind S., Philip M. J. Reckers, and Debra L. Sanders. "Increasing Tax Compliance in Washington State: A Field Experiment." *National Tax Journal* 63 (1), 2010, 7-32.
- Kagel, John H. "Laboratory Experiments," in *The Methods of Modern Experimental Economics*, edited by Guillaume R. Frechette and Andrew Schotter. New York, NY: Oxford University Press, forthcoming.
- Kagel, John H., and Alvin E. Roth (eds.). *The Handbook of Experimental Economics*. Princeton, NJ: Princeton University Press, 1995.
- Kessler, Judd, and Lise Vesterlund. "The External Validity of Laboratory Experiments: The Misleading Emphasis on Quantitative Effects," in *The Methods of Modern Experimental Economics*, edited by Guillaume R. Frechette and Andrew Schotter. New York, NY: Oxford University Press, forthcoming.
- Kleven, Henrik J., Martin B. Knudsen, Claus T. Kreiner, Soren Pedersen, and Emmanuel Saez. "Unwilling or Unable to Cheat? Evidence from a Randomized Tax Audit Experiment in Denmark." *Econometrica*, 79 (3), 2011, 651-692.
- Leamer, Edward E. "Let's Take the Con out of Econometrics." *The American Economic Review*, 73 (1), 1983, 31-43.
- List, John A. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions." *Journal of Political Economy*, 114 (1), 2006, 1-37.
- Levitt, Steven D., and John A. List. "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?" *Journal of Economic Perspectives*, 21 (2), 2007, 153-174.

- Manski, Charles F. "Economic Analysis of Social Interactions." *Journal of Economic Perspectives*, 14 (3), 2000, 115-136.
- McKee, Michael, James Alm, Todd Cherry, and Michael Jones. *Final Report for TIRNO-07-P-00683 on "Behavioral Tax Research."* Washington, D.C., 2008.
- Plott, Charles R. "Dimensions of Parallelism: Some Policy Applications of Experimental Methods," in *Laboratory Experimentation in Economics: Six Points of View*, edited by Alvin E. Roth. New York, NY: Cambridge University Press, 1987, 193-229.
- Pomeranz, Dina. "No taxation without Information: Deterrence and Self-enforcement in the Value Added Tax." NBER Working Paper 19199. Cambridge, MA: National Bureau of Economic Research, 2013.
- Roth, Alvin E. "Laboratory Experimentation in Economics," in *Advances in Economic Theory, Fifth World Congress*, edited by Truman Bewley. Cambridge, UK: Cambridge University Press, 1987, 269-299.
- Scheiber, Noam. "Freaks and Geeks." *The New Republic*, 2 April 2007, 27-31.
- Shadish, William R., Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin, Inc., 2002.
- Shogren, Jason, James Fox, Dermot Hayes, and John Roosen. "Observed Choices for Food Safety in Retail, Survey, and Auction Markets." *American Journal of Agricultural Economics*, 81 (5), 1999, 1192-1199.
- Slemrod, Joel, Marsha Blumenthal, and Charles Christian. "Taxpayer Response to an Increased Probability of Audit: Evidence from a Controlled Experiment in Minnesota." *Journal of Public Economics*, 79 (3), 2001, 455-483.
- Smith, Vernon L. "Experimental Economics: Induced Value Theory." *The American Economic Review Papers and Proceedings*, 66 (2), 1976, 274-279.
- Smith, Vernon L., and James Walker. "Monetary Rewards and Decision Cost in Experimental Economics." *Economic Inquiry*, 31 (2), 1993, 245-261.

TABLE 1
Summary Statistics for Taxpayer Sample ^a

	Taxable Income as Reported			Taxable Income That Should Have Been Reported			Mean Reporting Compliance Rate		
	N	Mean (\$)	Standard Deviation (\$)	Sum (\$ millions)	Mean (\$)	Standard Deviation (\$)	Sum (\$ millions)	Mean of Individuals	Overall Mean
Unweighted	1,101	5,461	12,081	6.0	25,277	132,064	27.8	0.313	0.216
Weighted	559,555	3,708	9,854	2,075.0	16,054	78,165	8,983.3	0.242	0.231

^a The data in this table reflect only the “raw” NRP audit adjustments, and do not account for any unreported income that the auditors did not detect.

TABLE 2
Summary Statistics for Experimental Sample

Audit Probability	Full Sample			Selected Sample			
	Number of Subjects	Number of Observations	Mean Reporting Compliance Rate	Number of Subjects	Number of Observations	Mean Reporting Compliance Rate	Overall Mean Reporting Compliance Rate
0.00	16	240	0.288	16	240	0.288	0.286
0.05	180	2,700	0.413	48	720	0.404	0.368
0.10	356	5,580	0.544	78	1,170	0.475	0.476
0.30	298	4,710	0.590	32	480	0.558	0.536
0.40	222	3,330	0.638	78	1,170	0.672	0.668
Total	1,072	16,560	0.551	252	3,780	0.521	0.517

TABLE 3A
Experimental Treatments – Information Services Experiments

Tax Liability Uncertain?	Information Services Provided?	
	No	Yes
No	T1	---
Yes	T2	T3

TABLE 3B
Experimental Treatments – Positive Inducements via Social Programs Experiments

Positive Inducements Provided?		
No	Yes, via Tax Credit	Yes, via Unemployment Benefits
T4	T5	T6

TABLE 4
Experimental Parameters

Parameter	Values
Income	Mean = 50, High = 100, Low = 10, Increment = 10
Audit Probability	0.3 and 0.4; 0.0 if Not File is selected
Fine Rate	150%, fixed across all sessions
Tax Rate	35%, fixed across all sessions
Tax Deduction	20%, with uncertainty (when present) via a uniform distribution
Tax Credit	Credit=30–0.6*Income, with uncertainty via a uniform distribution
Unemployment Probability	0.2 and 0.4 – fixed for a session
Unemployment Benefit	Benefits = 0.5, 0.6 times reported income average in the past 2 periods

TABLE 5
Descriptive Statistics for Student and Non-student Subjects

Treatment/Students/Non-students	Metric	Students	Non-students
Information Services	Age (years)	20.1	43.8
T1: 40 Students/18 Non-students	Gender (% male)	55.0	18.6
T2: 14 Students/20 Non-students	Dependent (% yes)	81.9	0.0
T3: 18 Students/21 Non-students	Prepare Own Tax (% yes)	27.7	44.1
	Number of Subjects	72	59
Positive Inducements via Social Programs	Age (years)	20.2	43.9
T4: 50 Students/30 Non-students	Gender (% male)	51.7	21.2
T5: 20 Students/38 Non-students	Dependent (% yes)	83.6	3.0
T6: 46 Students/32 Non-students	Prepare Own Tax (% yes)	36.2	48.1
	Number of Subjects	116	100

TABLE 6A
Estimates for Reporting Compliance – Information Services Experiments

Independent Variable	Dependent Variable – Tax Compliance Rate			
	Model 1	Model 2	Model 3	Model 4
Constant	0.9200** (0.0754)	0.9074** (0.076)	0.9337** (0.078)	0.9162** (0.105)
Period Income	-0.0016** (0.0004)	-0.0016** (0.0004)	-0.0017** (0.0003)	-0.0017** (0.0003)
Cumulative Wealth	-0.0003** (0.00003)	-0.0003** (0.00003)	-0.0003** (0.00003)	-0.0003** (0.00003)
Audit Probability	-0.1653 (0.194)	-0.1634 (0.194)	-0.1105 (0.183)	-0.1602 (0.269)
Lag Audit	0.1632** (0.022)	0.1630** (0.022)	-0.0105 (0.015)	-0.0104 (0.015)
Tax Liability Uncertainty	-0.0491** (0.023)	-0.0429* (0.020)	-0.2254** (0.042)	-0.2257** (0.042)
Tax Agency Information	0.0636** (0.0255)	0.0622** (0.024)	0.0752** (0.032)	0.0752** (0.035)
Student		0.0222 (0.020)	-0.0426 (0.037)	-0.0117 (0.135)
Student X Tax Liability Uncertainty			0.3071** (0.057)	0.3078** (0.057)
Student X Tax Agency Information			0.0780 (0.068)	0.0778 (0.068)
Student X Audit Probability				-0.0889 (0.363)
Wald χ^2	228.06**	229.43**	245.58**	247.40**
Panels	131	131	131	131
N	2489	2489	2489	2489

^a Panel estimations with clustered (subject level) standard errors. The dependent variable is the ratio of reported taxes to true taxes of individual i in period t . * and ** indicate significance at the 5 and 1 percent levels, respectively.

TABLE 6B

Estimates for Reporting Compliance – Positive Inducements via Social Programs Experiments

Independent Variable	Dependent Variable – Tax Compliance Rate			
	Model 1	Model 2	Model 3	Model 4
Constant	0.5429** (0.064)	0.5028** (0.081)	0.5537** (0.083)	0.4842** (0.069)
Period Income	-0.0007* (0.0004)	-0.0007* (0.0004)	-0.0007* (0.0003)	-0.0007* (0.0003)
Cumulative Wealth	-0.0001** (0.00003)	-0.0001** (0.00003)	-0.0001** (0.00003)	-0.0001** (0.00003)
Audit Probability	0.0807 (0.132)	0.0906 (0.132)	0.1072 (0.131)	0.1224 (0.156)
Lag Audit	0.0051 (0.015)	0.0049 (0.015)	0.0034 (0.015)	0.0050 (0.0154)
Tax Credit	0.1131** (0.047)	0.1376** (0.055)	0.1259** (0.053)	0.1392** (0.057)
Unemployment Benefit	0.2468** (0.088)	0.2541** (0.089)	0.2364** (0.112)	0.2382* (0.118)
Student		0.0475 (0.054)	0.0103 (0.0864)	-0.1053 (0.122)
Student X Tax Credit			0.1310 (0.098)	0.1313 (0.087)
Student X Unemployment Benefit			0.1182 (0.103)	0.1236 (0.102)
Student X Audit Probability				0.3250 (0.253)
Wald χ^2	803.52**	845.66**	894.84**	904.90**
Panels	216	216	216	216
N	4104	4104	4104	4104

^a Panel estimations with clustered (subject level) standard errors. The dependent variable is the ratio of reported taxes to true taxes of individual i in period t . * and ** indicate significance at the 5 and 1 percent levels, respectively.

FIGURE 1
Distribution of Reporting Compliance Rates for Taxpayer Sample

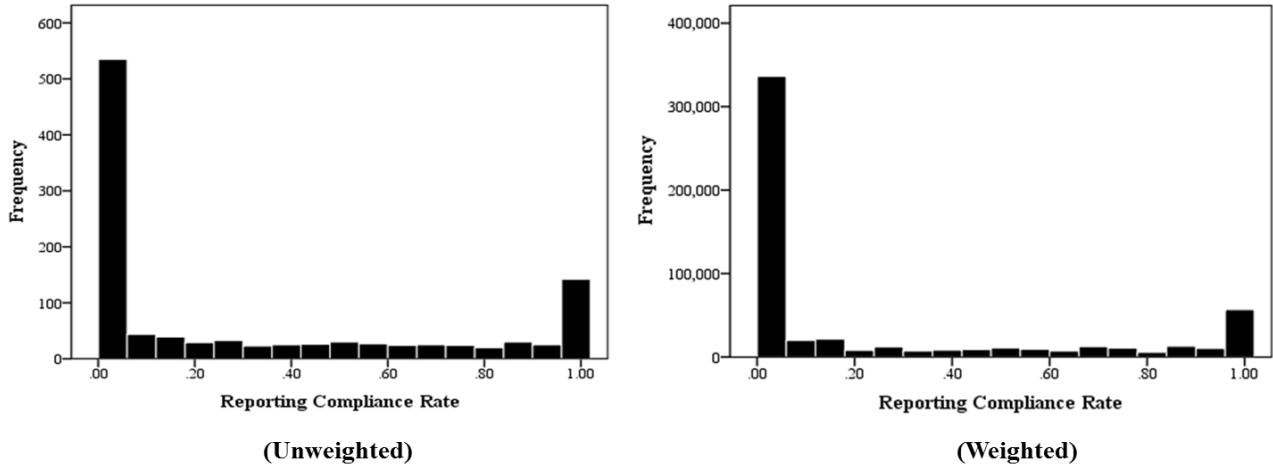
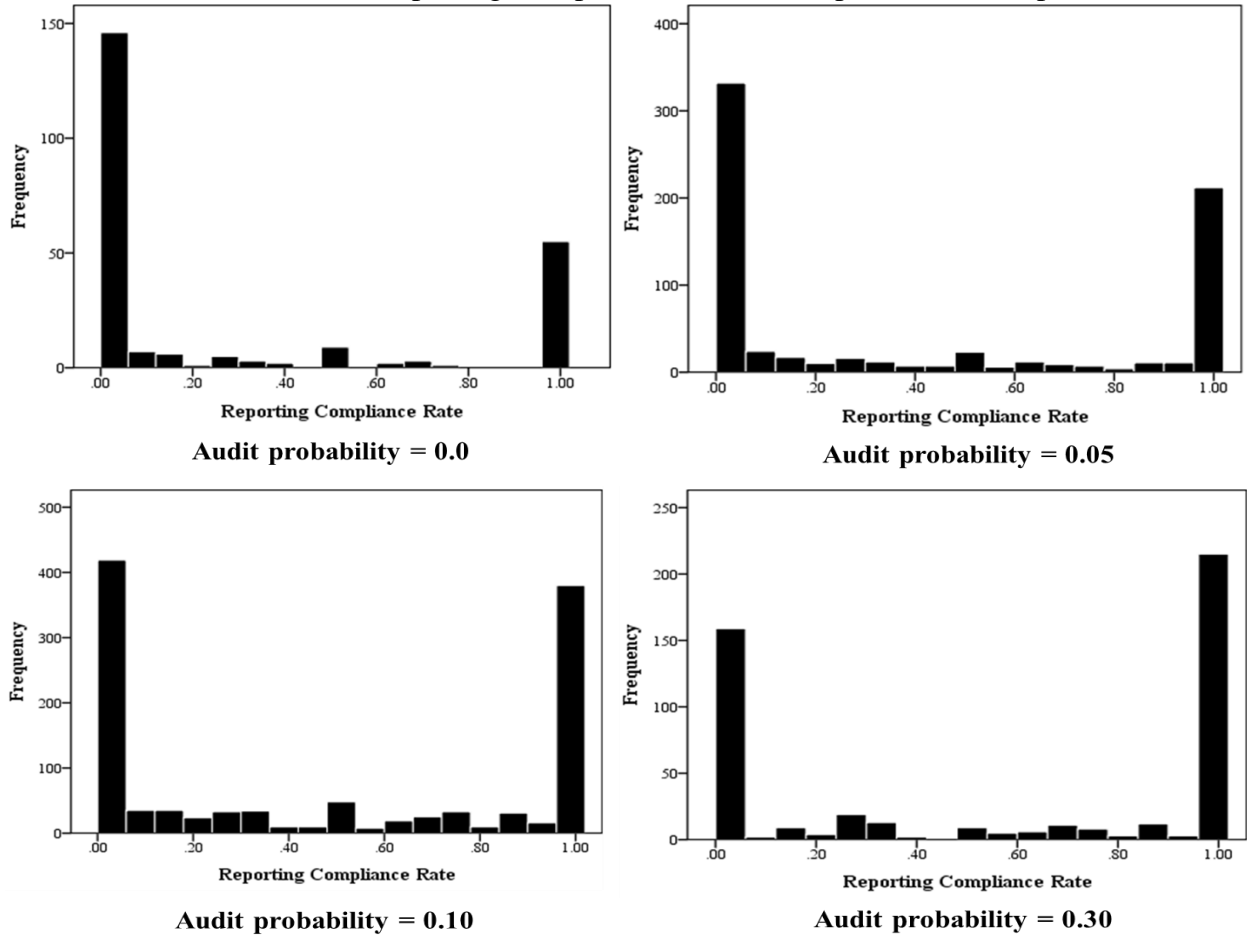


FIGURE 2
Distribution of Reporting Compliance Rates for Experimental Sample



APPENDIX: SAMPLE EXPERIMENTAL INSTRUCTIONS AND SCREEN SHOT
[POSITIVE INDUCEMENTS VIA SOCIAL PROGRAMS EXPERIMENTS – UNEMPLOYMENT BENEFITS]

Introduction

You are about to participate in an experiment in economic decision making. Please follow the instructions carefully, as the amount of money you earn in the experiment will depend on your decisions. At the end of today's session, you will be paid your earnings privately and in cash. Please do not communicate with other participants during the experiment unless instructed. Importantly, please refrain from verbally reacting to events that occur during the experiment.

Today's experiment will involve several decision "rounds". You will not know the number of rounds until the end of the experiment. The rounds are arranged into multiple series. After all decision rounds are finished, we will ask you to complete a questionnaire.

Aside from decisions in "training" rounds, each decision impacts your earnings, which means that it is very important to consider each decision carefully prior to making it. Each decision round is separate from the other rounds, in the sense that the decisions you make in one round will not affect the outcome or earnings of any other round. All money amounts are denominated in lab dollars, and will be exchanged at a rate of xxx lab dollars to US\$1 at the end of the experiment.

There are four parts to each decision round: the Income Earning Stage, the Tax Reporting Stage, the Audit Determination Stage, and the Round Summary Stage. We will now describe each part.

Income Earning Stage:

Each round or period you will complete a task that determines your income for the round. You will be required to sort the numbers 1 through 9 into the correct order. The task is timed. The person completing the task in the shortest time earns the highest income, the second fastest the second highest income and so on.

Tax Reporting Stage:

When the tax year has finished, you enter the tax reporting or filing stage. You will know your income and your allowable deductions and credits but these amounts are not known to the tax agency. You will fill out and file a tax form as you saw in the computer instructions.

After you choose income and deduction amounts to report, you click on the "FILE TAXES" button to submit your tax form. Your taxes are determined by subtracting what you report in deductions from what you report in income, and multiplying this difference by the tax rate of 35%. On your screen, this amount is included among the tax form calculations as "Reported Taxes".

There is a timer on the tax reporting screen. If you do not file the tax form before time runs out, this will be treated the same as if you submitted a form that reported 0 in income and 0 in deductions. In addition, your tax form will automatically be audited. In other words, it is not in your best interest to let the tax reporting screen time out!

Audit Determination Stage:

There is a chance you will randomly be selected for audit. You will know this chance prior to making your tax reporting decisions. The chance does not increase or decrease depending on your

current or past reporting choices or on the decision made by others in the group. This is a random selection process.

After you file the tax form, you will see an audit screen. While you are on this screen, the computer is randomly determining whether to select you for audit. This selection is done separately for each participant and each round.

If you are selected for audit, your reported income, credits, and deductions will be checked against your actual income, credits, and deductions. These amounts will be checked separately. If you underreported your taxes, all unpaid taxes will be discovered. If you are not audited, however, no unpaid taxes will be discovered.

If you are audited, you will have unpaid taxes if you reported too little in income or too much in deductions or credits. Unpaid taxes are calculated as the difference between your actual and reported amounts multiplied by the tax rate. Any unpaid taxes discovered in the audit must be paid back.

If you have unpaid taxes, a penalty of 150% will be assessed. What this means is that, if you are audited, for every lab dollar in unpaid taxes you will have to pay back the 1 dollar you owed, and in addition you will have to pay 0.5 lab dollars in penalties.

Round Summary Stage:

After the tax reporting decision, three things can happen: (1) you are not audited; (2) you are audited but you did not underreport your taxes; or (3) you are audited and you did underreport your taxes. Your earnings are, of course, the same for the first two scenarios. The computer will calculate your earnings for you.

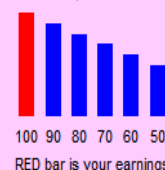
Unemployment

There is a chance that you will be unemployed in a round. The chance of this happening is shown on your screen as described in the computer instructions. If you are unemployed, you will not complete the income earning task in that round. Instead, you will receive unemployment benefits if you filed a tax form in the previous two rounds, calculated as 50% of the average income you reported in the previous two rounds on your filed tax forms. However, if you have not filed a tax form for the previous two rounds (both rounds), your unemployment benefits will be zero, and you will earn no income for the rounds you are unemployed.

Beginning the Experiment

We have now finished the instructions. We will continue on to a second training round. As with the first training round, your decisions in the training round will not affect your earnings. After the training round, you will have a final opportunity to ask questions.

APPENDIX FIGURE 1
Tax Form Screen with Deductions and Tax Credit

Subject 1	Tax Form	Round 1																																
<p>Group Earnings Summary</p> <p align="center">\$LAB</p>  <p>100 90 80 70 60 50 RED bar is your earnings</p>	<table border="1" style="width:100%; border-collapse: collapse;"> <thead> <tr> <th style="width: 60%;">Your Earnings Summary</th> <th style="width: 40%;">Actual</th> </tr> </thead> <tbody> <tr><td>Income earned</td><td align="right">100</td></tr> <tr><td>Deduction (actual is amount allowed)</td><td align="right">15</td></tr> <tr style="background-color: #ADD8E6;"><td>Taxable Income</td><td align="right">85</td></tr> <tr><td>Taxes Owed or Paid</td><td align="right">30</td></tr> <tr><td>Credits (actual is amount allowed)</td><td align="right">0</td></tr> <tr style="background-color: #ADD8E6;"><td>After Tax Income</td><td align="right">70</td></tr> <tr><td>Penalty (under reporting taxes owed)</td><td></td></tr> <tr><td>Penalty for late filing</td><td></td></tr> <tr><td>Cost of Tax Form</td><td></td></tr> <tr style="background-color: #ADD8E6;"><td>Take Home Income</td><td align="right">70</td></tr> </tbody> </table> <table border="1" style="width:100%; border-collapse: collapse;"> <thead> <tr> <th style="width: 50%;">Tax Policy</th> <th style="width: 50%;">Audit Probability</th> </tr> </thead> <tbody> <tr> <td>Tax Rate <small>(tax owed = rate x taxable income)</small></td> <td>35% The likelihood of an audit</td> </tr> <tr> <td>Enforcement Policy</td> <td>Audit Success <small>(percent discovered during audit)</small></td> </tr> <tr> <td>Penalty For Late Filing</td> <td>10% Income</td> </tr> <tr> <td>Penalty Rate <small>(penalty = rate x tax owed)</small></td> <td>50% 100%</td> </tr> </tbody> </table>	Your Earnings Summary	Actual	Income earned	100	Deduction (actual is amount allowed)	15	Taxable Income	85	Taxes Owed or Paid	30	Credits (actual is amount allowed)	0	After Tax Income	70	Penalty (under reporting taxes owed)		Penalty for late filing		Cost of Tax Form		Take Home Income	70	Tax Policy	Audit Probability	Tax Rate <small>(tax owed = rate x taxable income)</small>	35% The likelihood of an audit	Enforcement Policy	Audit Success <small>(percent discovered during audit)</small>	Penalty For Late Filing	10% Income	Penalty Rate <small>(penalty = rate x tax owed)</small>	50% 100%	<p align="center">Department Of Treasury</p> <p align="center">Individual Income Tax Return</p> <hr/> <p>Income 1 Income earned <input type="text" value="0"/> <input type="button" value="7"/> <input type="button" value="8"/> <input type="button" value="9"/> <input type="button" value="4"/> <input type="button" value="5"/> <input type="button" value="6"/> <input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="0"/> <input type="button" value="Clear"/></p> <hr/> <p>Deductions 2 Allowable reductions in income <input type="text" value="0"/> <input type="button" value="7"/> <input type="button" value="8"/> <input type="button" value="9"/> <input type="button" value="4"/> <input type="button" value="5"/> <input type="button" value="6"/> <input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="0"/> <input type="button" value="Clear"/></p> <hr/> <p>3 Taxable income <input type="text" value="0"/></p> <p>4 Initial taxes owed <input type="text" value="0"/></p> <hr/> <p>Tax Credits 5 Credit <input type="text" value="0"/> <input type="button" value="7"/> <input type="button" value="8"/> <input type="button" value="9"/> <input type="button" value="4"/> <input type="button" value="5"/> <input type="button" value="6"/> <input type="button" value="1"/> <input type="button" value="2"/> <input type="button" value="3"/> <input type="button" value="0"/> <input type="button" value="Clear"/></p> <hr/> <p>Taxes 6 Final taxes owed <input type="text" value="0"/></p> <p>7 Income after taxes <input type="text" value="0"/></p>
Your Earnings Summary	Actual																																	
Income earned	100																																	
Deduction (actual is amount allowed)	15																																	
Taxable Income	85																																	
Taxes Owed or Paid	30																																	
Credits (actual is amount allowed)	0																																	
After Tax Income	70																																	
Penalty (under reporting taxes owed)																																		
Penalty for late filing																																		
Cost of Tax Form																																		
Take Home Income	70																																	
Tax Policy	Audit Probability																																	
Tax Rate <small>(tax owed = rate x taxable income)</small>	35% The likelihood of an audit																																	
Enforcement Policy	Audit Success <small>(percent discovered during audit)</small>																																	
Penalty For Late Filing	10% Income																																	
Penalty Rate <small>(penalty = rate x tax owed)</small>	50% 100%																																	
<p>Time Remaining (Seconds) = 61</p> <div style="display: flex; justify-content: space-around;"> <input type="button" value="File"/> <input type="button" value="Not File"/> </div>																																		