EARNED INCOME TAX CREDIT EXPANSIONS
AND FILING BEHAVIOR AMONG ELIGIBLE INDIVIDUALS

A Thesis
submitted to the Faculty of the
Graduate School of Arts and Sciences
of Georgetown University
in partial fulfillment of the requirements for the
degree of
Master
in Public Policy

By

Jay Weismuller, B.A.

Washington, DC
April 11, 2016
This paper examines the relationship between expansions of Earned Income Tax Credit (EITC) benefits and federal tax return filing behavior among EITC eligible individuals. An estimated 13 to 18 percent of individuals who are eligible for EITC do not file tax returns, and therefore do not receive the credit. One understudied approach to reducing the EITC eligible nonfiler rate is increasing EITC benefits, which effectively increases filing incentives. This study uses panel data from the 2008 Survey on Income and Program Participation that track EITC eligible individuals before and after the EITC expansion in the American Recovery and Reinvestment Act of 2009 ($n = 111,057$). A cross-sectional Heckit model and fixed effects linear probability model estimate that a $100 increase in EITC is associated with a 5.1 to 5.9 percent increase in the 2009 filing propensity of 2007 EITC eligible nonfilers ($p < 0.000$). A generalized ordered logit model estimates that, among EITC eligible individuals without a filing requirement in 2007 and 2009, a $100 increase in EITC is associated with a 0.6 percent increase in the probability of persistent filing and a 0.4 percent decrease in the probability of persistent nonfiling across both years ($p < 0.000$). Greater participation should be counted among the potential benefits of EITC expansions.

**INDEX WORDS:** Earned Income Tax Credit, EITC, Eligible, Nonfiler, Participation
# Table of Contents

**Section**

1 Introduction ........................................ 1  
   1.1 EITC Eligible Nonfilers .......................... 1  
   1.2 The EITC Filing Decision ......................... 3  
   1.3 Study Contributions .............................. 5  

2 Data ................................................. 7  
   2.1 SIPP Survey Design ............................... 7  
   2.2 Analytic Data Set ................................ 8  
   2.3 Descriptive Statistics ............................ 12  

3 Econometric Methods .................................. 15  
   3.1 Identification Strategy ........................... 15  
   3.2 Cross-Sectional Heckit ............................ 18  
   3.3 Fixed Effects Linear Probability Model ............ 20  
   3.4 Generalized Ordered Logit ........................ 23  

4 Results .............................................. 25  
   4.1 2007 EITC Eligible Nonfilers .................... 25  
   4.2 EITC Eligible Individuals Without a Filing Requirement ..... 29  

5 Discussion .......................................... 33  
   5.1 Policy Implications ................................ 33  
   5.2 Study Limitations and Extensions ................. 35  

Bibliography ............................................ 43
LIST OF FIGURES

2.1 Timeline of Survey on Income and Program Participation administration, filing season deadlines, and tax policy developments from January 2007 to April 2011 ................................................. 11

4.1 Comparison of cross-sectional Heckit (Model 3) and fixed effects LPM (Model 4) coefficient and confidence interval estimates for EITC size and select tax variables ........................................ 27

4.2 Comparison of cross-sectional Heckit (Model 3) and fixed effects LPM (Model 4) predictions of filing propensity by EITC benefit size .... 28
LIST OF TABLES

2.1 Descriptive statistics for the full sample and EITC eligible nonfilers in 2008 and 2009 ................................................. 14
3.1 Unweighted correlations between tax control variables and dependent and key independent variable ........................................ 21
4.1 2009 filing propensity of 2007 EITC eligible nonfilers: average marginal effects of cross-sectional probit models and estimates of fixed effects linear probability model ........................................... 26
4.2 First-difference generalized ordered logit model of the relationship between change in EITC and predicted probability of filing in both 2007 and 2009 among EITC eligible individuals without a filing requirement ..................................................... 31
4.3 First-difference generalized ordered logit model of the relationship between change in EITC and predicted probability of not filing in both 2007 and 2009 among EITC eligible individuals without a filing requirement ..................................................... 32
A.1 Descriptive statistics for EITC eligible individuals without a federal filing requirement in 2008 and 2009 ........................................ 40
A.2 2009 filing propensity of 2007 EITC eligible nonfilers: marginal effects of cross-sectional Heckit model across the EITC distribution .............. 41
A.3 Fixed effects linear probability model of the relationship between EITC size and filing propensity among EITC eligible individuals without a filing requirement ..................................................... 42
SECTION 1

INTRODUCTION

The Earned Income Tax Credit (EITC) is a refundable tax credit that subsidizes the wages of low-income earners by providing tax refunds to filers when the value of the credit exceeds their tax liability. The EITC is the largest needs-tested anti-poverty cash assistance program in the United States, with 28.8 million tax return filers claiming $68.1 billion in EITC benefits in 2013 (Falk & Crandall-Hollick 2016). The policy goals of the EITC are to reduce poverty and increase labor supply among low-income individuals.

1.1 EITC Eligible Nonfilers

The effectiveness of EITC is limited by nonparticipation. The EITC participation rate is the share of EITC eligible individuals claiming the credit to the total number of EITC eligible individuals in a given time period.\(^1\) The 1990 EITC participation rate was estimated to be between 80 and 86 percent (Scholz 1994). In 1999, the estimated rate was 75 percent (GAO 2001). More recent estimates range from 75 percent in 2005 (Plueger 2005) to 53 percent in 2009 (Short et al. 2012). The majority of studies indicate that roughly three-fourths of EITC eligible individuals participate in the program (Dickert et al. 2005).

\(^1\)Participation can also be considered the share of EITC benefits claimed to total benefits individuals are eligible to claim, in dollar terms. For this study, which focuses on nonfiling, the share of participating individuals is more relevant than the share of claimed benefits.
EITC nonparticipants fall into one of two groups: nonclaimants or nonfilers. EITC nonclaimants file tax returns but do not claim the credit by filing a Form 8862. EITC nonfilers do not file tax returns at all. The majority of EITC nonparticipants are nonfilers. In tax year 2009, 59 percent of EITC nonparticipants were nonfilers and 41 percent were nonclaimants (Short et al. 2012).

Out of 39 million total nonfilers in tax year 2012, about 4 million were estimated to be potentially EITC eligible (Guyton et al. 2016). While recent estimates of the EITC eligible nonfiler rate are not available, Internal Revenue Service (IRS) estimates indicate that 12.8 to 17.8 percent of EITC eligible individuals did not file returns in 1996, with 2 to 3 million EITC eligible nonfilers failing to claim 2 to 3.5 billion in benefits (Maynard & Dollins 2002).

Efforts to increase EITC participation often use outreach to potentially eligible individuals. Examples include mailed brochures with filing reminders and claiming instructions. The IRS can target EITC nonclaimants with outreach because they file tax returns with identifying information. But identifying EITC nonfilers is more demanding. The problem of identifying nonfilers is so vexing that researchers have dubbed these individuals "ghosts" (Alm et al. 2012; Erard & Ho 2001), which have themselves always presented unique empirical challenges.

Researchers have studied the characteristics of EITC eligible nonfilers in order to better target outreach (Maynard & Dollins 2002). EITC nonfilers are disproportionately male, unmarried, and Hispanic (Scholz 1994). Thirty percent of EITC nonfilers in 2009 were Hispanic (Short et al. 2012). Eligible nonfilers are also disproportionately unbanked (Lim et al. 2011). Many EITC nonfilers have no qualifying children under EITC rules, with 43 percent in 2005 (Plueger 2005) and 40 percent in 1996 being eligible for the childless credit (Maynard & Dollins 2002).
Most EITC eligible nonfilers have especially low incomes and are eligible for small EITC amounts relative to the credit caps. In a stratified random sample of about 360,000 potentially EITC eligible nonfilers from 2011 and 2012, the average annual gross income was $7,350 (Guyton et al. 2016), and in 2005 roughly half of EITC nonfilers had adjusted gross income (AGI) under $8,000 (Plueger 2005). Among the 2011 and 2012 nonfilers, the average EITC eligibility was $576. By comparison, the maximum credit in 2012 was $3,169 for single-child, $5,236 for two-child, and $5,891 for three-child families.

In 2014 and 2015, IRS researchers conducted a series of randomized controlled trials to evaluate the effectiveness of outreach to EITC nonfilers in encouraging participation. These field studies estimate that simple outreach with postcards and brochures is associated with a 0.5 to 1 percent increase in filing among EITC nonfilers (Guyton et al. 2016).

1.2 THE EITC FILING DECISION

An understudied alternative to outreach is increasing EITC benefits. The standard economic model of taxpayer filing behavior treats the filing decision as a cost-benefit analysis weighing the benefits of avoiding tax against the burdens of filing and expected penalties for not filing (Erard & Ho 2001). This model can be expanded to include the filing incentives provided by tax credits. Expanding refundable EITC benefits for eligible individuals increases the benefits of filing relative to the costs. Economic theory therefore predicts that this would increase filing among EITC eligible individuals and decrease the EITC eligible nonfiler rate.

A key factor in tax return filing decisions is whether an individual has a federal filing requirement. Individuals are required to file tax returns if their gross income
exceeds the sum of their standard deduction and personal exemptions, set in 2016 at $10,350 for singles and $28,800 for a family of four. For higher-income individuals, having a filing requirement and being subject to penalties for not filing present substantial costs to being a nonfiler. A likely motivation for filing is avoiding these costs.

Approximately 73 percent of EITC eligible nonfilers from 2011 and 2012 had no filing requirement in 2014 (Guyton et al. 2016). In tax year 2005, 60 percent of EITC nonfilers did not have a filing requirement (Plueger 2005). Since these lower-income individuals do not face penalties for not filing returns, the costs of filing are more likely to exceed the benefits than for higher-income individuals. Consistent with economic theory, 30 to 39 percent of EITC eligible individuals below the filing threshold participate in EITC, while 89 percent of eligible individuals above the threshold participate (Blumenthal et al. 2005).

For low-income individuals without a filing requirement, legal non-filing penalties are not a factor in the filing decision. Still, when individuals eligible for refundable tax credits do not file returns, they pay an implicit penalty the size of their foregone benefits. Alm et al. (2012) argue that filing decision models should be extended to incorporate the filing incentives provided by tax credits. Under this extended model, the filing decisions of EITC eligible individuals below the filing threshold are likely a function of the burden of filing relative to the tax refunds for which they are eligible.

Considering that the majority of EITC eligible nonfilers do not have a filing requirement, it is likely that their filing decisions are also considerably influenced by refundable credits. In examining the relationship between EITC expansion and eligible nonfiling, it is important to estimate the relationship between EITC benefit
size and filing propensity specifically for EITC eligible individuals below the filing threshold.

1.3 Study Contributions

Two econometric studies estimate the relationship between EITC benefit size and the probability of filing a federal tax return among EITC eligible individuals. Scholz (1994) finds that a $100 increase in EITC is associated with a 2.6 percent increase in the probability of filing a federal tax return among all EITC eligible individuals. Blumenthal et al. (2005) find that a $100 EITC increase is associated with a 2.8 percent increase in filing propensity among EITC eligible individuals with children and a federal filing requirement. Neither of these studies estimate the relationship between EITC and filing propensity specifically for EITC eligible nonfilers or EITC eligible individuals without a filing requirement.

This study contributes to the literature by estimating the relationship between EITC benefit size and the probability of filing a federal tax return among EITC eligible nonfilers and EITC eligible individuals below the filing threshold. Panel data from the Census Bureau’s 2008 Survey on Income and Program Participation (SIPP) are used that track individual filing behavior before and after the EITC expansion passed as part of the American Recovery and Reinvestment Act of 2009 (ARRA). This expansion increased the maximum EITC credit for filers with three or more children and extended the phase-out for joint filers.

This study also contributes to the literature by estimating the relationship between EITC benefit size and filing behavior with more recent data than the 1990 SIPP data from Scholz (1994) and the 1988 administrative data from Blumenthal et al. (2005). There is evidence that the introduction of electronic filing in the 1990s
increased EITC claims (Kopczuk & Pop-Eleches 2007), and that public awareness of EITC benefits and eligibility has grown in recent decades (Caputo 2010). Both trends would be captured by the 2008 SIPP.

Section 2 covers the SIPP survey design, describes the construction of the analytic data set for this study, and reports descriptive statistics. Section 3 explains the econometric methods of the study, beginning with an identification strategy. It then specifies a cross-sectional Heckit model and a fixed effects linear probability model for 2007 EITC eligible nonfilers. Last, it specifies a generalized ordered logit model for EITC eligible individuals without a filing requirement. Section 4 reports and interprets the results. Section 5 concludes by discussing the implications of these results for policymakers and the limitations of the study.
2.1 SIPP Survey Design

The SIPP is administered by the U.S. Census Bureau. The goal of the survey is to facilitate research evaluating the effectiveness of public programs by providing comprehensive information about the income, labor force participation, social program participation, and demographic characteristics of U.S. individuals and households. The SIPP randomly samples the non-institutionalized, resident U.S. population using a multi-stage stratified design.

The 2008 SIPP survey was a series of continuous national panels administered between September 2008 and December 2013, with a total of 16 four-month interview waves. Sampled households were contacted every four months, once during each wave, and provided reports for the current month and the three preceding reference months. Rather than contacting all households in the same month, the SIPP used four staggered "rotation groups" that were contacted in different months across the four-month wave. Interviews were conducted primarily via telephone. Self-reports were gathered from all household members ages 15 and older, who also reported on behalf of household members younger than 15. The overall weighted response rate was 80.64 percent for Wave 1 and 74.07 percent for Wave 2 (Clark & Mack 2009).

Each SIPP wave consists of a core survey and a topical module questionnaire. The core survey is designed to measure the economic status of respondents with
questions centered around income and program participation. The topical module questions vary from wave to wave, covering topics ranging from taxes to medical expenses. This analysis uses the tax topical modules.

2.2 Analytic Data Set

The analytic data set for this study uses the four SIPP waves with tax topical modules: Waves 1, 2, 5, and 8. With these data, it can be determined whether individuals filed 2007 and 2009 tax returns. The other SIPP waves contain relevant income and demographic information, but due to the SIPP weighting scheme, only data from the waves including the tax modules can be used; otherwise, the sample cannot be projected to the population of EITC eligible individuals (see Maynard & Dollins 2002, Appendix C). In total, these four waves consist of 1,518,460 observations.

In Waves 1 and 2, which cover September 2008 to April 2009, individuals were asked whether they received the stimulus payments passed as part of the Economic Stimulus Act of 2008. These payments were administered by the IRS as tax rebates in 2008, so to receive them individuals must have filed 2007 tax returns. Eligible individuals received between $300 and $600 and joint filers received between $600 and $1200, with an additional $300 provided per qualifying child. To be eligible, filers had to have earned income and benefits greater than $3,000. The rebate phase-out began at $75,000 in AGI.

Eligible individuals who reported receiving the stimulus rebate in either Wave 1 or Wave 2 are coded as having filed a tax return. Eligible individuals who reported not receiving the stimulus rebate in either Wave 1 or Wave 2 are coded as having not filed a tax return. Though misreports of rebate receipt were certainly possible, given the salience of the payment, reporting is assumed to be reliable. The SIPP does
not provide a way to determine whether rebate ineligible individuals filed 2007 tax returns. This limitation is discussed in subsection 5.2.

Waves 5 and 8 cover January to April 2010 and January to April 2011, respectively. In both waves, individuals were asked directly whether they filed 2009 tax returns. Individuals who reported filing a 2009 return in either Wave 5 or Wave 8 are coded as having filed. Individuals who reported not filing in both waves are coded as having not filed.

The core SIPP survey is detailed enough to allow EITC eligibility and credit size to be determined. Previous studies have also used core data from SIPP surveys to estimate EITC levels (Maynard & Dollins 2002; Plueger 2005; Scholz 1994). The 2008 SIPP did ask individuals whether they received EITC and the amount they received, but researchers have cited concerns about the accuracy of similar EITC questions asked in previous SIPP panels (Maynard & Dollins 2002; Plueger 2005).

Rather than rely on individual EITC reports, this study uses the core survey income data to calculate individual EITC eligibility, as well as other tax characteristics. Individual tax liabilities under federal and state income tax laws are calculated using the National Bureau of Economic Research’s TAXSIM program (version 9) (Feenberg & Coutts 1993). TAXSIM output includes federal and state EITC credit size, federal and state tax liability, and eligibility for other federal and state tax credits.

An acknowledged source of measurement error in the SIPP and similar panel data sets is "seam bias." This bias results from individuals systematically reporting personal transitions and changes in interview months, rather than in the preceding reference months when the changes occurred (Rothstein & Valletta 2014). Seam bias has small effects on cross-sectional estimates based on all four rotation groups, because only one of the groups is on the seam at a given time (Clark & Mack 2009).
Still, because the reporting months are more accurate than reference months, a common and conservative strategy is to only keep the reporting month data (Schaefer 2013). Accordingly, the analytic data set for this study excludes reference months, leaving 219,102 observations.

Income in the SIPP is reported as monthly income. In order to calculate annual income for use in TAXSIM, monthly income has to be annualized. One option would be to average across months. For example, to proxy for tax year 2007 income, the two monthly incomes observed in Waves 1 and 2 could be averaged and annualized. An alternative is to annualize the income from the observed month closest to the filing season, under the assumption that this month’s income figures most prominently in the filing decision.

The latter approach is taken in this study. Wave 1, which ran from September to December 2008, is closest to the 2007 filing season. Wave 5, which ran from January to April 2010, includes the 2009 filing season. Waves 2 and 8 are dropped from the analytic data set after being used to construct the filing variables outlined above. The limitations of this approach are discussed in subsection 5.2.

Figure 2.1 presents a timeline showing how the four SIPP waves overlap with the 2007 and 2009 filing seasons, the 2008 stimulus payments, and the 2009 EITC expansion. Notes are provided regarding the construction of key variables in the analysis.

The final analytic sample consists of 173,984 observations in a two-year panel covering 2008 and 2009, with each individual having one observation per year. The resulting panel is unbalanced, with 105,663 observations in 2008 and 68,321 observations in 2009. 42,736 individuals are observed in 2008 but not in 2009, and 5,394 individuals are observed in 2009 but not in 2008. The sample contains a total of
Figure 2.1: Timeline of Survey on Income and Program Participation administration, filing season deadlines, and tax policy developments from January 2007 to April 2011
111,057 unique individual observations, with 62,927 individuals tracked across the two years.

The analytic data set includes SIPP probability sampling weights and replicate weights. The sampling weights compensate for the differential representation in the sample of low-income individuals, who were oversampled by the survey (Clark & Mack 2009). Individuals are assigned the sampling weights for the wave of their first entry into the two year analytic panel. Wave 1 weights are applied to the roughly 63,000 individuals who participated across the panel. The 48,000 individuals who only participated in a single wave are given the respective wave’s weight. SIPP replicate weights are also applied to each individual to compensate for the multi-stage stratified sampling design of the survey. Longitudinal weights are not incorporated in the analysis, which is discussed as a limitation in subsection 5.2.

The only data imported from outside the SIPP are IRS audit rates by AGI in 2008 and 2009, obtained from the IRS Statistics of Income (IRS 2008). Analyses are conducted using Stata/IC version 13.1.

2.3 DESCRIPTIVE STATISTICS

Table 2.3 provides descriptive statistics for the full analytic sample and for EITC eligible nonfilers. Appendix Table A.1 reports descriptive statistics for EITC eligible individuals below the filing threshold. In the full sample, 16 percent of individuals are EITC eligible in 2008 and 19 percent are eligible in 2009. Across the two years, between 69 and 76 percent of individuals have no federal filing requirement. The average EITC benefit for eligible nonfilers ranges from $571 in 2008 to $418 in 2009. The mean EITC benefit size for eligible individuals below the filing threshold is nearly identical to eligible nonfilers in both years. Nearly all EITC eligible nonfilers
are below the filing threshold in 2008 (94 percent) and 2009 (99 percent), and the average nonfiler has negative tax liability in both years.

While roughly 16 percent of the full sample is Hispanic across the two years, 28 percent of eligible nonfilers in the data are Hispanic in 2008 and 22 percent are Hispanic in 2009. Eligible nonfilers are also more likely to be single than the average individual in the full sample. While 41 percent of all individuals are married, roughly 20 to 33 percent of eligible nonfilers are married. Individuals in the full sample, nonfilers, and individuals below the threshold have a similar number of children on average.
Table 2.1: Descriptive statistics for the full sample and EITC eligible nonfilers in 2008 and 2009

<table>
<thead>
<tr>
<th>Binary Variables</th>
<th>All Individuals</th>
<th>All Individuals</th>
<th>EITC Eligible Nonfilers</th>
<th>EITC Eligible Nonfilers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2008 (Observations)</td>
<td>2009 (Observations)</td>
<td>2008</td>
<td>2009</td>
</tr>
<tr>
<td></td>
<td>105,663</td>
<td>68,321</td>
<td>3,930</td>
<td>4,943</td>
</tr>
<tr>
<td>EITC Eligible</td>
<td>17,135 (16.15)</td>
<td>13,549 (19.25)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>No Fed. Filing Requirement</td>
<td>73,990 (68.85)</td>
<td>51,841 (75.97)</td>
<td>3,639 (93.54)</td>
<td>4,815 (98.91)</td>
</tr>
<tr>
<td>ACTC Eligible</td>
<td>8,446 (7.70)</td>
<td>7,837 (10.59)</td>
<td>1,308 (31.75)</td>
<td>2,818 (55.72)</td>
</tr>
<tr>
<td>State EITC Eligible</td>
<td>6,036 (4.93)</td>
<td>5,992 (7.49)</td>
<td>1,379 (29.40)</td>
<td>2,068 (37.51)</td>
</tr>
<tr>
<td>No State Filing Requirement</td>
<td>74,907 (70.47)</td>
<td>49,939 (73.84)</td>
<td>2,991 (77.15)</td>
<td>3,526 (72.30)</td>
</tr>
<tr>
<td>Metropolitan Resident</td>
<td>81,509 (83.49)</td>
<td>52,557 (78.80)</td>
<td>3,130 (85.61)</td>
<td>3,815 (72.49)</td>
</tr>
<tr>
<td>White</td>
<td>82,703 (79.73)</td>
<td>53,914 (75.97)</td>
<td>3,067 (80.15)</td>
<td>3,948 (80.59)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>14,079 (15.70)</td>
<td>9,135 (15.65)</td>
<td>1,002 (28.44)</td>
<td>932 (21.87)</td>
</tr>
<tr>
<td>Male</td>
<td>50,837 (49.02)</td>
<td>32,895 (48.94)</td>
<td>1,922 (49.88)</td>
<td>1,922 (49.60)</td>
</tr>
<tr>
<td>High School Graduate</td>
<td>68,014 (65.40)</td>
<td>44,924 (65.92)</td>
<td>2,796 (76.72)</td>
<td>4,062 (82.22)</td>
</tr>
<tr>
<td>Family Household</td>
<td>88,742 (83.84)</td>
<td>57,597 (84.28)</td>
<td>3,501 (89.16)</td>
<td>4,516 (91.53)</td>
</tr>
<tr>
<td>Female Headed Household</td>
<td>25,628 (23.81)</td>
<td>16,647 (23.85)</td>
<td>989 (26.57)</td>
<td>964 (19.63)</td>
</tr>
<tr>
<td>Married</td>
<td>43,914 (41.30)</td>
<td>28,239 (40.90)</td>
<td>1,348 (33.18)</td>
<td>2,759 (19.63)</td>
</tr>
<tr>
<td>Continuous Variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EITC ($)</td>
<td>289.67 ± 894.20</td>
<td>437.68 ± 1,173.63</td>
<td>570.62 ± 1,317.31</td>
<td>417.85 ± 1,052.34</td>
</tr>
<tr>
<td>ACTC ($)</td>
<td>77.53 ± 342.19</td>
<td>135.75 ± 481.20</td>
<td>178.69 ± 545.09</td>
<td>112.6 ± 407.57</td>
</tr>
<tr>
<td>State EITC ($)</td>
<td>16.29 ± 121.22</td>
<td>28.65 ± 167.18</td>
<td>37.15 ± 190.77</td>
<td>23.5 ± 145.44</td>
</tr>
<tr>
<td>Fed. Inc. Tax Liability ($)</td>
<td>1,827.44 ± 8,463.10</td>
<td>2,124.46 ± 9,413.80</td>
<td>-882.50 ± 1,932.75</td>
<td>-527.35 ± 1,307.82</td>
</tr>
<tr>
<td>State Inc. Tax Liability ($)</td>
<td>571.18 ± 1915.02</td>
<td>472.32 ± 2,017.52</td>
<td>-6.33 ± 351.36</td>
<td>-4.37 ± 225.17</td>
</tr>
<tr>
<td>Federal AGI ($)</td>
<td>20,686.85 ± 39,172.38</td>
<td>21,700.55 ± 42,575.86</td>
<td>7085.15 ± 13390.14</td>
<td>4,089.61 ± 8,997.74</td>
</tr>
<tr>
<td>Age</td>
<td>36.58 ± 22.35</td>
<td>37.23 ± 22.90</td>
<td>34.02 ± 23.96</td>
<td>33.75 ± 24.91</td>
</tr>
<tr>
<td>Children under 18</td>
<td>1.20 ± 1.36</td>
<td>1.067 ± 1.32</td>
<td>1.29 ± 1.38</td>
<td>1.40 ± 1.44</td>
</tr>
</tbody>
</table>
3.1 Identification Strategy

Under the theoretical model discussed in subsection 1.2, the filing decisions of EITC eligible nonfilers and eligible individuals below the filing threshold are a function of the burden of filing relative to the tax refunds for which they are eligible. This study constructs several econometric models of the relationship between filing propensity and EITC benefit size on the basis of this theoretical model. The models control for several tax and demographic variables in order to properly identify this relationship.

Continuous measures of eligibility for specific tax credits in dollar terms are added as covariates, including variables for the Additional Child Tax Credit (ACTC), Child Tax Credit (CTC), state EITC, and total state tax credits. ACTC is partially refundable and state credits are partially or fully refundable, and so they are included directly in the models. CTC is the non-refundable portion of the child credit, which means it cannot directly provide a filing incentive. But CTC can bring down tax liability such that the refundable portion of EITC increases, so an interaction term between CTC and EITC is included in the models. An interaction between EITC and having a filing requirement is included under the assumption that EITC benefits provide different filing incentives when fully and when partially refundable.

Variables reflecting filing status and tax liability are also included. The models incorporate binary indicators for whether individuals had federal and state filing
requirements. The impact of filing requirements on filing propensity was discussed in subsection 1.2. Also added are continuous dollar measures of federal AGI, federal income tax liability, and state income tax liability, as well as the IRS audit rate in percentage terms. Income tax liability measures either filing incentives when the liability is negative or the costs of nonfiling when the liability is positive, as individuals who owe tax face penalties for not filing returns. The expected penalties for nonfiling are a function of the IRS audit rate.

The models also incorporate variables representing demographic characteristics. As outlined in subsection 1.1, EITC eligible nonfilers are disproportionately male, unmarried, and Hispanic, so binary indicators for all three are included in the model. Also included are binary indicators for metro status, white race, high school graduate status, family household status, female headed household status, and marital status. Number of children under 18 and age are included as continuous variables, with age-squared also added to account for an anticipated nonlinear relationship between age and filing propensity.

An ideal econometric model of the relationship between filing propensity and EITC benefit size would have four characteristics. First, it would incorporate SIPP weights. As mentioned above, the SIPP sampling weights adjust for the oversampling of low-income individuals and the replicate weights adjust for the multi-stage stratified sampling design of the survey.

Second, the model would control for unobserved heterogeneity with fixed effects. Unobserved individual and time effects likely have an impact on both EITC benefit size and filing propensity. An example is ability. Lower ability reduces wages, which in turn increases expected EITC benefits. Lower ability also increases the burden of filing by making tax complexity more, well, taxing. Omitting controls for unobserved heterogeneity likely biases models of the relationship between EITC benefits
and filing propensity. Fixed effects also control for macroeconomic factors like cost of living, which impact the marginal benefit of a given EITC increase.

Third, the model would be nonlinear. With a binary dependent variable, such as filing status, a linear probability model (LPM) has two disadvantages: it can generate fitted values outside the boundaries of the dependent variable, and it assumes constant marginal effects. For the current model, an LPM could predict a greater than 100 percent or less than 0 percent probability of an individual filing a return for certain EITC benefit levels. Nonlinear models such as probits and logits are therefore typically preferred to LPMs on curve-fitting grounds (Angrist & Pischke 2008). Further, economic theory suggests that the marginal effect of EITC increases on tax filing behavior would not be constant across all values of EITC, but decrease as a function of the diminishing marginal utility of benefits.

Fourth, the model would correct for the selection bias resulting from restricting the analysis to EITC eligible nonfilers or EITC eligible individuals without a filing requirement. This would be achieved with selection correction methods or a model that conditions on nonfiler or below-threshold status without truncating the sample.

This study does not offer a model for EITC eligible nonfilers that satisfies all four of the conditions above. Several issues arise in trying to estimate such a model for nonfilers with filing status as the dependent variable (see pg.24, footnote 2). As outlined below, the method adopted for this study is to compare two models with different relative virtues.

However, it is possible to meet these four conditions with a model of EITC eligible individuals below the filing threshold. Given the large overlap between the two populations and the similarity in their filing decisions, this approach is used as a way to indirectly estimate filing behavior among EITC eligible nonfilers.
The subsections below begin by outlining and comparing the advantages of two econometric models of filing behavior among EITC eligible nonfilers. Then the model for EITC eligible individuals below the filing threshold is discussed.

3.2 CROSS-SECTIONAL HECKIT

This subsection specifies a model of the relationship between EITC benefit size and the probability of filing a federal tax return among 2007 EITC eligible nonfilers. Using a cross-section of 2007 EITC nonfilers in 2009, MODEL 1 estimates the bivariate probit equation:

\[ P(\text{FiledReturn} = 1 | \text{EITC}) = \phi(\text{EITC}'\beta), \]

where \text{FiledReturn} is an indicator for whether an individual filed a 2009 return and \text{EITC} is a dollar measure of the EITC credit for which an individual was eligible in 2009. In order to estimate this relationship specifically for nonfilers, MODEL 1 is run on a subsample of 2007 EITC nonfilers in 2009, rather than all individuals in the data set.

The results of MODEL 1 cannot be causally interpreted because it fails to control for observed confounding factors that vary over time. Omitting from the model any observed characteristics that are correlated with both filing propensity and EITC benefit size biases the estimated coefficient on \text{EITC}. There are two groups of observed omitted variables that could potentially bias MODEL 1: tax controls relating to the costs and benefits of filing and demographic characteristics.

For example, one tax variable omitted from the model is ACTC benefit size. As a refundable credit, ACTC provides a benefit to low-income individuals who choose to file. Further, ACTC benefit size is mechanically linked to EITC benefit size because
eligibility for both is determined by income and number of qualifying children. Economic theory therefore suggests ACTC benefit size is correlated with both filing propensity and EITC size, so omitting it from the model likely biases the EITC coefficient.

Table 3.2 reports Pearson and point-biserial correlation coefficients for the filing and EITC variables and a set of key tax control variables. All correlations are statistically significant, with the exception of state income tax liability and filing a tax return. Notably, most of the individual coefficients are below conventional correlation thresholds (|r| < 0.3), though four coefficients are above the threshold. The tax variables likely confound jointly, if at all.

MODEL 2 controls for observed confounding tax and demographic factors. It estimates the multivariate probit equation:

\[
P(\text{FiledReturn} = 1 \mid EITC, X_{tax}, X_{dem}) = \phi(EITC' \beta_1, X_{tax}' \beta_2, X_{dem}' \beta_3),
\]  

(3.2)

where \(X_{tax}\) is a vector containing variables that reflect the costs and benefits of filing a tax return for EITC eligible nonfilers and \(X_{dem}\) is a vector of demographic characteristics. The vectors contain the full set of control variables defined in subsection 3.1.

Because MODEL 2 is run on a truncated sample of 2007 EITC nonfilers, its estimates are subject to selection bias. The presence of selection bias can be viewed as an omitted variable problem in the selected sample (Wooldridge 2002). To correct for this bias, MODEL 3 uses a one-step maximum likelihood estimation (MLE) of a two-step Heckman sample selection model.

The standard two-step Heckman correction would adjust for sample selection by estimating the first stage probit equation:

\[
P(EITC_{nonfiler} = 1 \mid z) = \phi(z \gamma),
\]  

(3.3)
where $EITC_{nonfiler}$ is an indicator for whether an individual is an EITC eligible nonfiler, $z$ is a vector of explanatory variables for nonfiler status, and $\gamma$ is a vector of unknown parameters related to nonfiler status. The second Heckman stage would then estimate the probit equation:

$$P(\text{Filed}=1|EITC, X_{tax}, X_{dem}, \hat{\lambda}_i) = \phi(EITC'\beta_1, X_{tax}'\beta_2, X_{dem}'\beta_3, \hat{\lambda}_i'\beta_4),$$  \hspace{1cm} (3.4)

where $\hat{\lambda}_i$ is the inverse Mills ratio obtained from equation 3.3 by $\hat{\lambda}_i = \lambda(z, \hat{\gamma})$. With the inclusion of the inverse Mills ratio as an additional covariate, the estimators in equation 3.4 are consistent and approximately normally distributed (Wooldridge 2013).

The standard errors of MODEL 1 to MODEL 3 are weighted using SIPP’s probability sampling weights and replicate weights. The two-step Heckman correction procedure is incompatible with weighting, but with the assumption of normally distributed errors, the MLE used by MODEL 3 can perform the standard correction in one step (Wooldridge 2002). Heckman-corrected probit models are referred to as "Heckit" models in the econometrics literature.

### 3.3 Fixed Effects Linear Probability Model

MODEL 3 does not take full advantage of the SIPP panel data. Given repeated observations for individuals over time, a fixed effects model can control for unobserved individual effects by including binary variables for each individual. Unobserved time effects can be controlled for by including binary variables for each time period. With individual effects, a regression model controls for all time-invariant individual characteristics. By including time fixed effects, a model can control for all characteristics that vary over time but are fixed across individuals at a given time.
Table 3.1: Unweighted correlations between tax control variables and dependent and key independent variable

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>Filed Return $r$ (s.e.)</th>
<th>EITC Size ($) $r$ (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Federal Filing Requirement</td>
<td>-0.122 (0.000)</td>
<td>0.203 (0.000)</td>
</tr>
<tr>
<td>EITC*No Filing Requirement</td>
<td>0.087 (0.000)</td>
<td>0.999 (0.000)</td>
</tr>
<tr>
<td>Additional Child Tax Credit</td>
<td>0.048 (0.000)</td>
<td>0.620 (0.000)</td>
</tr>
<tr>
<td>EITC*Child Tax Credit</td>
<td>0.042 (0.000)</td>
<td>0.318 (0.000)</td>
</tr>
<tr>
<td>Federal Income Tax Liability</td>
<td>-0.038 (0.000)</td>
<td>0.231 (0.000)</td>
</tr>
<tr>
<td>Federal Adjusted Gross Income</td>
<td>0.034 (0.000)</td>
<td>-0.026 (0.000)</td>
</tr>
<tr>
<td>IRS Audit Rate</td>
<td>-0.162 (0.000)</td>
<td>-0.247 (0.000)</td>
</tr>
<tr>
<td>State EITC</td>
<td>0.037 (0.000)</td>
<td>0.457 (0.000)</td>
</tr>
<tr>
<td>No State Filing Requirement</td>
<td>-0.130 (0.000)</td>
<td>0.019 (0.000)</td>
</tr>
<tr>
<td>Total State Tax Credits</td>
<td>0.015 (0.000)</td>
<td>0.286 (0.000)</td>
</tr>
<tr>
<td>State Income Tax Liability</td>
<td>-0.001 (0.806)</td>
<td>-0.108 (0.000)</td>
</tr>
</tbody>
</table>

Notes: $r$ is a Pearson’s or a point-biserial correlation coefficient, depending on the variable type. Unweighted standard errors in parentheses.
Fixed effects models therefore have the advantage of controlling for unobserved heterogeneity (Angrist & Pischke 2008). As mentioned in subsection 3.1, omitting controls for unobserved heterogeneity likely biases models of the relationship between EITC benefits and filing propensity.

MODEL 4 is a fixed effects linear probability model (LPM) of the relationship between filing propensity and EITC size. It estimates the equation:

\[ \text{Filed}_{it} = \alpha_i + \text{EITC}_{it}' \beta_1 + X'_{\text{tax},it} \beta_2 + X'_{\text{dem},it} + \lambda_t + \varepsilon_{it}, \]  

(3.5)

where subscripts \( i \) and \( t \) denote individual and time indices and \( \lambda \) is the year fixed effect. The regression is run on the subsample of 2007 EITC nonfilers, but with the full two years of panel data, rather than the 2009 cross-section in MODEL 3.

MODEL 4 has four notable disadvantages. The first two were discussed above: LPMs can generate fitted values outside the range of the binary dependent variable and assume constant marginal effects. The third disadvantage is that, though the fixed effects LPM in equation 3.5 can incorporate probability sampling weights, it cannot use replicate weights that adjust for the SIPP’s complex survey design. Its standard error estimates are therefore likely biased. Last, fixed effects are incompatible with the Heckman correction method. Because MODEL 4 is run on a truncated panel of 2007 nonfilers, its estimates are biased by sample selection.

The disadvantage of linearity could be avoided by using a fixed effects logit model, which is nonlinear.\(^1\) However, a fixed effects logit model has a critical disadvantage: it cannot incorporate either sampling or replicate weights. This means both its point and standard error estimates would be biased. Analysis of SIPP data, at the least, requires use of sampling weights.

\(^1\)Probit models can only take on a random effects specification that uses generalized least squares. The random effects model assumes that the unobserved effects and the model covariates are conditionally independent, which is not a reasonable assumption for the current model. See subsection 3.1.
It is worth noting that, as Angrist & Pischke (2008) argue, LPM coefficients and probit marginal effects tend, as an empirical matter, to be similar across the non-linear function. Though not a formal theorem, they argue this is “fairly robustly true.” In this vein, the approach taken in this study is to run linear fixed effects in Model 4 and compare point estimates and statistical significance with the cross-sectional Heckit Model 3.

3.4 Generalized Ordered Logit

Model 5 estimates the first-difference ordered logit model:

$$\ln\left(\frac{\Delta filed}{1 - \Delta filed}\right) = \Delta \delta_0 + \Delta EITC_i \beta_1 + \Delta X_{tax,i} \beta_2 + \Delta X_{dem,i} \beta_3 + \Delta u_i, \quad (3.6)$$

where each variable and vector is the difference between its 2009 and 2008 values. With this approach, Model 5 controls for individual-specific time-invariant and individual-invariant time-specific effects, because factors that are constant across 2008 and 2009 are "differenced away" in equation 3.6 (Wooldridge 2013). The model is effectively run on a 2009 cross-section, but it takes advantage of the panel data by controlling for individual and time fixed effects.

The SIPP panel is well-suited to a first-difference model because it tracks EITC nonfilers before and after the ARRA EITC expansion. A general condition for first-differences is that $\Delta x_i$ must have some variation across individuals. If the key independent variable does not change over time, the model cannot pick out and separate the invariant effects’ relationships to the dependent and key independent variable (Wooldridge 2013). With the ARRA EITC expansion, Model 5 has the variation needed to estimate first-differences.
The dependent variable $\Delta filed$ takes on the values:

$$
\Delta filed = \begin{cases} 
1, & \text{if filed in both years,} \\
0, & \text{if filed in one year,} \\
-1, & \text{if filed in neither year.}
\end{cases}
$$

(3.7)

This variable is coded by keeping $Filed = 1$ for filers in 2009, but recoding $Filed = 0$ for 2007 filers, so that after first-differences, both-year filers are coded $1 - 0 = 1$, both-year nonfilers are coded $0 - 1 = -1$, and single-year filers are coded $0 - 0 = 1 - 1 = 0$.\(^2\)

MODEL 5 has all four advantages discussed in subsection 3.1. Unlike the fixed effects MODEL 4, the first-difference MODEL 5 can incorporate SIPP probability and replicate weights. First-differences also control for unobserved heterogeneity. Further, the model takes a nonlinear function form. Last, fixed effects logit models estimate the conditional probabilities such as $P(\text{Filed } 07&09 | \Delta EITC)$, rather than the constant and average marginal effects of LPMs and probits. The model can therefore be run on the full analytic sample, and by conditioning on EITC eligibility and below-threshold status, can predict filing probabilities for the relevant population without truncating the sample.

MODEL 5 is technically a generalized ordered logit model. A test of the proportional odds assumption for ordered logit rejected the null hypothesis of no statistical difference between model coefficients ($p < 0.00$). The generalized ordered logit model in this study does not make the proportional odds assumption.

\(^2\)This coding strategy cannot be used for nonfilers, because the "0" they have for filing status in at least one of the years means (a), the subtraction of the 2009 from the 2008 dependent variable does not result in enough dependent variable variation for first-differences to control for fixed effects, and (b) it is not possible to order the dependent variable so that higher categorical values are actually "higher" than the lower values. This is why MODEL 5 can only be used for EITC eligible individuals below the filing threshold.
Section 4

Results

4.1 2007 EITC Eligible Nonfilers

Column (1) of Table 4.1 reports the average marginal effects of Model 1, a bivariate probit regression of filing propensity on EITC benefit size among 2007 EITC eligible nonfilers. On average, a $100 increase in EITC is associated with a 0.5 percent increase in the probability that a 2007 EITC nonfiler filed a tax return in 2009 ($p < 0.00$).

Column (2) provides the average marginal effects for Model 2, which adds in the full set of tax and demographic covariates. Controlling for these observed confounders, a $100 increase in EITC is associated on average with a 5.8 percent increase in the probability that a 2007 EITC nonfiler filed in 2009 ($p < 0.00$).

Column (3) reports the results for Model 3, a Heckit with full tax and demographic controls. Adjusting for sample selection, on average a $100 increase in EITC is associated with a 5.9 percent increase in the probability that a 2007 EITC nonfiler filed in 2009 ($p < 0.00$).

Column (4) of Table 4.1 provides the estimates of Model 4, a fixed effects linear probability model regressing filing propensity on EITC benefits for 2007 eligible nonfilers. Controlling for the set of observed tax and demographic variables, as well as unobserved individual and year fixed effects, a $100 increase in EITC is associated with a 5.1 percent increase in the probability of filing a return ($p < 0.00$).
Table 4.1: 2009 filing propensity of 2007 EITC eligible nonfilers: average marginal effects of cross-sectional probit models and estimates of fixed effects linear probability model

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>Probit (1)</th>
<th>(2)</th>
<th>(3)</th>
<th>LPM (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Controls</td>
<td>Full Controls</td>
<td>Heckman</td>
<td>Fixed Effects</td>
</tr>
<tr>
<td></td>
<td>Avg. Mfx (s.e.)</td>
<td>Avg. Mfx (s.e.)</td>
<td>Avg. Mfx (s.e.)</td>
<td>( \beta ) (s.e.)</td>
</tr>
<tr>
<td>EITC</td>
<td>0.005*** 0.000</td>
<td>0.058*** 0.014</td>
<td>0.059*** 0.014</td>
<td>0.051*** 0.012</td>
</tr>
<tr>
<td>No Federal Filing Requirement</td>
<td>0.172*** 0.041</td>
<td>0.169*** 0.043</td>
<td>0.147*** 0.036</td>
<td></td>
</tr>
<tr>
<td>EITC*No Filing Requirement</td>
<td>-0.050*** 0.014</td>
<td>-0.051*** 0.014</td>
<td>-0.046*** 0.012</td>
<td></td>
</tr>
<tr>
<td>Additional Child Tax Credit</td>
<td>-0.017*** 0.004</td>
<td>-0.017*** 0.004</td>
<td>-0.010*** 0.002</td>
<td></td>
</tr>
<tr>
<td>EITC*Child Tax Credit</td>
<td>-0.001*** 0.000</td>
<td>-0.001*** 0.000</td>
<td>-0.000 0.000</td>
<td></td>
</tr>
<tr>
<td>Federal Income Tax Liability</td>
<td>-0.002*** 0.001</td>
<td>-0.002*** 0.001</td>
<td>-0.001 0.001</td>
<td></td>
</tr>
<tr>
<td>Federal AGI</td>
<td>0.000*** 0.000</td>
<td>0.000*** 0.000</td>
<td>0.000*** 0.000</td>
<td></td>
</tr>
<tr>
<td>IRS Audit Rate</td>
<td>0.019* 0.011</td>
<td>0.017** 0.009</td>
<td>-0.011 0.009</td>
<td></td>
</tr>
<tr>
<td>State EITC</td>
<td>-0.005 0.006</td>
<td>-0.004 0.005</td>
<td>-0.002 0.006</td>
<td></td>
</tr>
<tr>
<td>No State Filing Requirement</td>
<td>-0.012 0.036</td>
<td>-0.001 0.025</td>
<td>0.043 0.027</td>
<td></td>
</tr>
<tr>
<td>Total State Tax Credits</td>
<td>0.003 0.005</td>
<td>0.004 0.004</td>
<td>0.006 0.005</td>
<td></td>
</tr>
<tr>
<td>State Income Tax Liability</td>
<td>-0.002 0.002</td>
<td>-0.002 0.002</td>
<td>0.000 0.003</td>
<td></td>
</tr>
<tr>
<td>Fixed effects</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Weighted s.e.</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Observations</td>
<td>2,362</td>
<td>2,242</td>
<td>2,816</td>
<td>5,925</td>
</tr>
<tr>
<td>Number of id</td>
<td></td>
<td></td>
<td></td>
<td>3,724</td>
</tr>
</tbody>
</table>

Notes: Columns (1)-(3) report average marginal effects and delta-method standard errors. Column (4) reports heteroskedasticity robust standard errors. All tax variables are scaled to hundreds of dollars. Unlisted controls in columns (1)-(4) are metro status, age-squared, family household status, female headed household status, marital status, and number of children under 18. Columns (1)-(3) also include control variables for white race, Hispanic ethnicity, age, and high school graduate status.

*** \( p < 0.01 \), ** \( p < 0.05 \), * \( p < 0.1 \)
Figure 4.1: Comparison of cross-sectional Heckit (Model 3) and fixed effects LPM (Model 4) coefficient and confidence interval estimates for EITC size and select tax variables.

Figure 4.1 compares the coefficient estimates and confidence intervals of Model 3 and Model 4 for the EITC and select tax control variables. The overlapping 95% confidence intervals show that many of the point estimates of the two models are not statistically different.

In Figure 4.2, individual EITC eligibility in dollar terms is plotted against the probability of filing predicted by each model. Though the average marginal effects of the two models are close, the graph shows that the Heckit model exhibits different
marginal effects across the EITC distribution that are not present in the constant effects LPM.

Table A.2 in the Appendix reports the marginal effects of MODEL 3 at the 25th, 50th, and 75th percentiles of the distribution of EITC benefits among 2007 nonfilers. Controlling for the set of covariates, a $100 increase in EITC is associated with a filing propensity increase of 2.7 percent at the 25th percentile, 4.2 percent increase at the 50th percentile, and 1.1 percent increase at the 75th percentile. As predicted by economic theory, the marginal effects are not constant across the EITC distribution. All three estimates are statistically significant ($p < 0.00$).
4.2 EITC Eligible Individuals Without a Filing Requirement

Column (1) of Table 4.2 reports the predicted conditional probabilities of Model 5, a first-difference generalized ordered logit model of the relationship between EITC benefit changes and filing propensity among EITC eligible individuals below the filing threshold. Column (1) does not use the set of tax and demographic controls, but does control for unobserved heterogeneity. For successive increments of $100 EITC increases between 2008 and 2009 ranging from $0 to $500, the change in EITC is associated on average with a 0.7 percentage point increase in the probability of filing in both 2008 and 2009 ($p < 0.00$).

Column (2) of Table 4.2 incorporates the first-difference estimators of the set of tax and demographic controls. Conditioning on these controls set at their means, a $100 increase in EITC within the $0 to $500 range is associated on average with a 0.6 percentage point increase in the probability of filing in both years ($p < 0.00$).

Column (1) of Table 4.3 reports the predicted conditional probabilities for not filing in both 2007 and 2009, given specific levels of EITC change. Column (1) does not incorporate controls. Controlling for unobserved heterogeneity, within the $0 to $500 increase range, a $100 EITC increase is associated on average with a 0.4 percentage point decrease in the probability of not filing in both years ($p < 0.00$).

Column (2) of Table 4.3 controls for the set of tax and demographic factors using first-difference estimators. Controlling for these covariates set at their means, a $100 increase in EITC within the $0 to $500 range is associated with a 0.3 percentage point decrease in the probability of not filing in both 2007 and 2009 ($p < 0.00$).

Appendix Table A.3 reports the results of a fixed effects linear probability model run on a truncated sample of EITC eligible individuals without filing requirements. Controlling for the observed set of tax and demographic covariates and for unobs-
served heterogeneity, a $100 increase in EITC is associated with a 4 percent increase in filing propensity. Notably, this is roughly 1 to 2 percentage points smaller than the estimates of \textsc{Model 3} and \textsc{Model 4}.

The LPM estimate is also notably larger than the predicted conditional probabilities of \textsc{Model 5}. Even if the 2007 and 2009 filing decisions are conditionally independent, which is highly unlikely, the LPM would estimate that a $100 increase is associated with a $0.04 \times 0.04 = 0.016 \approx 1.6\%$ increase in filing propensity among EITC eligible individuals below the filing threshold, throughout the distribution of EITC benefit changes. Compare this to the \textsc{Model 5} results showing a 0.6 percent increase in persistent filing and a 0.3 percent decrease in persistent nonfiling is associated with a $100$ EITC increase.
Table 4.2: First-difference generalized ordered logit model of the relationship between change in EITC and predicted probability of filing in both 2007 and 2009 among EITC eligible individuals without a filing requirement

<table>
<thead>
<tr>
<th>EITC Change 2007 to 2009</th>
<th>(1) No Controls</th>
<th>(2) Full Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>$00</td>
<td>0.315*** (0.000)</td>
<td>—</td>
</tr>
<tr>
<td>$100</td>
<td>0.322*** (0.000)</td>
<td>0.7</td>
</tr>
<tr>
<td>$200</td>
<td>0.329*** (0.000)</td>
<td>0.7</td>
</tr>
<tr>
<td>$300</td>
<td>0.337*** (0.000)</td>
<td>0.8</td>
</tr>
<tr>
<td>$400</td>
<td>0.344*** (0.000)</td>
<td>0.7</td>
</tr>
<tr>
<td>$500</td>
<td>0.351*** (0.000)</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Fixed effects: yes  yes
Weighted s.e.: yes  yes
Observations: 37,690  37,690
Prob > F: 0.0000  0.0000

Notes: The left column provides the level EITC of change from 2007 to 2009. Column (1) reports $P(\text{Filed 07\&09} \mid \Delta EITC)$, the predicted probability of filing in both years conditional on a given EITC level of change, without controls. Column (2) includes as controls the first-difference estimators for EITC eligibility, ACTC, EITC*CTC, federal AGI, federal income tax liability, not having a federal filing requirement, state EITC, total state tax credits, state income tax liability, not having a state filing requirement, the IRS audit rate, marital status, and number of children under 18. To the right of columns (1) and (2) are the changes in probability percentage points from the previous level of EITC change.

*** p<0.01
Table 4.3: First-difference generalized ordered logit model of the relationship between change in EITC and predicted probability of *not filing* in both 2007 and 2009 among EITC eligible individuals without a filing requirement

<table>
<thead>
<tr>
<th>EITC Change 2007 to 2009</th>
<th>(1) No Controls</th>
<th>(2) Full controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0</td>
<td>0.215*** (0.000)</td>
<td>—</td>
</tr>
<tr>
<td>$100</td>
<td>0.211*** (0.000)</td>
<td>-0.4</td>
</tr>
<tr>
<td>$200</td>
<td>0.207*** (0.000)</td>
<td>-0.4</td>
</tr>
<tr>
<td>$300</td>
<td>0.203*** (0.000)</td>
<td>-0.4</td>
</tr>
<tr>
<td>$400</td>
<td>0.200*** (0.000)</td>
<td>-0.3</td>
</tr>
<tr>
<td>$500</td>
<td>0.196*** (0.000)</td>
<td>-0.4</td>
</tr>
</tbody>
</table>

Fixed effects: yes
Weighted s.e.: yes
Observations: 37,690
Prob > F: 0.0000

Notes: The left column provides the level of change in EITC from 2007 to 2009. Column (1) reports \( P(\text{NotFiled 07&09} | \Delta \text{EITC}) \), the predicted probability of not filing in both years conditional on a given level of EITC change, without controls. Column (2) includes as controls the first-difference estimators for EITC eligibility, ACTC, EITC*CTC, federal AGI, federal tax liability, not having a federal filing requirement, state EITC, total state tax credits, state tax liability, not having a state filing requirement, the IRS audit rate, marital status, and number of children under 18. To the right of columns (1) and (2) are the changes in probability percentage points from the previous level of EITC change.

*** p<0.01
Section 5

Discussion

5.1 Policy Implications

The marginal effects estimated by the Heckit model and fixed effects LPM are similar, despite the differences in specification. Together, the models suggest that the filing behavior of 2007 EITC eligible nonfilers are responsive to increases in EITC benefits. The fixed effects LPM estimated for EITC eligible individuals without a filing requirement also indicates an economically significant relationship between filing propensity and EITC benefit size. These estimates of a 4 to 6 percent increase in filing propensity for a $100 EITC increase are larger than the estimates in Scholz (1994) and Blumenthal et al. (2005), which were 2.6 and 2.8 percent, respectively.

The first-difference logit model estimated a much weaker relationship than the other models. One possible explanation for this could be that filing propensity is less responsive at the levels of EITC change conditioned on in the model’s probability predictions. The filing behavior of eligible individuals may be less responsive in the $0 to $500 range of EITC increases than in other parts of the EITC distribution. The Heckit estimates from Appendix Table A.2 support this explanation. Still, future EITC expansions are most likely to increase benefits within the $0 to $500 range. The ARRA, for example, increased the credit cap for filers with three or more children by $629. Regardless of whether EITC eligible individuals have larger marginal filing
responses further out in the distribution of EITC changes, as a practical policy matter these results are not particularly significant.

The most likely explanation of the divergent logit estimates is that none of the other models incorporate weighted errors, fixed effects, nonlinear functional form, and selection correction. Model 5 is the most conservative of the models considered in this study. The estimates it provides of the relationship between EITC benefit size and filing propensity among eligible individuals are the least likely to be biased. The logit model also estimates the relationship between EITC benefit size and persistent filing and nonfiling across two years, which means its results are least likely to be inflated by one-off instances of filing by otherwise consistent nonfilers.

Strictly in EITC participation terms, the opportunity cost of an increase in EITC benefits is a foregone increase in outreach to likely nonparticipants. As noted in subsection 1.1, recent randomized controlled trials administered by IRS researchers indicate that outreach to likely EITC eligible nonfilers with inexpensive mailers increases filing by 0.5 to 1 percent (Guyton et al. 2016). Comparing those results to the results of this study suggests that EITC increases are a far less cost-effective way to decrease the EITC eligible nonfiler rate than outreach.

Of course, policymakers may be motivated to increase EITC for reasons orthogonal to EITC participation. Despite their different point estimates, each model in this study found a positive and statistically significant relationship between EITC benefit size and the probability of filing a federal tax return. If EITC expansions do indeed increase filing among EITC eligible individuals, tax policymakers considering EITC increases should count greater participation among the potential benefits, in addition to and in interaction with any effects EITC has on poverty rates and labor supply.
5.2 Study Limitations and Extensions

This study has limitations of two kinds. The first set concern the SIPP in general, while the second set concern issues specific to the filing and EITC variables used in the analysis.

5.2.1 SIPP-related Limitations

The first study limitation is that individuals do not accurately report income in the SIPP. Maynard & Dollins (2002) note that the 1996 SIPP underestimated overall income. Though wage, salary, and social security income are reported fairly accurately in the 1996 survey, self-employment income, interest income, and dividend income are under-reported. If some individuals who appear EITC eligible in the 2008 SIPP are actually ineligible due to unreported income, the results of this study would be biased. Blumenthal et al. (2005) and Scholz (1994) both discuss this source of bias in estimating the relationship between EITC benefit size and filing propensity. Blumenthal et al. (2005) avoided the problem by using administrative data; but administrative data lack observations of nonfilers. Scholz (1994) skirted the issue by matching SIPP survey data to IRS administrative records. Unfortunately, securing IRS cooperation in linking SIPP to administrative data lies beyond the scope of the present author’s influence. As such, this stands as a potential source of bias in the study.

A second limitation stems from this study’s estimation of annual income based on a single month’s income. Monthly income is volatile and a given month’s income does not necessarily represent a typical month’s income. If individuals do not earn in a particular month but do work throughout the year, this approach could count EITC
eligible individuals as having no income, and therefore ineligible for EITC. If individuals receive a lower amount of income than typical in a given month, this approach could count EITC ineligible individuals as being low-income enough to qualify for the credit. An alternative to this approach would be averaging across the months of data available for a given year, and then annualizing on that basis. This is the approach taken by Rothstein & Valletta (2014). Rather than dropping Waves 2 and 8, these data could be kept in the analytic sample and used in estimating income. The relative advantage of the approach taken in the study is outlined in subsection 2.2.

This study also did not use SIPP longitudinal weights. These weights make non-response adjustments to compensate for attrition from the panel, as well as post-stratification adjustments that make weighted sample totals mirror the U.S. population for key variables in the data set (Census Bureau 2009). Because the analytic sample is not longitudinally weighted, the estimates of this study are likely biased by non-random attrition. One way to extend this study would be to incorporate longitudinal weights.

5.2.2 Tax-related Limitations

The ideal unit of analysis for this study is the tax filing unit, rather than the individual. But the units of analysis in the SIPP are individuals and households. Sometimes individuals more closely approximate the tax filing unit, and sometimes households do. Following the EITC nonfiler literature, this study opted to use individuals as the unit of analysis. An extension of this study could run the models on a household sample and compare the results with those from the individual sample.

The filing dependent variable has many missing values, mostly due to non-response in the tax topical module questions. In the Wave 5 and 8 modules, for
example, 33 to 34 percent of individuals didn’t answer the 2009 return filing question. In the analytic data set, 39.87 percent of observations are missing the 2007 filing variable and 36.97 of observations are missing the 2009 filing variable. If the non-response to these questions was not random, then these missing values bias the model estimates. A possible way to expand this study is to impute missing values for these individuals.

Another issue relates to the construction of the 2007 filing variable with using the stimulus rebate variable. 2007 filing status cannot be determined for rebate ineligible but EITC eligible individuals. About 4,000 individuals in the analytic data set had AGI between $0 and $3,000. Excluding these individuals from the analysis could bias model estimates if extremely low-income individuals have a different response to EITC filing incentives. This issue is especially acute considering that half of EITC eligible nonfilers had AGI under $8,000 in 2005 (Plueger 2005) and that, in the analytic sample, EITC nonfilers had average AGI of $7,085 in 2008 and $4,089 in 2009.

A baseline of comparison throughout this study was outreach efforts targeting EITC eligible nonfilers. However, receiving EITC outreach is associated both with EITC eligibility and the probability of filing a return. Yet the impacts of outreach efforts are not controlled for in any of the models. It is hard to imagine what data could be used to this end, but this study could perhaps be extended by discovering an instrument for outreach.

One source of potential endogeneity in this study is EITC work incentives. Depending on substitution and income effects, EITC can provide a work incentive during its phase-in by increasing subsidies as earnings increase. These incentives could lead individuals to increase their incomes over the filing threshold, which could increase their filing propensity. An estimate of how EITC functions as a filing
incentive separately from how it functions as a work incentive must control for the self-selection of individuals into the group with a filing requirement. One way to improve this study would be to estimate a selection correction equation like a Heckman model that could control for EITC eligible individuals working themselves into having a filing requirement.
Table A.1: Descriptive statistics for EITC eligible individuals without a federal filing requirement in 2008 and 2009

<table>
<thead>
<tr>
<th>Year</th>
<th>Eligible Individuals Without Filing Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2008</td>
</tr>
<tr>
<td>Observations</td>
<td>73,990</td>
</tr>
</tbody>
</table>

Binary Variables

<table>
<thead>
<tr>
<th></th>
<th>Count (%)</th>
<th>Count (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EITC Eligible</td>
<td>15906 (22.07)</td>
<td>12410 (24.35)</td>
</tr>
<tr>
<td>No Fed. Filing Requirement</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ACTC Eligible</td>
<td>8200 (11.23)</td>
<td>7183 (13.92)</td>
</tr>
<tr>
<td>State EITC Eligible</td>
<td>5496 (6.51)</td>
<td>5120 (9.21)</td>
</tr>
<tr>
<td>No State Filing Requirement</td>
<td>68623 (92.73)</td>
<td>46151 (89.22)</td>
</tr>
<tr>
<td>Metropolitan Resident</td>
<td>56250 (82.21)</td>
<td>39389 (81.70)</td>
</tr>
<tr>
<td>White</td>
<td>56744 (78.18)</td>
<td>40178 (78.85)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>11609 (18.19)</td>
<td>7987 (17.96)</td>
</tr>
<tr>
<td>Male</td>
<td>32666 (44.75)</td>
<td>23248 (45.43)</td>
</tr>
<tr>
<td>High School Graduate</td>
<td>38066 (52.18)</td>
<td>29230 (56.35)</td>
</tr>
<tr>
<td>Family Household</td>
<td>65450 (88.76)</td>
<td>45643 (88.48)</td>
</tr>
<tr>
<td>Female Headed Household</td>
<td>18681 (24.59)</td>
<td>12902 (24.36)</td>
</tr>
<tr>
<td>Married</td>
<td>26352 (35.69)</td>
<td>19364 (37.03)</td>
</tr>
</tbody>
</table>

Continuous Variables

<table>
<thead>
<tr>
<th></th>
<th>Mean ± Std. Dev.</th>
<th>Mean ± Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>EITC ($)</td>
<td>570.62 ± 1317.31</td>
<td>417.85 ± 1052.34</td>
</tr>
<tr>
<td>ACTC ($)</td>
<td>178.69 ± 545.09</td>
<td>112.6 ± 407.570</td>
</tr>
<tr>
<td>State EITC ($)</td>
<td>37.15 ± 190.77</td>
<td>23.51 ± 145.44</td>
</tr>
<tr>
<td>Fed. Inc. Tax Liability ($)</td>
<td>-882.50 ± 1932.75</td>
<td>-527.35 ± 1307.82</td>
</tr>
<tr>
<td>State Inc. Tax Liability ($)</td>
<td>-6.33 ± 351.36</td>
<td>-4.37 ± 225.17</td>
</tr>
<tr>
<td>Federal AGI ($)</td>
<td>7085.15 ± 13390.14</td>
<td>4089.61 ± 8997.74</td>
</tr>
<tr>
<td>Age</td>
<td>34.02 ± 23.96</td>
<td>33.73 ± 24.91</td>
</tr>
<tr>
<td>Children under 18</td>
<td>1.29 ± 1.38</td>
<td>1.40 ± 1.44</td>
</tr>
</tbody>
</table>
Table A.2: 2009 filing propensity of 2007 EITC eligible nonfilers: marginal effects of cross-sectional Heckit model across the EITC distribution

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1) 25th EITC Percentile Marginal Effect (s.e.)</th>
<th>(2) 50th EITC Percentile Marginal Effect (s.e.)</th>
<th>(3) 75th EITC Percentile Marginal Effect (s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EITC</td>
<td>0.027*** (0.005)</td>
<td>0.042*** (0.012)</td>
<td>0.011*** (0.001)</td>
</tr>
<tr>
<td>Observations</td>
<td>2,816</td>
<td>2,816</td>
<td>2,816</td>
</tr>
</tbody>
</table>

Notes: Columns (1)-(3) report marginal effects at the 25th, 50th, and 75th percentiles of the EITC benefits distribution. Delta-method standard errors reported in parentheses. EITC is scaled to hundreds of dollars. Unlisted controls in columns (1)-(3) are metro status, white race, Hispanic ethnicity, age, age-squared, family household status, female headed household status, high school graduate status, marital status, and number of children under 18.

*** p<0.01, ** p<0.05, * p<0.1
Table A.3: Fixed effects linear probability model of the relationship between EITC size and filing propensity among EITC eligible individuals without a filing requirement

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>(1)</th>
<th>β</th>
<th>(s.e.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EITC</td>
<td>0.040***</td>
<td>(0.005)</td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.407</td>
<td>(0.305)</td>
<td></td>
</tr>
</tbody>
</table>

Fixed effects yes
Weighted s.e. no

Observations 20,938
Number of id 16,821
$R^2$ 0.164

Notes: Heteroskedasticity robust standard errors in parentheses. Unlisted controls include ACTC, EITC*CTC, federal AGI, federal income tax liability, the IRS audit rate, state EITC size, not having a state filing requirement, total state tax credits, state income tax liability, metro status, age-squared, family household status, female headed household status, marital status, and number of children under 18.

*** p<0.01, ** p<0.05, * p<0.1


Lim, Younghee, Michelle Livermore, and Belinda Creel Davis. 2011. "Tax Filing and Other Financial Behaviors of EITC-Eligible Households: Differences


