Predicting Intentional Tax Error Using Open Source Literature and Data

Ju-Sung Lee and Kathleen M. Carley

November 5, 2009
CMU-ISR-09-125

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Center for the Computational Analysis of Social and Organizational Systems

CASOS technical report

This work was supported by the Internal Revenue Service (IRS) and the National Science Foundation (NSF) IGERT Grant.

Keywords: tax evasion, non-compliance, intentional error, meta analysis

Abstract

Intentional non-compliance in providing accurate income tax returns, also known as 'tax evasion' or 'intentional error', has been studied from both attitudinal and socio-demographic perspectives. A significant portion of previous research employs a common set of indicators, which we can exploit by pooling meta-analytically with the hopes of obtaining a unified, well-predicting model of intentional error. Towards this end, we turn to a large, nationally representative data source, namely the Census Bureau's Public-Use Microdata Samples (PUMS), as our source of covariance between the socio-demographic covariates of interest. Additionally, the same source offers data on potential opportunities of evasion for each PUMS respondent (or agent), in certain line item/taxpayer categories, allowing us to construct distinct error models for these categories. Furthermore, we extend the error model to include attitudinal meta-analysis, by linking the General Social Survey (GSS) to the PUMS through imputation of a GSS covariate that identifies respondents who are more likely to break the law. Our meta-analysis requires an in-depth re-analysis of the selection of previously published results on non-compliance. The result is a comprehensive model of non-compliance that fits historical, published data and that can be applied generically and to specific tax issues.

Contents

1 Introduction

Taxation is one of the oldest and most common collective action dilemmas in human history. Most of us do not enjoy paying taxes and, if given the choice to forego paying (legally) for, say, the past calendar year, most of us would gladly embrace the boon. The burden of the taxpayer, which is to relinquish a portion of earned resources towards the efficacy of the state, or society, is so diffuse over a large population that the temptation to avoid paying — or, in game-theoretic terms, to 'free-ride' — is ever-present, especially in taxation systems that are partly or wholly voluntary such as the United States'. A lá the 'tragedy of commons' (Hardin, 1968), tax evasion, though illicit, is locally rational; that is, it is easy to understand how one might justify the behavior when the consequences of a single act is minimal. However, repeated acts and increasing prevalence (and increasing seriousness of the act) will render the behavior systemically irrational, as it can then significantly undermine the government's effectiveness and diminish the fidelity of society's infrastructure, an outcome unwanted by everyone, including the tax evaders.[1]

Over the past few decades, researchers have examined the variety of conditions — psychological motives and predispositions, structural or social influences, socio-demographic correlates, etc. — that might explain why (or at least predict when) individuals end up purposely evading, above and beyond the mere 'rational' excuse. That is, while the inclination to evade is a foregone conclusion for most, how and why the act materializes is not so trivially understood. Furthermore, there remains the minority of individuals for whom taxpaying is not only civic duty, but privilege as well.[2]

As a purported collective action dynamic, tax evasion ought to be regarded as one of a specific brand of 'free-riding', or 'defection' (another game-theoretic term), commonly known is 'criminal behavior', instead of being treated as a special behavior isolated to certain income classes and uniquely dealing with economic loss; that is to say, evasion, generally placed under the umbrella of 'white collar crime', should not be compartmentalized to a specific class or subgroup or even definition, such that it imposes the presumption that the underlying motives are esoteric (Hirschi and Gottfredson, 1987; Knight and Knight, 1992). In fact, Mason and Calvin (1978) observed that $\frac{1}{3}$ of their evading respondents would not be considered 'white collar', by any definition of the phrase.

As with most counter-cooperative behaviors, the norming of both evasion and compliance have been realized through social influence. At the interpersonal level, researchers consistently find peer influence to be associated with evasion (Collins et al., 1992; Elffers et al., 1987; Webley et al., 2001; Vogel, 1974). On the other hand, wider norming effects are evident in cohesive communities especially those rooted in religion, such as a church, which maintains a relatively high standard of ethical conduct. While Andreoni et al. (1998) outlines such a mechanism, Torgler (2006) and Grasmick et al. (1991) empirically find religiosity to be linked to tax compliance, specifically through the threat of public shaming, while Stack and Kposowa (2006) report mixed findings. Still, the threat of emotional pain through shaming

[1] We are not considering here those individuals who wish to effect a radical change in the structure of society.

[2] To recall the phrase, "buying a civilization."

has been verified to be a potent effector (or constraint) of behavior (Coricelli et al., 2007), and researchers have recently been formally modeling the social interaction effect on evasion (Cowell, 1992; Fortin et al., 2007).

Conversely, some evidence shows that, absent the imposition of a strong moral norm or community, tax evasion becomes generally regarded as only a minor offense to law and society, standing mid-way in a list of criminal offenses containing the most heinous (e.g. murder) to the most innocuous (e.g. jay-walking) (Burton et al., 2005), all this a natural consequence of diffused responsibility and part and parcel to collective dilemmas. That non-compliance is considered so light an offense allows any temptation or inclination to commit it to be susceptible to a myriad of external conditions, excuses, or rationales. For instance, disaffection with the tax system, or with the behavior of its authorities, and disbelief in its fairness are all well-observed antecedents to non-compliant behavior (Frey and Feld, 2002; Webley et al., 2001; Porcano, 1988).

Unlike the canonical, unrestricted 'commons', there does often, in most nation-states, exist systems and institutions for enforcing tax compliance, namely through the execution of punitive measures on the more egregious evaders. As such, any inclination to evade is countered by the fear of having one's non-compliant behavior being formally discovered through an audit (e.g. Mason and Calvin, 1978). Since the rate of audits tend to be quite low, their preventative effect is perceptual, often based on hearsay, well-publicized cases, and, of course, whether an evader was personally audited or not. Still, there seems to be enough accuracy in the perceived risk such that a change in actual audit rates can affect non-compliance rates (Klepper and Nagin, 1989; Dubin et al., 1990; Andreoni et al., 1998). Since successful detection is often dependent on the ability for tax authorities to observe or verify the act of non-compliance, a significant change in the manner of income acquisition, e.g. from unmatchable sources such as investments, can contribute to non-compliance (Bloomquist, 2004).

However, the human decision-making process is fickle, prone to both predictable and unpredictable irrationalities, often borne out of temporary or persistent emotive states. So, not surprisingly, the assessment of one's risk of being audited, despite actually being low, can become exaggerated as a response to salient or frequent samples (of auditing events), a dynamic espoused by prospect theory (Kahneman and Tversky, 1979) and confirmed by Alm et al. (1992). Also, in accordance to prospect theory, Robben et al. (1990) find that how the tax is framed (i.e refund vs. payment, or gain vs. loss) can determine the level of compliance. And, as a tempting risk, tax evasion is undoubtedly connected to opportunity; that is, individuals with more, and more conducive, opportunities are more likely evade (Wahlund, 1992; Collins et al., 1992; Porcano, 1988); this is obvious.

Also expected and believable is the existence of personality types that are prone to non-normative behaviors such as tax evasion. Both psychological and sociological literature point to fixed predispositions or personalities that promote the commission of non-normative and risk-seeking behaviors, such as tax evasion. These individuals are said to be tolerant of deviant and/or illicit behaviors in themselves and others (Elffers et al., 1987; Wahlund, 1992; Collins et al., 1992) and also, self-oriented, as a result of selfishness or competitiveness

(Elffers et al., 1987); this point is highlighted in the New York Times by Goleman (1988) who accuse evaders of being "selfish to the bottom line".

Despite the variety in motivations behind tax evasion, evaders tend to fall into certain socio-demographic categories; these patterns (summarized later in this paper) are more-or-less consistent across the reviewed literature and clearly point to associations between the socio-demographic type and latent factors behind tax compliant behavior, both external (such as income) and internal (such as personality). While many of the surveyed papers ignore the role of socio-demographic indicators, a sufficient number of investigators deem them important enough to warrant their inclusion in their statistical models. These papers offer the advantage of aligned comparison as well as meta-analysis; findings from the behavioral and attitudinal variables in most of the sources cannot be directly compared.

As such, we seek to infer a model of intentional non-compliance (i.e. intentional error) incorporating findings from many of the aforemented literature, which span several decades and survey both U.S. and non-U.S. populations. Our model makes predictions of error from gender, age, education, and income class, as these were the primary socio-demographic predictors examined in prior research. An extension of the model explores the effect of one particular, relevant attitudinal covariate: an individual's tolerance of law-breaking. Furthermore, the model is being expanded to include other socio-demographic indicators, such as marital status and self-employment, and also behavioral indicators, such as the use of a paid preparer, all of which have been demonstrated to be significantly predictive of non-compliance.

Our main approach comprises a meta-analytic pooling of results from past studies into a single, comprehensive logistic regression model, while concurrently matching those empirical results as best as possible. We employ a variety of numerical approaches including Newton-Raphson optimization as our primary instrument for fitting models to the multiple marginal reports of empirical error (i.e. the reported proportion of error from sub-populations, e.g. males or high school graduate). More complicated study results, arising from, in one case, a large regression model and, in another, a large structural equation model, require us to resort to more flexible, heuristic optimization, such as simulated annealing and other variants of gradient descent (stochastic and expectation maximization), in order to transform the empirical results into usable estimates.

Furthermore, our source of unbiased covariance among the predictor variables, necessary for the meta-analysis, is the U.S. Census Bureaus Public-Use Microdata Samples (PUMS), which also provides us with population samples of seven taxpayer categories of interest. From these samples, we generate separate line item/taxpayer category error models. Additionally, we refer to several data years of the General Social Survey (GSS) to supplement parts of our prediction, e.g. from tolerance of law-breaking, using weaker inference, like multiple imputation, due to the absence of these covariates in the PUMS.

2 Data Sources

2.1 Empirical Rates of Error

We survey a collection of studies, most of which employ socio-demographic indicators in predicting non-compliance. The empirical error rates, obtained from multiple countries and across several decades, exhibit considerable variance:

% Error	n	Source	Country	Notes
10%	284	Houston and Tran (2000)	Australia	
19.6%	188	Antonides and Robben (1995)	Netherlands	
22-25%	N/A	Collins et al. (1992)	United States	review of 18 surveys
23.7%	125	Porcano (1988)	United States	
24.8%	800	Mason and Calvin (1978)	United States	
25%	1797	Vogel (1974)	Sweden	
29.9%	125	Porcano (1988)	United States	
37%	1427	Wahlund (1992)	Sweden	
50%	240	Collins et al. (1992)	United States	

Naturally, these rates are dependent on kinds of non-compliance studied, the common ones being under-reporting of income, over-reporting of deductions, and non-filing; not all studies explore the same types. But, even when controlling for the type of error, there still remains a noticeable range, which has been acknowledged by others including Clotfelter (1983) who estimated a range of 20–58% of taxpayers who evade taxes.

2.2 Summary of Findings

The following table enumerates the covariates examined in the studies of interest; we incorporate most of these into our meta-analysis:

Source	Pop.	Age	Sex	Educ	Inc	S.E.	Occ	Prep	Net	Mar
Erard			?			✓		?		✓
Antonides and Robben	✓			✓		✓				✓
Porcano		✓	✓		✓	?	✓			✓
Wahlund	✓	✓			✓	✓	?			✓
Mason and Calvin	✓	?	✓		?		?			
Houston and Tran	✓	✓	✓	✓	✓	✓		✓		
Collins et al.	✓	✓	✓	✓	✓			✓		
Vogel	✓	✓	✓	✓	?	✓	?		✓	

where

✓ = study employs covariate — ? = Coarse inference possible
Pop. = Population Rate of Error — Net = Social Network Effect
S.E. = Self-Employment — Mar = Marital Status
Occ = Occupation Type/Prestige — Prep = Paid Preparer Use

For the most part, there is a high degree of consistency in these papers' findings, which we summarize here:

- Sex/Gender - males are consistently more likely to evade, purportedly due to self-confident, competitive, and anti-authoritarian attitudes.

- Age - older individuals are less likely to evade, partly due to increasing risk aversion.

- Education - education reduces propensity to evade; however a curvilinear effect has been detected in one study.

- Income - low income and very high income individuals more likely to evade, partly due to increased opportunities from either under-the-table income or investment-based income.

- Self-Employment - confers increased opportunity to evade.

- Professional Preparer Use - studies show higher non-compliance when professional preparer is employed.

- Social Network - there is a higher likelihood to evade when peers evade (measured probabilistically).

2.3 Public-Use Microdata Samples (PUMS)

For all of the error models in this paper, we require a basis of covariance between the key socio-demographic indicators. Since our source papers, at best, provide marginal rates, we resort to a separate, relatively unbiased source of data/covariance: the U.S. Census Bureau's Public-Use Microdata Samples, which contain the socio-demographic covariates of interest as well as additional indicators that allow us to infer certain taxpayer categories (e.g. tips, self-employment, farm ownership, etc.), thus providing us with a tax evasion opportunity structure.[3] Despite containing only individuals residing in housing units, we suspect this qualification adequately reflects the population sampled in our source papers. In order to compromise between computation time and generalizability from a robust (i.e. large) data set, we sample from the PUMS a nationally representative data set of size, $n_{\text{PUMS}} = 10{,}000$; in later sections, when n is unspecificed, we imply n_{PUMS}. Also, for generalizability and compatibility to the data in our source papers, we impose a discrete categorization, or 'bins', of the PUMS covariates:

[3]For further information, refer to the PUMS documentation (United States Census Bureau, 2003).

covariate	bin #	description
Sex	0	Female
	1	Male
Age	0	< 30
	1	30 – 60
	2	60+
Education	0	Less than High School
	1	At Least a High School Graduate
	2	Bacheler's Degree
	3	Post-Graduate
Income	0	No Earnings or Less than \$0
	1	$<$ \$15,000
	2	$<$ \$30,000
	3	$<$ \$50,000
	4	$<$ \$80,000
	5	$<$ \$120,000
	6	$\geq$ \$120,000

2.4 The General Social Survey (GSS)

The General Social Survey has been administered almost yearly since the 1970s and assesses various cultural, behavioral, attitudinal, and socio-demographic trends in the non-institutionalized, adult household population of the United States.[4] In several survey years, data was collected on taxpaying attitudes as well as another important indicator, attitudes towards obedience to the law. We will employ both of these covariates in constructing our error models.

3 Terminology and Notation

We offer a brief primer on the specific terms and various statistical and mathematical notation and symbols used throughout this paper:

- The term 'intentional error' is completely interchangeable with 'tax evasion'. Its use stems from the IRS' classification of incorrect portions of tax returns as being one of two type of errors: intentional or inadvertent. While we use the term 'non-compliance' also synonymously with 'intentional error', in other writings, it might be used to indicate either kind of error.

- A bold-typed variable, e.g. $\boldsymbol{X}$, or a set of values held within a set of parentheses, e.g. $(x_0, \ldots, x_n)$, denotes a vector of values.

[4] Additional information on the GSS maybe found at the website: http://www.norc.org/GSS+Website

- $\mathrm{N}(\mu, \sigma^2)$ denotes the Normal distribution (also known as a Gaussian) with parameters mean μ and variance σ^2.

- $\mathrm{logit}[x]$ is the log odds of x, $\log\left[\dfrac{x}{1-x}\right]$; conversely, $\mathrm{logit}^{-1}[x]$ represents the inverse-logit, $\dfrac{\exp(x)}{1+\exp(x)} = \dfrac{1}{1+e^{-x}}$

- L denotes the likelihood and $\mathcal{L}$ denotes the log of a likelihood, or log-likelihood.

- $X \sim Y$ can denote that random variable X is distributed as defined by distribution Y, or it can denote the likelihood (i.e. probability or density) of the value X being drawn from distribution Y, depending on the context. For example, $x \sim \mathrm{N}(\mu, \sigma^2)$ is equivalent to $\mathrm{N}(x|\mu, \sigma^2)$ or $\mathcal{L} = \log[\mathrm{N}(x|\mu, \sigma^2)]$. So, naturally, if we have a vector $\boldsymbol{X}$ and $\boldsymbol{X} \sim \mathrm{N}(\mu, \sigma^2)$ then the log-likelihood of the fit is $\mathcal{L} = \sum_i \log[\mathrm{N}(X_i|\mu, \sigma^2)]$.

- The significance stars in regression models follows the standard nomenclature:

$$p < \begin{cases} 0.001 & \text{if `***'} \\ 0.01 & \text{if `**'} \\ 0.05 & \text{if `*'} \\ 0.1 & \text{if `^'} \end{cases}$$

- σ denotes standard deviation.

- p is a probability, while ρ (rho) is a Pearson correlation.

- Non-italic sub/super-scripts is a variable label while italicized ones are variables; e.g. in $x_{i,tp}$, tp is a placeholder for one of eight taxpayer categories and $x_{i,tp}$ is flag of whether or not agent i qualifies for taxpayer category tp. n_i^{tp} indicates the number of taxpayer categories agent i falls under; here, 'tp' merely indicates that n is a count of something related to taxpayer categories.

- We often employ the indicator function $\mathcal{I}(\ldots)$ in which '...' denotes some true/false conditional:

$$\mathcal{I}(\ldots) = \begin{cases} 1 & \text{if } \ldots \text{ (i.e. conditional is true)} \\ 0 & \text{otherwise} \end{cases}$$

- We will, on occasion, denote a mean with a functional expression, $\mathrm{mean}(x)$, rather than the symbolic over-line, $\overline{x}$, especially when paired with the functional expression for standard deviation, $\mathrm{sd}(x)$. On occasion we will employ the symbolic functional expressions: $\mu(x)$ and $\sigma(x)$.

4 Meta-Analysis

In this paper, we attempt to infer a statistical model of intentional error by combining findings (notably marginal statistics) from similar, overlapping covariates found our source empirical papers. Generally, we will relegate the details for any intermediate analytical steps, particularly ones in which we align a paper's results to our desired model, to Appendix A.

Our primary model, which we will call the 'generic error model, is a single logistic regression model employing the socio-demographic covariates common to our studies. In particular, we seek a set of coefficients:

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$$

duly labeled

$$\boldsymbol{\beta} = (\beta_{\text{Intercept}}, \beta_{\text{Sex}}, \beta_{\text{Age}}, \beta_{\text{Education}}, \beta_{\text{Income}})$$

such that a logistic regression prediction will best fit whichever empirical data and findings we see fit to include in the meta-analytical model:

$$\text{logit}[p(y > 0)] = \beta_0 + \beta_1 x_{\text{Sex}} + \beta_2 x_{\text{Age}} + \beta_3 x_{\text{Educ}} + \beta_4 x_{\text{Inc}} \tag{1}$$

where

$$y = \text{the number of acts of intentional error by a single individual}$$

and so

$$p(y > 0) = \text{the probability of commission of at least one intentional error}$$

4.1 Fitting to Marginals

4.1.1 Simple Example

One of our main fitting strategies exploits the marginal statistics reported in most of the sources; that is, we seek a model that best concurrently fits these marginal statistics. For instance, our sources offer several different error commission rates by males. The marginal statistic is:

$$p_{\text{Male}} = p(y > 0 | x_{\text{Sex}} = 1), \text{ where } x_{\text{Sex}} = 1 \text{ is equivalent to } x_{\text{Sex}} = \text{Male}$$

and the data we have is:

p_{Male}	n_{Male}	Source
0.323	709	Vogel (1974)
0.067	171	Houston and Tran (2000)
0.293	377	Mason and Calvin (1978)

Naturally, we employ the normal approximation to the logit of the probability to find the maximum-likelihood estimate, $\hat{p}_{\text{Male}}$:

$$\mathcal{L} = \sum_{s \in \{\text{Vogel, Houston, Mason}\}} \log\left[\text{N}\left(\text{logit}[\hat{p}_{\text{Male}}] | \mu = \text{logit}[p_{s,\text{Male}}], \sigma^2 = \sigma_s^2\right)\right]$$

where the variance is the standard variance surrounding the logit of a probability:

$$\sigma_s^2 = \frac{1}{(n_s)(p_{s,\text{Male}})(1 - p_{s,\text{Male}})}$$

For this example, we will provide a bit of elaboration of the inner term for the first summand (i.e. Vogel), which offers the likelihood of our estimate for males $\hat{p}_{\text{Male}}$ fitting Vogel's marginal probability:

$$\begin{aligned}
&\text{N}(\text{logit}[\hat{p}_{\text{Male}}|\mu_{\text{Vogel}} = \text{logit}[0.323], \sigma^2_{\text{Vogel}} = [709 \cdot (1 - 0.323) \cdot 0.323]^{-1}) \\
&\quad = \frac{1}{\sigma_{\text{Vogel}}\sqrt{2\pi}} \exp\left\{-\frac{(\hat{p}_{\text{Male}} - \mu_{\text{Vogel}})^2}{2\sigma^2_{\text{Vogel}}}\right\} \\
&\quad = 61.85 \cdot \exp\left\{-\frac{(\hat{p}_{\text{Male}} - (-0.740))^2}{8.32 \times 10^{-5}}\right\}
\end{aligned}$$

While it is possible to algebraically obtain the maximum-likelihood estimate of $\hat{p}_{\text{Male}}$ for this simple example, we will need to resort to numerical optimatization methods when we later solve for multiple marginals and multiple β coefficients. So, here, we employ the Newton-Raphson optimization and obtain the following solution:

$$\begin{aligned}
\text{logit}[\hat{p}_{\text{Male}}] &= -0.868 \text{ and } \sigma_{\text{Male}} = 0.0640 \text{ (i.e. the standard error around logit)} \\
\hat{p}_{\text{Male}} &= \quad 0.295 \text{ and } \sigma_{\text{Male}} = 0.0133 \text{ (i.e. the standard error around prob.)}
\end{aligned}$$

The normal approximation to the logit of a binomial is just that, an approximation, and has its limitations, especially for probabilities close to the extremities (0 or 1) or for low sample sizes. In this example, the low probability of $p_{\text{Houston,Male}}$ results in a slight mismatch to the real solution which can be obtained in one of three ways:

method	p	σ	Notes
weighted binomial	0.2792	0.01265	
logistic regression	0.2778	0.01265	converted from logit: $\mu = -0.955, \sigma = 0.0629$
beta approximation	0.2792	0.01265	using Newton-Raphson
normal approximation	0.2956	0.01333	using Newton-Raphson

The normal approximation to the logit results in a $|0.2956 - 0.2792|/0.2792 = 0.059$ or 5.9% error. We expect that the error in our estimates will be lower due to larger sample sizes as well as estimating from probabilities higher than 10%.[5]

[5]The beta approximation is more accurate and will be employed in future versions of the model.

4.1.2 General Marginal Model

The general expression for the marginal likelihood of our covariates across all sources is as follows:

$$\text{logit}[\overline{q}_{jk}|x_j = k] \sim \text{N}(\mu = \text{logit}[p_{ijk}|x_j = k], \sigma^2 = [n_{ijk}p_{ijk}(1 - p_{ijk})]^{-1}) \tag{2}$$

where i indexes our sources that report some of the marginal intentional error rates and j indexes our covariates of interest, i.e.

$$i \in \{\text{Houston, Mason, Vogel}\}$$
$$j \in \{\text{Sex, Age, Education, Income}\}$$

and k indexes separate bins/categories for each covariate (wherever applicable),

$$k \in \begin{cases} \{0, 1\} & \text{if } j = \text{Sex} \\ \{0, 1, 2\} & \text{if } j = \text{Age} \\ \{0, 1, 2, 3\} & \text{if } j = \text{Education} \\ \{0, 1, 2, 3, 4, 5, 6\} & \text{if } j = \text{Income} \end{cases}$$

and

$$\begin{aligned} n_{ijk} &= \#\text{ of respondents falling in bin } k \text{ of covariate } j \text{ in source } i \\ p_{ijk} &= \text{reported intentional error probability for bin } k \text{ of covariate } j \text{ in source } i \\ \overline{q}_{jk} &= \text{average predicted/fitted probability of error for bin } k \text{ of covariate } j \\ x_j &= \text{refers to the values of covariate } k \text{ of our PUMS sample} \end{aligned}$$

Finally, we have our predicted mean marginal probability for the covariate category:

$$\overline{q}_{jk} = \frac{\sum_{i=1}^{i=n} \mathcal{I}(x_{ij} = k) \cdot \text{logit}^{-1}[\boldsymbol{\beta}\hat{\boldsymbol{x}}_i]}{\sum_{i=1}^{i=n} \mathcal{I}(x_{ij} = k)}$$

where we incorporate the intercept multiplier of 1:

$$\hat{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i) = (1, x_{i,\text{Sex}}, x_{i,\text{Age}}, x_{i,\text{Education}}, x_{i,\text{Income}})$$

and our prediction for each agent i is then

$$\boldsymbol{\beta}\hat{\boldsymbol{x}}_i = \beta_0 + \beta_1 x_{i,\text{Sex}} + \beta_2 x_{i,\text{Age}} + \beta_3 x_{i,\text{Educ}} + \beta_4 x_{i,\text{Inc}}$$

and the indicator function, applying the prediction to those agents whose category j has value k,

$$\mathcal{I}(x_{ij} = k) = \begin{cases} 0 & \text{if } x_{ij} \neq k \\ 1 & \text{if } x_{ij} = k \end{cases}$$

and, x_i and n, above, respectively refer to the respondents in and size of our PUMS sample. Refer to the Appendices A.1, A.2, and A.3 for further details on how we obtain the marginal data from the source papers.

4.2 Fitting to Sample Population Error Rates

Several of the studies also report overall intentional error rates for their sample population; we incorporate these into the likelihood, assuming they are not automatically inferred from the covariate meta-analysis. Antonides and Robben (1995) report a sample tax evasion rate of 19.6% in their 188 respondents. Also, Wahlund (1992) reports 37% of his 430 respondent pool evaded taxes.[6] And finally, 50% of the respondents from Collins et al.'s study report some evasion.[7] Hence, we incorporate each of these population sample error rates into our estimation:

$$\text{logit}[\overline{p}(y>0)] \sim \text{N}(\mu = \text{logit}[p_i], \sigma^2 = [n_i p_i (1-p_i)]^{-1})$$

where $i \in \{\text{Antonides, Wahlund, Collins}\}$ and

$$\begin{aligned} p_{\text{Antonides}} &= 0.196 \text{ and } n_{\text{Antonides}} = 188 \\ p_{\text{Wahlund}} &= 0.370 \text{ and } \quad n_{\text{Wahlund}} = 430 \\ p_{\text{Collins}} &= 0.500 \text{ and } \quad n_{\text{Collins}} = 240 \cdot \kappa_{\text{C}}^n \end{aligned}$$

and our PUMS population sample error rate is:

$$\overline{p}(y>0) = \frac{\sum_{i=1}^{n} \text{logit}^{-1}[\boldsymbol{\beta}\boldsymbol{x}_i]}{n}$$

Later, we will discuss differentially weighting the empirical rates based on how relevant and similar the sampled population is to the PUMS; thus, the Collins sample, which is both recent and U.S.-based, will receive a special weighting κ_{C}^n. So now, the log-likelihood from fitting these population error rates are added together, and added eventually to the overall log-likelihood:

$$\mathcal{L}_{\text{Pop}} = \sum_{i \in \{\text{Antonides, Wahlund, Collins}\}} \log[\text{N}(\overline{p}(y>0) | \mu = \text{logit}[p_i], \sigma^2 = [n_i p_i (1-p_i)]^{-1})]$$

4.3 Fitting to Empirical Coefficients

Instead of marginal covariate error rates, Collins et al. (1992) report multiple regression coefficients predicting counts of evasion per respondent. Since we cannot fit our predictive means to to any corresponding empirical means, we instead fit our expected/predicted regression coefficients $\boldsymbol{\beta}$ directly onto those of Collins et al.:

$$\boldsymbol{\beta_{-0}} \sim \text{N}(\boldsymbol{\beta}_{-0,\text{Collins}}, \boldsymbol{\sigma}^2_{\text{Collins}})$$

[6] We eschew the fit to Wahlund's population sample error rate when we include the 'obey law' covariate, in Section 8, since fitting to the latter already includes the former.

[7] We maintain the Collins' population error rate, even when we include the Collins covariates due to omission of the intercept.

where -0 subscript indicates that we omit the intercept. The log-likelihood is then:

$$\mathcal{L}_{\text{Collins}} = \sum_{i=0}^{4} \log[\text{N}(\beta_i | \beta_{i,\text{Collins}}, \sigma^2_{i,\text{Collins}})]$$

where $i \in \{1, 2, 3, 4\}$ maps on to {Sex, Age, Education, Income}. Details on how we obtain $\boldsymbol{\beta}_{-0,\text{Collins}}$ and $\boldsymbol{\sigma}_{-0,\text{Collins}}$ may be found in Appendix A.4.

4.4 Meta-Analytical Log-Likelihood

The complete log-likelihood is then the sum of the aforementioned log-likelihoods, $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{Marginal}} + \mathcal{L}_{\text{Pop}} + \mathcal{L}_{\text{Collins}}$$

Note, $\mathcal{L}_{\text{Marginal}}$ was not explicitly reported but easily derived from Eq. 2.

5 Results of Meta-Analysis

Since several of our empirical sources do not survey a recent U.S. population sample, hence their marginals and estimates may not be generalizable to our PUMS sample, it behooves us to weight their findings during our estimation process such that greater weight is given to those populations which are recent and/or U.S.-based. As such, we introduce two scaling factors/multipliers, κ^n_{C} and κ^n_{M}, applied to the sample sizes of the Collins et al. and Mason and Calvin studies (both U.S.-based), where $n_{\text{Collins}} = 240$ and $n_{\text{Mason}} = 800$, respectively. Since the Collins et al. study occurred recently (and has a relatively low sample size), we offer it a weight of $\kappa^n_{\text{C}} = 8$, while doubling the influence of the Mason and Calvin study, $\kappa^n_{\text{M}} = 2$. Furthermore, the Vogel study reported a curvilinear (i.e. inverted V-shape) prediction from 'education' on intentional error. We believe this effect to be significant enough to warrant investigation a separate model.

So, in Table 1, we present the four model variants, obtained by fitting the two sets of κ^n and two treatments of 'education' with our $n = 10{,}000$ PUMS sub-sample. The second model of each pair (b) highlights Vogel's curvilinear effect from 'education'; while this model is superior to the first model, the difference in log-likelihood indicates only a marginally better fit for the unweighted fit. However, for the weighted condition, the non-linear 'education' model is a substantially better fit, obliging us to consider this variant as being closer to the 'true' model. We obtain the predicted probabilities, $\boldsymbol{p}$, by applying each model, $\boldsymbol{\beta}$, to the covariate values for each of our PUMS sample agents, $\hat{\boldsymbol{x}}_i$, while not forgetting the intercept prefix:

$$p_i = p(y_i > 0) = \text{logit}^{-1}[\boldsymbol{\beta}\hat{\boldsymbol{x}}_i]$$

And, we can offer the summary statistics for the distribution of predicted probabilities from

	Unweighted: $\kappa_{\mathrm{C}}^n = 1, \kappa_{\mathrm{M}}^n = 1$: $n_{\mathrm{Collins}} = 240, n_{\mathrm{Mason}} = 800$		Weighted: $\kappa_{\mathrm{C}}^n = 8, \kappa_{\mathrm{M}}^n = 2$: $n_{\mathrm{Collins}} = 1920, n_{\mathrm{Mason}} = 1600$	
Predictor	Model #1a	Model #1b	Model #2a	Model #2b
Intercept	-0.922^{***} (0.141)	-0.842^{***} (0.145)	-0.415^{***} (0.108)	$-0.230^{\hat{}}$ (0.118)
Sex	0.411^{**} (0.141)	0.383^{**} (0.116)	0.339^{***} (0.087)	0.377^{***} (0.088)
Age	-0.583^{***} (0.114)	-0.598^{***} (0.110)	-0.531^{***} (0.089)	-0.522^{***} (0.091)
Education	0.080 (0.133)		-0.045 (0.070)	
$\lvert\text{Educ} - 1\rvert$		$-0.220^{\hat{}}$ (0.113)		-0.472^{***} (0.113)
Income	0.065 (0.142)	0.141^{*} (0.070)	-0.026 (0.055)	-0.028 (0.047)
$\mathcal{L}$	-92	-90	-324	-316
n	10000	10000	10000	10000

Table 1: Four Model Variants

each model:

	Summary Statistics for $p(y < 0)$					
	Min.	25%	Median	Mean	75%	Max.
Model #1a	0.1102	0.2041	0.2774	0.2694	0.3229	0.5130
Model #1b	0.0631	0.1797	0.2431	0.2497	0.3121	0.5959
Model #2a	0.1488	0.2482	0.3040	0.3069	0.3684	0.4811
Model #2b	0.0558	0.1930	0.2272	0.2655	0.3203	0.5367

Interestingly, the non-linear 'education' covariate confers a greater range of fitted probabilities for committing non-compliance, while reducing the mean, suggesting that the strictly linear model might be biased. While we expect the mean error rate to climb under the weighted models due to the imposition of the Collins' error rate, the non-linearity in the (b) models mitigates this effect.

6 Intentional Error in Line Items/Taxpayer Categories

We now wish to predict intentional error occurring from several sources of opportunity on the tax return, particularly errors in line items pertaining to several taxpayer categories. We aim to express distinct 'line item' error models for each of these categories, but first, we introduce our approach with a single category. For the analyses in this section, we employ the same

PUMS sub-sample, supplemented with taxpayer category assignment flags for each agent. While some of these categories are taken directly from the PUMS respondent data (e.g. self-employment), others were strongly inferred (e.g. tip-earners from appropriate occupations), while the remaining are moderately presumptive (e.g. EIC/EITC line items flagged for all who qualify).[8]

6.1 Univariate Solution

Given that our intentional error model involves multiple covariates, we can easily infer the change in error commission probability of any one of the taxpayer categories across a single covariate. Essentially, the remaining covariates, not involved in the univariate inference, provide enough covariance structure to allow us to make some statement on the impact of the focal covariate on the commission of error in a particular line-item category.[9] We offer several examples before moving on to the multivariate solution.

In our first example, we focus on the first taxpayer category 'Tips' and infer the naïve probability that an individual who falls into that category will commit an error from a line-item associated with that category; here, there is no predictive covariate. We first observe the following when we apply the PUMS sub-sample to our predictive, generic error model from Eq. 1:

$$p(y > 0|x^{\text{Tips}} = 0) = 0.253$$
$$p(y > 0|x^{\text{Tips}} = 1) = 0.355$$

where y is the number of intentional errors. Essentially, we observe that the mean probability of some error given no 'Tips' is 25.3% and with 'Tips' is 35.5%; the socio-demographic error model already predicts higher non-compliance for tip earners. So, we define two sources of intentional error, 'Tips' and '–Tips', where the latter denotes error from any non-'Tips' source, and elaborate on the above estimates:

$$p(y > 0|x^{\text{Tips}} = 0) = 0.253 = 1 - \qquad (1 - p(y^{-\text{Tips}} > 0))$$
$$p(y > 0|x^{\text{Tips}} = 1) = 0.355 = 1 - (1 - p(y^{\text{Tips}} > 0) \cdot (1 - p(y^{-\text{Tips}} > 0))$$

We easily solve for $p(y^{\text{Tips}} > 0)$ and obtain 0.137; that is, assuming independence, a taxpayer who falls under the 'Tips' category has a 13.7% chance of committing an intentional error in line-items related to 'Tips' and a 25.3% chance of committing an error through non-'Tips' line-items.[10]

We can extend this approach to assess the impact of gender/sex on the proclivity towards intentional error commission of, again, 'Tips'. Essentially, we wish to find the parameters

[8] EIC/EITC is the acronym for the Earned Income (Tax) Credit.

[9] It remains to be seen how this constriction on the covariance, due to use of a direct logit model, influences the validity of the results.

[10] There is some evidence, particularly from Collins et al. (1992) and Wahlund (1992), that there is modest dependency between errors; one explanation is that an individual with multiple opportunities is likely to commite more errors. We will relax the independence assumption in later writings.

for the following model:

$$\text{logit}\big[p(y^{\text{Tips}} > 0 | x^{\text{Tips}} = 1)\big] = \beta_0 + \beta_1 \cdot x_{\text{Sex}} \tag{3}$$

Our data provides us with the following summary statistics:

Sample Size, n

Tips \ Sex	0 = F	1 = M
0	4889	4806
1	230	75

Mean Error Rate, $\overline{p}$

Tips \ Sex	0 = F	1 = M
0	0.191	0.316
1	0.296	0.536

Firstly, we observe more females than males earn tips, which is what we would expect from an unbiased sample. However, in accordance to the intentional error model, male tip-earners are more likely to commit acts of non-compliance than female tip-earners (i.e. $0.536 > 0.296$). We elaborate on the mean error rate table:

Given $x_{\text{Sex}} = 0$:

$$\begin{aligned} p(y > 0 | x^{\text{Tips}} = 0) &= 0.191 = 1 - \qquad\qquad (1 - p(y^{-\text{Tips}} > 0)) \\ p(y > 0 | x^{\text{Tips}} = 1) &= 0.296 = 1 - (1 - p(y^{\text{Tips}} > 0) \cdot (1 - p(y^{-\text{Tips}} > 0) \end{aligned}$$

Given $x_{\text{Sex}} = 1$:

$$\begin{aligned} p(y > 0 | x^{\text{Tips}} = 0) &= 0.316 = 1 - \qquad\qquad (1 - p(y^{-\text{Tips}} > 0)) \\ p(y > 0 | x^{\text{Tips}} = 1) &= 0.536 = 1 - (1 - p(y^{\text{Tips}} > 0) \cdot (1 - p(y^{-\text{Tips}} > 0) \end{aligned}$$

Solving for each x_{Sex} conditional, we obtain:

$$\begin{aligned} p(y^{\text{Tips}} > 0 | x^{\text{Tips}} = 1, x_{\text{Sex}} = 0) = 0.130 \\ p(y^{\text{Tips}} > 0 | x^{\text{Tips}} = 1, x_{\text{Sex}} = 1) = 0.322 \end{aligned}$$

giving us the following solution for Eq. 3:

$$\begin{aligned} \beta_0 &= -1.901 = \text{logit}[0.130] \\ \beta_1 &= \quad 1.156 = \text{logit}[0.322] - \text{logit}[0.130] \end{aligned}$$

We can continue this analysis and infer the effect of sex/gender on non-'Tips' error commission. However, we must admit a weakness to this method: the covariance afforded by our intentional error model can only go so far, being effective for only univariate or small multivariate models. Our goal of inferring line-item error commission using all of the key covariates (i.e. Sex, Age, Education, and Income) is non-sensical with the above method, because our intentional error model neither 1) is contingent on taxpayer category nor 2) does it have additional sources of covariance. So, for any unique socio-demographic category, defined by a combination of the four covariates, the predicted error is identical for categorical or non-categorical taxpayers, alike. Therefore, we must turn to a different approach if we wish to offer a predictive line-item model that employs all four socio-demographic covariates.

6.2 Multivariate Solution

As with the univariate inference, we first obtain error probabilities for our subsample using the intentional error model on the socio-demographic covariates: Sex, Age, Education, and Income. Our goal is to construct separate predictive models for each of the taxpayer categories using the same socio-demographic covariates as the intentional error model; that is, each of the models will reflect the probability of intentional error occurring in the line-items associated with each taxpayer category.

Our nomenclature for the taxpayer categories (tp) of interest is as follows, noting presumptive categories:

Taxpayer/Line-Item Category, tp, ...	comprises individuals who ...
Tips	receive tips
SEmp	are self-employed or pay self-employment tax
EIC	(probably) take the Earned Income (Tax) Credit
SLns	(might) have student loans
Cap	report capital gains (i.e. own a house)
Frm	own a farm
SSB	receive social security benefits (due to appropriate age)

In Figure 1, we present, graphically, the proportions for each of the taxpayer category for several sub-samples for comparison purposes and to demonstrate the modest degree of variance across cities and sample sizes. We present our models, reiterating Eq. 1 as the source of the dependent data. A similar formulation is employed for the predictive component, which fits specifically for only categorical taxpayers:

$$\text{logit}(p_i) = \beta_0 + \beta_{\text{Sex}} \cdot x_{i,\text{Sex}} + \beta_{\text{Age}} \cdot x_{i,\text{Age}} + \dots \tag{4}$$

$$\text{logit}(q_i^{tp} | x_i^{tp} = 1) = \alpha_0^{tp} + \alpha_{\text{Sex}}^{tp} \cdot x_{i,\text{Sex}} + \alpha_{\text{Age}}^{tp} \cdot x_{i,\text{Age}} + \dots \tag{5}$$

where p_i is shorthand for $p(y_i > 0)$, the probability of commission of *at least one* intentional error for agent i, and q_i^{tp} (similarly abbreviated) is the predicted probability that an error occurs in the line-item(s) associated with a single taxpayer category, tp, where $tp \in \{\text{Tips, SEmp, EIC, SLn, Cap, Frm, SSB, Misc}\}$, given that agent i falls in that particular category. As such, the second model fits only those agents who can possibly commit error from one of the seven categories.

Note, in addition to the explicit taxpayer categories, we include a miscellaneous category, 'Misc', since we will also have to model errors occurring in line-items not covered in the seven, aforementioned taxpayer categories.[11] Hence, the model for $tp =$ 'Misc' will be close to the intentional error model; the intercept will be the only parameter we fit in the estimation

[11] An alternative simpler approach assumes that all errors will occur in the provided line-item categories. However, since our data contains agents who do not fall into any of the seven taxpayer categories, we can estimate the rate of error incurred through line-items outside those categories and avoid this over-simplification.

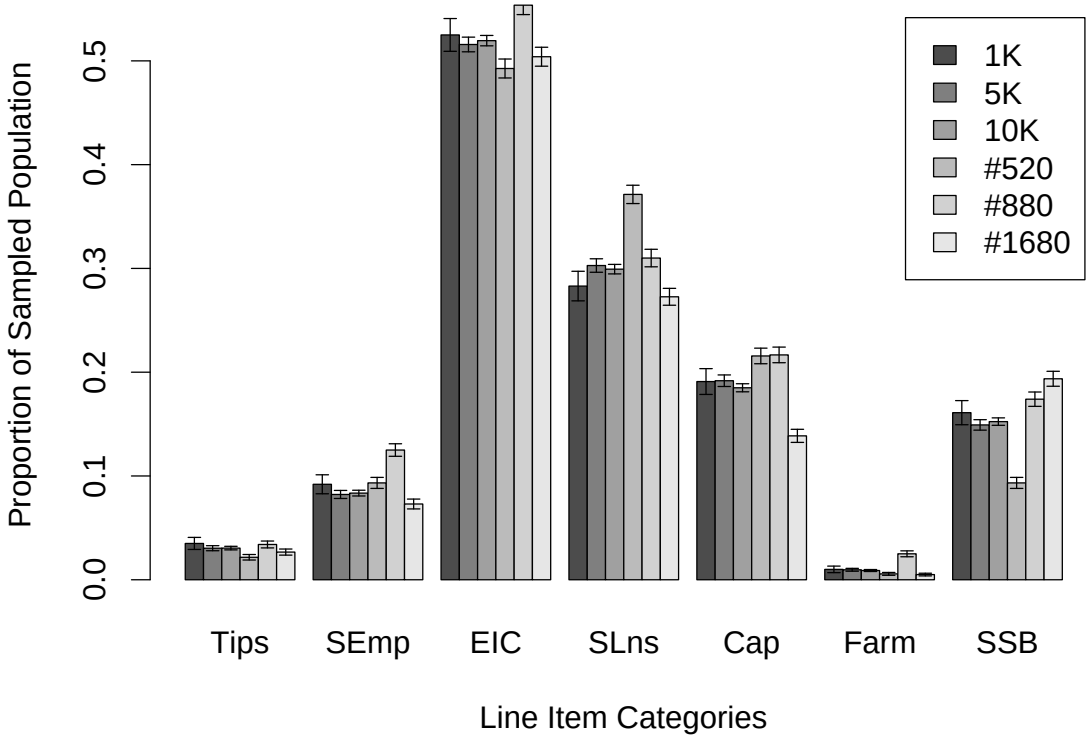


Figure 1: Taxpayer/Line-Item Categories. *The error bars represent the nominal standard deviation surrounding a proportion. The third bar of each set represents the proportions from our main* $n = 10{,}000$ *subsample. The latter three bars represent data from specific city samples, labeled by their City ID #.*

process. While, realistically, the rate of 'Misc' error likely covaries with the rate of error from the explicit categories, we have no way of modeling this; hence, we assume independence across all categories, including 'Misc'.

The second model, Eq. 5, predicts the probability of error commission for each category. Since, at this stage, we assume independence, we can directly calculate the probability of *at least one* error as:

$$p_i \approx q_i = 1 - \prod_{tp}(1 - q_i^{tp}) \tag{6}$$

where, again, $tp \in \{\text{Tips, SEmp, EIC, SLn, Cap, Frm, SSB, Misc}\}$ and i refers to a specific agent/PUMS respondent. Basically, the probability of some error is the probability of the non-occurrence of no errors. If an agent does not qualify for taxpayer category tp, the probability is naturally 0, i.e. $q_i^{tp} = 0 | x_i^{tp} = 0$. So, essentially, we are seeking a matrix of coefficients/estimates (i.e. a model for each of the seven plus one categories):

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_0^{\text{Tips}} & \alpha_1^{\text{Tips}} & \dots & \alpha_4^{\text{Tips}} \\ \vdots & \ddots & \dots & \vdots \\ \alpha_0^{\text{SSB}} & \alpha_1^{\text{SSB}} & \dots & \alpha_4^{\text{SSB}} \\ \alpha_0^{\text{Misc}} & \alpha_1^{\text{Misc}} & \dots & \alpha_4^{\text{Misc}} \end{bmatrix}$$

Out of 168 unique socio-demographic combinations, our sample contains 158.[12] With number of bins per covariate, reiterated here:

	Covariate			
	Sex	Age	Educ	Inc
$m_{\text{Covariate}}$	2	3	4	7

we indicate each unique agent combination using a unidimensional scalar, uid_i:[13]

$$\begin{aligned} f_{\text{uid}}(\boldsymbol{x}_i) = uid_i &= x_{i,\text{Sex}} m_{\text{Age}} m_{\text{Educ}} m_{\text{Inc}} + x_{i,\text{Age}} m_{\text{Educ}} m_{\text{Inc}} + x_{i,\text{Educ}} m_{\text{Inc}} + x_{\text{Inc}} \\ &= x_{i,\text{Sex}}(3)(4)(7) + x_{i,\text{Age}}(4)(7) + x_{i,\text{Educ}}(7) + x_{i,\text{Inc}} \\ &= (84)x_{i,\text{Sex}} + (28)x_{i,\text{Age}} + (7)x_{i,\text{Educ}} + x_{i,\text{Inc}} \end{aligned} \tag{7}$$

Conversely, $\boldsymbol{x}_{uid}$ indicates a vector $(x_{\text{Sex}}, x_{\text{Age}}, x_{\text{Educ}}, x_{\text{Inc}})$ corresponding to to the uid scalar. Now, we collate the predicted probabilities q_i for all agents i who have the same uid:

$$\boldsymbol{Q}_{uid} = \{q_{j,uid}\} \text{ where } j \in \{f_{\text{uid}}(\boldsymbol{x}_i) = uid\}$$

Giving us the fitting likelihood for each agent within a given uid:

$$\text{logit}[\boldsymbol{Q}_{uid}] \sim \text{N}(\text{logit}[p_{uid}], \sigma_{uid}^2) \tag{8}$$

[12] In future writings, we will perform our inference with a much larger PUMS sub-sample, which will likely contain more socio-demographic combinations.

[13] Note that we distinguish 'uid', a function label, and '*uid*', an identifier variable/value.

where σ^2_{uid} can be one of a) the standard error surrounding the fit for $\boldsymbol{x}_{uid}$ or b) the variance of the logit of the predicted generic error fit (i.e. logit$[p_{uid}]$) computed with the count of data points in the PUMS that constitute the *uid*. More specifically, each of these variance sources are:

a) $\sum_{i=0}^{4}(x_i^{uid})^2\text{Var}(x_i^{uid}) + 2\sum_{i,j;i<j}^{4} x_i^{uid}x_j^{uid}\text{Cov}(x_i^{\text{uid}}, x_j^{\text{uid}})$; here, we employ the covariance (and variance) of the generic error Model #2a in Table 1.

b) $[n_{uid}p_{uid}(1-p_{uid})]^{-1}$ where $n_{uid} = |\boldsymbol{x}^{uid}|$, the count of PUMS agents that fall under *uid*.

However, since we are dealing with probabilities, that is, the binomial distribution, the final log-likelihood ought to consider just the mean fit for each *uid*, computed across all *uid*'s, rather than each individual q_i:

$$\overline{Q}_{uid} = \frac{\sum_{i=1}^{n_{uid}} \mathcal{I}(f_{\text{uid}}(\boldsymbol{x}_i) = uid) \cdot q_{i,uid}}{n_{uid}}$$

where our count of PUMS agents per *uid* is $n_{uid} = \sum \mathcal{I}(f_{\text{uid}}(\boldsymbol{x}_i) = uid)$. And now, the new likelihood substitutes $\overline{Q}_{uid}$ for the vector $\boldsymbol{Q}_{uid}$ in Eq. 8.

Furthermore, we employ two additional fits, which account for the higher chance of error commission given more opportunities. The first fits the Pearson correlation of opportunity as defined by the number of taxpayer categories an agent falls into, and the second, fits specifically the probability of error committed by those self-employed, as we have direct empirical data for this taxpayer category. First, we claim that the number of taxpayer categories for which a single agent qualifies, not counting 'Misc', adequately serves as a proxy for the degree of opportunity to commit error. So, for each agent i:

$$n_i^{\text{tp}} = \sum_{tp} \mathcal{I}(x_{i,tp} = 1)$$

From an analysis of Wahlund's findings (see Appendix A.5), we obtain a correlation of $\rho_{\text{W}} = 0.329$ between the degree of opportunity and at least one act of evasion. However, since there is some uncertainty as to how accurate this correlation is for taxpaying population of the United States in the year 2000, we consider a modification to the canonical variance around ρ when fitting our line item error models:[14]

$$\hat{\rho} \sim \text{N}\left(\mu = \rho_{\text{W}}, \sigma^2 = (\kappa^{\rho}_{\text{W}})^2 \cdot \frac{1-\rho^2_{\text{W}}}{n_{\text{W}} - 2}\right)$$

[14]However, we can say with high certainty that there ought to be both a positive and a moderate to large correlation.

where

$$
\begin{aligned}
n_{\mathrm{W}} &= \text{the size of Wahlund's sample, 430} \\
\kappa^{\rho}_{\mathrm{W}} &= \text{a presumptive scaling factor/multiplier of the standard deviation of } \rho \\
\hat{\rho} &= \text{our predicted correlation between the vector of counts of} \\
&\quad \text{opportunity, } \boldsymbol{n}^{\mathrm{tp}}, \text{ and the probability of error } \boldsymbol{P} \text{ for all agents}
\end{aligned}
$$

In our next addition, we fit our joint line item predictions to the marginal rates of error for self-employment, such that these predictions conform to their empirical analogue:

$$\operatorname{logit}\left[\overline{q}^{\mathrm{SEmp}=j}\right] \sim \mathrm{N}\left(\mu = \operatorname{logit}\left[p_s^{\mathrm{SEmp}=j}\right], (\sigma_s^{\mathrm{SEmp}=j})^2 = [n_s^{\mathrm{SEmp}=j} p(1-p)]^{-1}\right)$$

where the mean marginal prediction of error for self-employment status is:

$$\overline{q}^{\mathrm{SEmp\,=\,j}} = \frac{\sum_i q_i \cdot \mathcal{I}(x_i^{\mathrm{SEmp}} = j)}{\sum_i \mathcal{I}(x_i^{\mathrm{SEmp}} = j)}$$

and the marginal rates of error for each source and each self-employment status are:[15]

s	j	p_s^{SEmp}	n_s^{SEmp}
Vogel	0	0.279	967
Vogel	1	0.371	106
Houston	0	0.084	144
Houston	1	0.188	79

That is, we are fitting our self-employment predictions to four likelihoods, two for each self-employment status times two for each source. Furthermore, we examine the same fit to self-employment, but under the Beta distribution:[16]

$$\overline{q}^{\mathrm{SEmp=j}} \sim \mathrm{Beta}\left(\alpha = p_s^{\mathrm{SEmp}=j} n_s^{\mathrm{SEmp}=j} + 1, \beta = (1 - p_s^{\mathrm{SEmp}=j}) n_s^{\mathrm{SEmp}=j}) + 1\right)$$

6.3 Methodology

Due to the large parameter space, inferring a set of line item error models requires two steps:

1. We employ a gradient descent/stochastic maximization algorithm to reach a solution sufficiently close to the mode of the multivariate distribution. Using a predefined shape parameter, s, for each of the forty β coefficients. This will confine the size of the jump to a new value; typically, $s = 0.001$ seems sufficient:
   (a) Draw a random 'jumping' delta from a Gaussian: $\delta \sim \mathrm{N}(0, s^2)$.
   (b) Obtain the $\mathcal{L}$ (i.e. fit of the entire set of line item models) for each of $(\beta - \delta, \beta, \beta + \delta)$, three values.

[15] This data is reproduced from Appendices A.1 and A.3.

[16] In fact, we employ the beta fit for our reported line item error models.

(c) Select the change (or no change) that corresponds to the highest $\mathcal{L}$.

(d) Repeat (a)–(c) until no changes are accepted for the entire set of coefficients.

2. Then, we employ the Newton-Raphson optimization procedure, using the result of the stochastic maximization as our initial point, in order to find the true mode as well as the variance surrounding the mode, under the assumption of normality.[17]

7 Line Item Model Results

We present summary results for several different approaches and parameterizations towards fitting the set of line item models in Table 2. These variations/condition include:

1. varying the weight on the Collins data by increasing the sample size by the multiplier, $\kappa_{\mathrm{C}}^{n} \in \{1, 4, 8\}$.

2. fitting to the means (and standard error) for each *uid* (μ) or the overall mean of the log-likelihoods from the former fits, ($\overline{\mathcal{L}}$).

3. relaxing the variance around each *uid* fit (i.e. μ) so that they are proportional to the variance around each logit, $\propto\sigma^2$, or equalizing the variance, $=\sigma^2$, which assumes all *uid*'s are represented by the same number of data points.

4. imposing different priors, Cauchy or normal, on the model coefficients, varying the shape or variance. Refer to Appendix C for a comparison of the different priors.

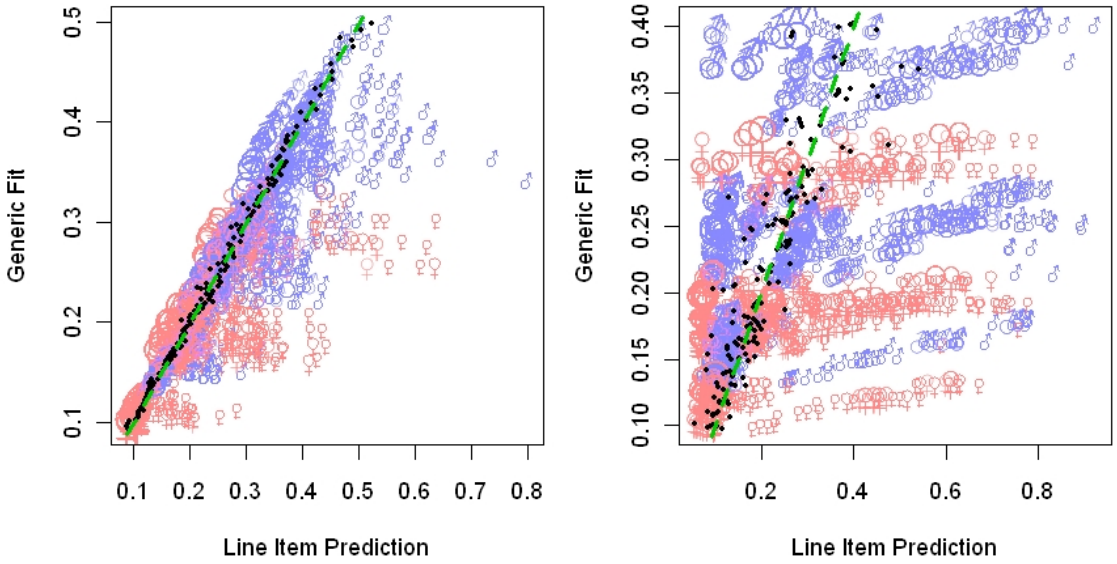
5. varying the uncertainty around the opportunity/evasion correlation ρ with a multiplier $\kappa_{\mathrm{W}}^{\rho}$.

Predicted error probabilities for each line item are computed using Eq. 5 in the same manner as we generated the generic error probabilities. The predicted 'any' error probability is easily calculated with Eq. 6, the summary of which we present in Table 2. One notable observation in Table 2 is that an increase in the error probability spread seems necessary for the opportunity/error correlation to fit. In Figure 2, we naturally see that expanding the variance around the original prediction (right plot), increases the spread of the line item model predictions, with the consequence of less fitting means for each *uid*. Furthermore, we can easily see the distinction in error between males and females. For further detail, we report the line item error models fitted to three of the above conditions. The first one employs the mean fit (μ) to the standard error, while increasing the uncertainty around the

[17] While the posterior distributions might possibly depart widely from the normal, we suspect the divergence is not serious, obviating the need to employ MCMC Bayesian inference methods.

					Summary of Error: $p(y > 0)$						
	fitting	prior		$\kappa^{\rho}_{\mathrm{W}} =$		1st			3rd		
κ^{n}_{C}	method	dist.	shape	$\sigma_{\rho}\times$	Min	Qu.	Med	μ	Qu.	Max.	ρ
1	μ	N	3	2	0.09	0.17	0.23	0.24	0.30	0.78	0.049
4	μ	N	3	2	0.10	0.16	0.23	0.23	0.28	0.82	0.066
8	μ	N	3	2	0.09	0.17	0.23	0.24	0.28	0.85	0.072
8	μ	N	10	5	0.10	0.18	0.23	0.23	0.28	0.78	0.043
8	$\overline{\mathcal{L}}$	Cauchy	10	5	0.09	0.17	0.23	0.23	0.27	0.69	0.046
8	μ	Cauchy	10	1	0.09	0.15	0.20	0.23	0.27	0.99	0.137
8	μ	Cauchy	3, 2.5	1	0.09	0.15	0.21	0.24	0.28	0.91	0.128
8	μ	Cauchy	3, 2.5	0.01	0.00	0.07	0.15	0.23	0.32	1.00	0.329
8	$\mu, \propto \sigma^2$	Cauchy	3, 2.5	1	0.06	0.10	0.18	0.24	0.29	0.93	0.258
8	$\overline{\mathcal{L}}, \propto \sigma^2$	Cauchy	3, 2.5	1	0.03	0.11	0.19	0.23	0.30	0.85	0.274
8	$\mu, \propto \sigma^2$	Cauchy	3, 2.5	0.1	0.00	0.07	0.18	0.23	0.34	0.98	0.328
8	$\mu, = \sigma^2$	Cauchy	3, 2.5	1	0.06	0.14	0.20	0.25	0.29	0.94	0.206

Table 2: Summary of Error Prediction Distributions from Line Item Models for Different Fitting Parameterizations. *The spread of predicted probabilities arising from the line item models is shown for each of the different model conditions. The predicted error/opportunity correlation, ρ, is also reported.*



(a) Original Fit (b) Wider Variance

Figure 2: Line Item Fit. *Each graph plots the line item prediction (as a probability on the X-axis) against the accompanying generic error model for each uid (Y-axis). The original fit corresponds to the first line in Table 2, wherein the line item models aim to remain within the standard error of each generic fit. The right plot incorporates a wider variance around the fits as determined by the number of data points that exist. The black points depict our predicted mean for each agent uid category and the green line shows the where the perfect fit would lie. The sizes of the red ♀ (female) and blue ♂ (male) points are exponentially proportional to number of data points that exist for the category.*

correlation to opportunity:

$\kappa^n_{\mathrm{C}} = 8$, fit $= \mu$, $\kappa^\rho_{\mathrm{W}} = 2$
prior dist. $= \mathrm{N}(0, \sigma^2 = 3^2)$

Variable	Tips	SEmp	EIC	SLns	Cap	Frm	SSB	Misc
Intercept	−1.34	−0.46	−3.10*	−2.48*	−1.55	−0.21	−0.95	−1.12***
	(2.25)	(1.84)	(1.35)	(1.10)	(1.45)	(2.61)	(2.62)	(0.20)
Sex	−0.21	−0.30	−0.13	0.18	0.75	0.16	−0.29	0.39***
	(2.75)	(1.85)	(2.16)	(0.64)	(1.00)	(2.50)	(2.19)	(0.09)
Age	−0.83	−0.90	−0.96	−0.41	−0.87	−0.85	−1.77	−0.44***
	(2.42)	(1.39)	(2.44)	(0.57)	(1.47)	(1.91)	(1.49)	(0.12)
Education	−0.48	−0.66	−0.83	−0.12	−0.84	−0.06	−0.02	−0.09ˆ
	(2.24)	(1.12)	(1.75)	(0.40)	(1.11)	(1.54)	(1.39)	(0.05)
Income	−0.80	−0.16	−0.53	−0.06	−0.25	−0.05	−0.29	0.01
	(3.00)	(0.57)	(3.30)	(0.24)	(0.46)	(1.10)	(0.88)	(0.03)

The next two incorporates several enhancements: 1) fitting to the empirical variance (proportional or equal), 2) a Cauchy prior to the coefficients, 3) canonical deviation around the opportunity correlation:

$\kappa^n_{\mathrm{C}} = 8, \mathrm{fit} = \{\mu, \propto \sigma^2\}, \kappa^\rho_{\mathrm{W}} = 1$
prior dist. $= \mathrm{Cauchy}(x_0 = 0; \gamma_0 = 3, \gamma_{-0} = 2.5)$

Variable	Tips	SEmp	EIC	SLns	Cap	Frm	SSB	Misc
Intercept	−0.71	−0.74	−1.74**	−1.49	0.15	0.02	−0.85	−2.73***
	(1.89)	(1.76)	(0.62)	(1.05)	(1.07)	(1.98)	(2.04)	(0.77)
Sex	0.04	−0.35	0.49	0.35	0.72	0.10	−0.06	0.35
	(1.77)	(1.59)	(0.60)	(0.58)	(0.92)	(1.70)	(1.45)	(0.42)
Age	−0.51	−0.66	−1.58	0.01	0.13	−0.13	−1.32	0.04
	(1.60)	(1.50)	(1.25)	(0.59)	(0.79)	(1.62)	(1.29)	(0.46)
Education	−0.23	−0.74	−0.75	−0.19	−0.67	−0.18	0.15	−0.07
	(1.53)	(1.64)	(0.84)	(0.42)	(0.82)	(1.55)	(1.20)	(0.32)
Income	−0.47	0.04	−0.26	−0.00	−0.15	0.31	−0.14	0.09
	(1.85)	(0.73)	(0.58)	(0.22)	(0.32)	(1.38)	(0.77)	(0.15)

	same as above except fit = $\{\mu, = \sigma^2\}$							
Variable	Tips	SEmp	EIC	SLns	Cap	Frm	SSB	Misc
Intercept	0.09	−0.33	−3.59ˆ	−1.36	0.46	0.06	−0.98	−1.77***
	(1.84)	(1.76)	(2.16)	(1.09)	(1.03)	(1.99)	(1.95)	(0.34)
Sex	−0.46	−0.44	0.32	0.19	0.79	−0.12	0.05	0.40ˆ
	(1.74)	(1.56)	(0.96)	(0.55)	(0.83)	(1.69)	(1.04)	(0.22)
Age	−0.22	−0.59	0.19	−0.05	0.17	−0.26	−0.93	−0.45*
	(1.57)	(1.46)	(1.03)	(0.56)	(0.86)	(1.53)	(0.98)	(0.21)
Education	−0.11	−0.56	0.00	−0.23	−1.08ˆ	−0.19	0.12	−0.09
	(1.28)	(1.20)	(0.52)	(0.48)	(0.65)	(1.35)	(0.61)	(0.16)
Income	−0.60	−0.23	−0.69	−0.11	−0.55ˆ	0.04	−0.23	0.11ˆ
	(0.69)	(0.51)	(1.80)	(0.19)	(0.31)	(0.96)	(0.39)	(0.06)

The broad standard errors reflects the variance in both the empirical marginal statistics and the opportunity structure (i.e. the covariance in the PUMS between the taxpayer categories and the socio-demographics). With a few exceptions, the direction of the coefficients align with the generic error model. Those exceptions are insignificant and, at best, might indicate a possible reversal in trend; for instance, we find that females who receive tips and social security benefits are more likely to commit error than their male counterparts, which may or may not be an artifact of their over-representation in these categories.

8 Error Models with Obey Law

There is considerable evidence that links an individual's attitudes to crime and obedience to the law to tax evasion. The background analysis for connecting intentional error to both 'attitudes to crime' and 'obedience to the law' may be found in Appendices A.6.1 through A.6.3, with the aggregation of empirical findings reported A.6.2 and the estimation/imputation approach detailed in the A.6.3.

8.1 Predicting with Imputation

In this section, we examine several different 'obey law' models:

1. The 'Basic' model is the original, generic error model, which omits 'obey law'.
2. 'No Obey' includes the imputed 'obey law' only as a covariate (i.e. we do not fit to the inferred obew law/evasion correlation). This condition also employs the beta prior during imputation
3. 'No W' includes obey law as a covariate and fit; however, it ignores the Wahlund population error rate fit as it is redundant. As above, this condition also employs the beta prior during imputation.
4. 'Prior' employs the beta prior during imputation and both the 'obey law' and Wahlund population error fit.

5. 'M.I.' executes multiple imputation without the beta prior (i.e. drawing 'obey law' from a binomial parameterized by the GSS *uid* means); the Wahlund population error rate is also included.

6. 'Mean' dichotomizes each agent's predicted obey law probability, $p > 0.5$, to obtain the 0/1 obey law response; this condition is offered as a control or naïve case.

In the Table 3, we first present results for the unweighted condition, $\kappa_C^n = 1, \kappa_M^n = 1$, i.e. the Collins and Mason estimates are treated as equally as important as the findings from other countries. We notice that Age's contribution towards error is apparently unaffected

Predictor	Basic	No Obey	No W	Prior	M.I.	Mean
Intercept	-0.922^{***}	-0.515	-0.068	-0.031	-0.027	0.101
	(0.141)	(0.717)	(0.177)	(0.178)	(0.176)	(0.219)
Sex	0.411^{**}	0.368^{*}	0.328^{*}	0.322^{*}	0.311^{*}	0.051
	(0.141)	(0.160)	(0.152)	(0.153)	(0.153)	(0.150)
Age	-0.583^{***}	-0.577^{***}	-0.576^{***}	-0.573^{***}	-0.568^{***}	-0.376^{**}
	(0.114)	(0.124)	(0.126)	(0.127)	(0.125)	(0.128)
Education	0.080	0.000	-0.048	-0.040	-0.077	$-0.245^{\hat{}}$
	(0.133)	(0.192)	(0.140)	(0.140)	(0.140)	(0.147)
Income	0.065	0.064	0.046	0.040	0.054	-0.011
	(0.142)	(0.144)	(0.147)	(0.147)	(0.146)	(0.141)
Obey Law		-0.770	-1.514^{***}	-1.521^{***}	-1.516^{***}	-1.358^{***}
		(1.355)	(0.142)	(0.143)	(0.144)	(0.200)
$\mathcal{L}$	-92	-91	-120	-126	-127	-141
$n = 10{,}000$ (for all)						

Table 3: Unweighted Obey Law Models: $\kappa_{\mathrm{C}}^n = 1, \kappa_{\mathrm{M}}^n = 1$

by the inclusion of 'obey law'. Instead, the effect sizes of intercept and sex are diminished, hinting that these are the active areas of the covariance between 'obey law' and the socio-demographic covariates. This is confirmed in the 'No Obey' model, in which the mere inclusion of the imputed 'obey law' covariate, fitted to the original likelihood, impacts the original coefficients; the lack of the 'obey law' likelihood renders the covariate insignificant. Not surprisingly, once we engage the 'obey law' likelihood, the covariate attains prominence in both significance and effect size.

We see the same pattern when we weight the U.S.-based data (i.e. $\kappa_{\mathrm{C}}^n = 8, \kappa_{\mathrm{M}}^n = 2$) in Table 4. Curiously, 'education' attains prominence in the 'obey law' models, partly as a response to the jump in the intercept, and perhaps partly due to the inherent connection between it and 'obey law'; they are significantly correlated. Furthermore, the increased weight on the Collins et al. likelihood lends to the increase in effect size for the 'education'

Predictor	Basic	No Obey	No W	Prior	M.I.	Mean
Intercept	−0.415***	0.034	0.519**	0.534***	0.550***	0.837***
	(0.108)	(0.495)	(0.158)	(0.159)	(0.159)	(0.191)
Sex	0.339***	0.285**	0.228*	0.225*	0.217*	−0.205ˆ
	(0.087)	(0.106)	(0.096)	(0.097)	(0.097)	(0.114)
Age	−0.531***	−0.515***	−0.517***	−0.515***	−0.506***	−0.268**
	(0.089)	(0.093)	(0.099)	(0.099)	(0.098)	(0.103)
Education	−0.045	−0.141	−0.253**	−0.252**	−0.291***	−0.634***
	(0.070)	(0.133)	(0.080)	(0.081)	(0.083)	(0.111)
Income	−0.026	−0.027	−0.031	−0.032	−0.023	−0.026
	(0.055)	(0.056)	(0.056)	(0.056)	(0.056)	(0.056)
Obey Law		−0.796	−1.626***	−1.631***	−1.642***	−1.886***
		(0.862)	(0.148)	(0.149)	(0.151)	(0.205)
$\mathcal{L}$	−324	−324	−341	−342	−343	−345
$n = 10{,}000$ (for all)						

Table 4: Weighted Obey Law Models: $\kappa^n_{\mathrm{C}} = 8, \kappa^n_{\mathrm{M}} = 2$

coefficient; their coefficient is comparable to what we are observing here.[18] Once, again these patterns recur with the non-linear 'education' covariate, which we highlight:

	Unweighted: $\kappa^n_{\mathrm{C}} = 1, \kappa^n_{\mathrm{M}} = 1$		Weighted: $\kappa^n_{\mathrm{C}} = 8, \kappa^n_{\mathrm{M}} = 2$	
Predictor	Basic	No W	Basic	No W
Intercept	−0.842***	0.076	−0.230ˆ	0.582***
	(0.145)	(0.189)	(0.118)	(0.155)
Sex	0.383**	0.375**	0.377***	0.291**
	(0.116)	(0.131)	(0.088)	(0.095)
Age	−0.598***	−0.562***	−0.522***	−0.521***
	(0.110)	(0.122)	(0.091)	(0.099)
|Education − 1|	−0.220ˆ	−0.264*	−0.472***	−0.522***
	(0.113)	(0.125)	(0.113)	(0.123)
Income	0.141*	0.008	−0.028	−0.090ˆ
	(0.070)	(0.081)	(0.047)	(0.049)
Obey Law		−1.515***		−1.496***
		(0.142)		(0.137)
$\mathcal{L}$	−90	−118	−316	−337
$n = 10{,}000$ (for all)				

[18] See Appendix A.4.

In fact, the fit of the weighted, non-linear eduation model now permits the significance of 'income'. In Table 5, we review the spread of error probabilities from each of the 'obey law'

κ^n_{C}	κ^n_{M}	Model	V.Ed	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	1	Basic		11.0	20.4	27.7	26.9	32.3	51.3
1	1	No Obey		7.8	17.2	26.0	27.0	36.9	57.0
1	1	No W		5.3	13.8	24.3	28.4	43.3	61.8
1	1	Prior		5.6	14.2	24.8	28.9	44.1	61.9
1	1	M.I.		5.2	14.1	25.8	28.9	43.8	63.2
1	1	Mean		5.9	17.0	31.7	30.0	38.0	53.8
8	2	Basic		14.9	24.8	30.4	30.7	36.8	48.1
8	2	No Obey		8.9	20.9	30.1	30.7	39.1	57.9
8	**2**	**No W**		**4.5**	**14.7**	**30.5**	**31.4**	**46.1**	**67.8**
8	2	Prior		4.6	14.8	30.8	31.7	46.4	68.1
8	2	M.I.		4.5	14.9	30.9	31.7	46.9	68.3
8	2	Mean		2.8	16.8	32.6	32.5	46.4	68.7
1	1	Basic	✓	7.7	20.1	26.6	27.1	33.7	59.6
1	1	Prior	✓	4.3	13.1	27.1	29.1	41.9	63.3
8	2	Basic	✓	8.6	21.8	30.8	31.0	39.3	53.7
8	2	Prior	✓	3.1	16.8	29.0	32.0	47.7	70.9

Table 5: Summary Statistics for Error as % across Obey Law Models. *Statistics are presented as percentages (%). V.Ed refers to the non-linear treatment of Education. Our flagship model is bold-typed:* $\kappa^n_{\mathrm{C}} = 8, \kappa^n_{\mathrm{M}} = 2$, model='No W'.

models. As we expect, the inclusion of both the Wahlund population and 'obey law' fits (in the 'Prior' and 'M.I.' models) should increase the mean and maximum error probabilities. However, what we do not necessary expect is for both the mean error and range of probabilities to increase when we substitute the Wahlund fit with the 'obey law' fit (i.e. going from the 'Basic' model to 'No W'), suggesting that 'obey law' is considerably informative to error prediction.

8.2 Predicting with the GSS

Alternatively, we can construct generic error models (with and without 'obey law') using the GSS data aligned to PUMS bins. While there exist some differences in the marginal proportions of the socio-demographic categories, this does not necessarily mean the predictive models will differ. So, we perform an identical meta-analytic fit as we perform in Table 1, using GSS data from the survey years which we employed in the 'obey law' analysis of Appendix A.6.1. In Table 6, we find sufficient parity between the GSS-based models and the analogous PUMS-based generic error models, with one slight exception. The non-linear 'education' coefficient for the unweighted model, here, is insignificant, whereas in the PUMS-based model, it was not; hence, the differential covariance can affect the model results. We

	Unweighted: $\kappa^n_{\mathrm{C}} = 1, \kappa^n_{\mathrm{M}} = 1$		Weighted: $\kappa^n_{\mathrm{C}} = 8, \kappa^n_{\mathrm{M}} = 2$	
Predictor	/ Ed.	V Ed.	/ Ed.	V Ed.
Intercept	-0.653^{***} (0.183)	-0.632^{***} (0.156)	-0.262^{*} (0.113)	$-0.202^{\hat{}}$ (0.114)
Sex	0.460^{***} (0.129)	0.470^{***} (0.116)	0.343^{***} (0.085)	0.382^{***} (0.086)
Age	-0.619^{***} (0.105)	-0.596^{***} (0.110)	-0.646^{***} (0.089)	-0.513^{***} (0.094)
Education	-0.009 (0.104)		-0.092 (0.067)	
|Education − 1|		-0.074 (0.113)		-0.430^{***} (0.104)
Income	0.017 (0.109)	0.012 (0.069)	-0.042 (0.042)	-0.046 (0.038)
$\mathcal{L}$	−91	−91	−320	−312
$n = 16{,}488$ (for all)				

Table 6: GSS-Based Generic Error Models. *We distinguish the standard, linear treatment of 'education',* '/ Ed.' *from the non-linear,* 'V Ed.'.

now include 'obey law' in the GSS meta-analysis looking at both the cases when we omit and include the fit to 'obey law' correlation, 'No Obey' and 'No W', respectively:

	Unweighted: $\kappa^n_{\mathrm{C}} = 1, \kappa^n_{\mathrm{M}} = 1$		Weighted: $\kappa^n_{\mathrm{C}} = 8, \kappa^n_{\mathrm{M}} = 2$	
Predictor	No Obey	No W	No Obey	No W
Intercept	-0.040 (0.407)	0.040 (0.204)	-0.103 (0.383)	0.425^{**} (0.135)
Sex	0.376^{**} (0.145)	0.353^{*} (0.144)	0.326^{**} (0.109)	0.229^{*} (0.093)
Age	-0.610^{***} (0.119)	-0.595^{***} (0.116)	-0.638^{***} (0.093)	-0.621^{***} (0.096)
Education	-0.266 (0.194)	$-0.216^{\hat{}}$ (0.119)	-0.174 (0.156)	-0.382^{***} (0.083)
Income	0.036 (0.111)	0.027 (0.112)	-0.025 (0.042)	-0.016 (0.041)
Obey Law	-1.586 (1.030)	-1.499^{***} (0.124)	-0.417 (0.841)	-1.616^{***} (0.133)
$\mathcal{L}$	−89	−113	−318	−328
$n = 4{,}261$ (for all)				

The pattern mirrors our earlier findings with the PUMS-based 'obey law' error models. The additional fit is required for the 'obey law' covariate to exhibit significance; interestingly, its effect size in the unweighted condition, ~ -1.59 is just as large as it is in the weighted, fitting condition ~ -1.62 suggesting that direct use of the data better specifies the predictive process, as we would expect in a comparison between actual and imputed data. However, the high degree of similarity between the earlier imputed models and these models validates the former approach. Another difference worth mentioning is potency of 'age'. Finally, we examine the same models, but with the non-linear treatment of 'education':

	Unweighted: $\kappa^n_{\mathrm{C}} = 1, \kappa^n_{\mathrm{M}} = 1$		Weighted: $\kappa^n_{\mathrm{C}} = 8, \kappa^n_{\mathrm{M}} = 2$	
Predictor	No Obey	No W	No Obey	No W
Intercept	-0.456	0.217	$-0.392^{\hat{}}$	0.443^{***}
	(0.395)	(0.192)	(0.225)	(0.134)
Sex	0.483^{***}	0.483^{***}	0.429^{***}	0.317^{***}
	(0.119)	(0.129)	(0.091)	(0.092)
Age	-0.577^{***}	-0.523^{***}	-0.532^{***}	-0.425^{***}
	(0.114)	(0.122)	(0.102)	(0.102)
$\lvert\text{Education} - 1\rvert$	-0.062	-0.149	-0.414^{***}	-0.507^{***}
	(0.114)	(0.123)	(0.107)	(0.113)
Income	-0.026	$-0.126^{\hat{}}$	-0.030	-0.077^{*}
	(0.098)	(0.074)	(0.040)	(0.037)
Obey Law	-0.338	-1.469^{***}	0.308	-1.435^{***}
	(0.630)	(0.125)	(0.392)	(0.122)
$\mathcal{L}$	-90	-114	-311	-330

$n = 4{,}261$ (for all)

Once again, with the exception of some differences in the effect sizes of some covariates (such as 'age'), these GSS-based models compare well to the PUMS-based model, and with slightly greater emphasis on 'income'.

9 Conclusion and Discussion

In this paper, we first sought to meta-analytically infer a unified model predicting taxpayer intentional error from four socio-demographic and one attitudinal variable. This endeavor not only required a meta-analytic fit from several empirical sources, but also substantial modification of the reported statistics such that they were aligned with the PUMS covariates; the sheer volume of analysis in the appendices is a testament to the complexity of this process.

We find enough consistency in most of the empirical marginal statistics to construct a moderately predictive model, demonstrated by the significance of some of the model estimates. However, the disparate nature of populations that our source studies surveyed infuses this endeavor with imprecision as evidenced by some inconsistent patterns between the empirical error commission and the main covariates. Ultimately, we find certain covariates,

namely 'sex' and 'age', to consistently retain their significance across the different variants of generic error models we inferred, while the predictive power of 'education' and 'income' remain tenuous. Furthermore, much of the power from any of these models is lost when we attempt to draw out the predictive variance into eight (7+1) separate line item models, which contains only a handful of significant coefficients. Still, our assumption of independence is likely to be found to be partly responsible for this lack of predictive effectiveness; we will revisit this issue in future writings.

As our models are based off a nationally representative United States population sample, they can be applied towards measuring non-compliance behavior in specific U.S. subpopulations, such as different cities or regions, which might vary socio-demographically. Such comparative analysis would undoubtedly be informative to policy-makers who seek effective interventions to reduce the tax gap.

In the next stage, we intend to supplement our models with additional findings, particularly those that reflect the evasion behavior of recent, U.S. populations as well as additional important predictors, which we mention back in Section 2.2, such as marital status, use of paid preparer, self-employment (as a predictor), and social network effects.[19]

Appendix

These appendices primarily include the process by which we align reported statistics to the PUMS covariates as well as additional, necessary inference, such as combining error rates from specific kinds of non-compliance into a single rate and the inferring correlations between error and intervening predictors, such as opportunity and 'obey law'; this material is covered in Appendix A. Appendix B reviews some of the prior work on the role of social influence in intentional error; these findings, while not explicitly used in the analysis, will inform our future work. Appendix C details the priors (or prior distributions) we imposed on our line item model coefficients, to insure we were able converge on a model.

Appendices D–E detail some earlier work, which we deem to be 'background material'. These latter sections illustrate earlier approaches, some of which are more precise, but less practical, than the methods employed in the body of the paper, and others which are too imprecise, but demonstrates the evolution of our methodology.

A Adjusting Empirical Findings for Meta Analysis

In order to fit our intentional error model to all data sources, we need to align them such that we are fitting to consistent covariates. Primarily, educational categories across the source nations require mapping to U.S. categories and income of respondents in these different countries need to be converted into 2000 U.S. dollars.

[19]For a discussion on social network effects on tax evasion found in the literature, refer to Appendix B.

A.1 Houston and Tran (2000)

Houston and Tran report under-reporting of income and over-reporting of deductions separately. From their reports of non-compliance, we can infer some estimate of joint error commission, which we require for our intentional error model that focuses on, instead, *some* commission of either or both forms of non-compliance:

$$\begin{aligned} \text{Prop. under-reporting income} &= 5.5\% \\ \text{Prop. over-reporting deductions} &= 6.5\% \\ \text{Prop. committing either} &= 7.1\% \end{aligned}$$

These proportions break down as follows:

		Over-reporting Deductions		
		No	Yes	Total
Under-reporting Income	No	0.929	0.016	
	Yes	0.006	0.049	0.055
	Total		0.065	0.071

And, with a sample size $n = 284$, the table yields a rather high correlation of $\rho_{\text{RR}} = 0.790\,(p < 0.001)$ between under-reporting income and over-reporting deductions. When we examine the authors' direct-questioning (DQ) responses, we obtain a much lower correlation, $\rho_{\text{DQ}} = 0.299\,(p < 0.001)$. Note that Wahlund (1992) finds the correlation between the two behaviors to be $\rho_{\text{W}} = 0.16$. For now, we will use the mean $\overline{\rho} = 0.416$ as our estimate of joint non-compliance behavior for the Houston and Tran data.

The following table reiterates the Houston/Tran estimates and summarizes covariate categories, adjusted to reflect both any level non-compliance as well as further concordance to our data (e.g. 2000 U.S. dollars for income).[20]

[20]For education categories, the Australian "non-tertiary" and "tertiary" categories map to "no college" and "at least some college". For income, we employ the AUS–to–USD exchange rate in 1992, averaged over all months, which is 0.680, and the CPI conversion factor for 1992–to–2000 USD, which is 1.192.

Covariate	Covariate Categories	Proportion Committing ... Under-reporting	Over-claiming	Either/ Both	σ_{Any}	n
Sex	Female	0.083	0.112	0.150	(0.0532)	101
	Male	0.040	0.039	0.062	(0.0355)	180
Age	18–45	0.063	0.102	0.128	(0.0494)	111
	46+	0.050	0.041	0.071	(0.0370)	170
Education	No college	0.064	0.094	0.122	(0.0439)	138
	At least some college	0.048	0.038	0.067	(0.0403)	142
Income	$\leq$ \$30,781.47 USD	0.069	0.036	0.083	(0.0370)	175
	$>$ \$30,781.47 USD	0.036	0.022	0.046	(0.0456)	104
Self-Employed	No	0.036	0.075	0.088	(0.0411)	144
	Yes	0.165	0.092	0.197	(0.0629)	79
Paid Preparer	No	0.008	0.008	0.013	(0.0582)	57
	Yes	0.062	0.074	0.105	(0.0339)	223

The above table employs a $\rho = 0.416$, the mean of the correlations of Wahlund, Houston RR, and DQ. The inclusion of DQ is debatable and an equally valid construction comprises Wahlund and Houston RR, or just Houston RR. The randomized-response data do not exhibit typical standard deviations for proportions. Instead, due to the nature of the questioning, they incur wider uncertainty. We take the mean of the biased variance, as measured by Eq. 11 in Appendix D.1, for both under-reporting income and over-reporting deductions.[21] Interestingly, in this data, females commit more acts of non-compliance than males, which is opposite of what the other studies have found. Furthermore, the proportion of respondents committing some non-compliance falls far below similar proportions reported in the other studies. The omission of other types of non-compliance in this study cannot account for this difference; that is, even the rates of specific types of non-compliance are lower for this study than say Mason and Calvin's work, suggesting some other hidden variable is responsible; perhaps Australians are far more honest than Europeans or Americans?

Our intentional error likelihood model will strive to fit coefficients taking into account the above proportions for each category, along with estimates from the other studies, which will be incorporated in a similar fashion; we detail this in Section 4. For instance,

$$\text{logit}[\overline{p}(y > 0 \,|\, x_{\text{Sex}} = \text{`Female'})] \sim \text{N}\big(\mu = \text{logit}(0.150), \sigma^2 = 0.417^2\big)$$
$$\text{logit}[\overline{p}(y > 0 \,|\, x_{\text{Sex}} = \text{`Male'})] \sim \text{N}\big(\mu = \text{logit}(0.062), \sigma^2 = 0.610^2\big)$$

where σ^2 for the Houston marginals is derived using the standard deviation of a logit for Any (or Either/Both):

$$\sigma^2 = \left(\frac{\sigma_{\text{Any}}}{p(\text{Any}) \cdot (1 - p(\text{Any}))}\right)^2$$

[21] We might also examine a maximum-likelihood estimate, which is likely to be equivalent to some weighted mean.

A.2 Mason and Calvin (1978)

Mason and Calvin administered a survey of tax evasion behavior to $n = 800$ adults in the state of Oregon.

		Mean covariate for any violation	
Covariate	Bins	Non-Evader	Evader
Sex	1=male, 2=female	1.56	1.43
Age	groups 1 to 6	4.53	4.02
Income	groups 0 to 13	8.52	7.96

We can easily transform the gender breakdown into rates of error for each sex for the 'Any Violoation' category. Given that we have $\mu_{\text{Non-Evasion}} = 1.56$ and $\mu_{\text{Evasion}} = 1.43$, where Male = 1 and Female = 2, and $p(y > 0) = 0.242$, we seek the breakdown proportions (a, b, c, d):

	Sex	
	Female	Male
No Evasion	a	b
Evasion	c	d

where

$$\begin{aligned} \frac{2a + 1b}{a + b} &= 1.56 \\ \frac{2c + 1d}{c + d} &= 1.43 \\ c + d &= 0.242 \\ a + b + c + d &= 1 \end{aligned}$$

We obtain the solution:

$$a = 0.42448, b = 0.33352, c = 0.10406, d = 0.13794$$

Hence,

$$\begin{aligned} p_{\text{Male}} &= d/(b + d) = 0.13794/(0.33352 + 0.13794) = 0.2925805 \\ p_{\text{Female}} &= c/(a + c) = 0.10406/(0.42448 + 0.10406) = 0.1968820 \\ n_{\text{Male}} &= 800(a + c) = 423 \\ n_{\text{Female}} &= 800(b + d) = 377 \end{aligned}$$

We now have the within category error rates for gender:

Covariate	Bins	% Tax Evasion	n
Sex	Female	19.7	377
	Male	29.3	423

Converting the other Mason/Calvin covariates, particularly Age and Income, is not trivial since, the paper employs custom bins, which the authors neglected to document. One method of obtaining error estimates is to employ gradient descent on a simulated data set, as we do in the Wahlund analysis in Appendix A.5. We will attend to this exercise at a later date.

A.3 Vogel (1974)

Vogel studied tax evasion behavior and attitudes for a sample of the Swedish population. Here, we reproduce the portions of his data which are relevant to our work. Fortunately, his paper reports marginal statistics which we can directly employ in this paper.

Covariate	Bins	% Tax Fraud	n
Sex	Female	21.7	506
	Male	32.3	709
Age	20-29	38.8	288
	30-39	31.5	230
	40-49	29.5	258
	50-59	19.6	226
	60-70	16.5	214
Education	Less Than H.S.	24.4	649
	High School	33.9	287
	College	30.1	276
Self-Employment	No	27.9	967
	Yes	37.1	106

A.4 Collins et al. (1992)

The authors report marginal proportions from each of the socio-demographic covariates; however, their proportions (reported here as %) and those of our PUMS sample substantially differ:

	Sex		Age			
Source	Female	Male	<25	25-44	45-65	65+
Collins	39	60	7	44	31	18
PUMS	51	49	13	42	30	14

	Education				
Source	<H.S.	H.S.	Some college	College graduate	Post-graduate
Collins	9	23	32	19	17
PUMS	20	27	23	22	9

	Income (K = \$1,000)					
Source	≤15K	(15K,30K]	(30K,50K]	(50K,75K]	(75K,100K]	>100K
Collins	15	36	32	10	3	4
PUMS	50	27	14	5	1	2

As one can easily verify, the categorical breakdown of Collins sample departs from our PUMS sample. We explored subsampling from the PUMS by assigning weights such that drawn subsamples yield breakdowns identical to the Collins data; however, we found that this

process yielded a variety of covariance structures. Hence, we choose the weighting solution that yields a similar covariance as the PUMS data, thereby producing coefficients close to what we would if we use the uncorrected PUMS sample.

We employ simulating annealing to find weights for each unique combination of our discrete covariates, which correspond to a sampling of which marginal proportions correspond to those of the Collins data.[22] A δ parameter controls the rate at which we alter the weights (via multiplication or division) We offer a statistical summary of our ten sampled sets of weights for two settings of δ:

Sample	δ	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Wide	0.90	0.0424	0.4783	1.0000	2.8010	2.0910	114.6000
Restricted	0.99	0.1740	0.5527	0.9044	1.6420	1.6360	27.8500

Each sample weight assigned to a unique covariate combination is divided in the PUMS sample by the corresponding frequency of that combination. For instance, if the weight for the combination {Sex = 0, Age = 1, Educ = 1, Income = 5} is 1.3 and there are 13 such individuals, then the weight for each individual is $\frac{1.3}{13}$ or 0.1.

A.4.1 Zero-Inflated Poisson

Collins et al. report that 50% of their respondents commit some form of evasion while the mean number of acts is 1.65. As it is, a unimodal distribution (i.e. single peak) cannot account for these results. So, we naturally assume a zero-inflated Poisson, or ZIP, distribution for the counts of non-compliant acts for their respondents. The maximum-likelihood estimation for obtain θ and λ which, respectively, denote the probability of some error and the count of errors, given some error, is detailed here:

$$p \sim \mathrm{N}\left(0.5, \sigma^2 = \frac{(1-0.5)\cdot 0.5}{240}\right)$$
$$\mu \sim \mathrm{N}\left(1.65, \sigma^2 = \frac{1.875^2}{240}\right)$$

where

$$\begin{aligned} p &= 1 - \theta + \mathrm{Pois}(0|\lambda) \\ &= 1 - \theta + e^{-\lambda} \end{aligned}$$

and

$$\mu = (1-\theta)\cdot \sum_{n=1}^{\infty} n \cdot \frac{e^{-\lambda}\lambda^n}{n!}$$

[22] Instead of finding weights for all 240 combinations, we opt to weight only the 190 combinations found in our PUMS sample.

We find the maximum-likelihood at $\theta = 0.478$ and $\lambda = 3.161$ and draw $n = 9482$ count samples from the ZIP, using the multinomial, and randomly assign these counts of non-compliance to the PUMS sample; the PUMS sample is smaller as we ignore teenagers under the age of eighteen. We employ simulated annealing to find maximally fitting permutations of these assignments. In order to asssess the fit, we regress the assigned data, y, similarly to Collins et al, with:[23]

$$y = \beta_0 + \beta_1 x_{\text{Sex}} + \beta_2 x_{\text{Age}} + \beta_2 x_{\text{Educ}} + \beta_3 x_{\text{Income}}$$

and fit to normal likelihood surrounding each of the non-intercept coefficients:

$$\hat{\boldsymbol{\beta}}_{-0} \sim \text{N}(\boldsymbol{\beta}_{-0}, \boldsymbol{\sigma}^2_{-0})$$

where (as Collins et al. report) $\boldsymbol{\beta}_{-0} = (0.279, -0.089, -0.392, -0.169)$ and (we assume) $\boldsymbol{\sigma}_{-0} = \boldsymbol{\beta}_{-0} \cdot (0.8, 0.8, 0.5, 0.8)$; we base the standard errors on the reported significance of each coefficient, a naïve assumption to be sure.

Since these results are applicable to a Collins weighted sample, we need to reverse the weights. So, we infer a Collins error model with data that reflect the PUMS, not Collins, marginal, by employing a weighted logistic regression, using converse weights:

Variable	Restricted	Wide	Both	Both ($n = 240$)
Intercept	0.920***	0.955***	0.937***	1.005
	(0.134)	(0.228)	(0.188)	(0.739)
Sex	0.224*	0.239	0.231ˆ	0.272
	(0.089)	(0.157)	(0.127)	(0.478)
Age	−0.064	−0.081	−0.073	−0.059
	(0.054)	(0.101)	(0.081)	(0.325)
Educ	−0.339***	−0.350***	−0.345***	−0.399ˆ
	(0.035)	(0.059)	(0.049)	(0.222)
Income	−0.143***	−0.140**	−0.142***	−0.150
	(0.027)	(0.053)	(0.042)	(0.209)
n_{draws}	200	200	400	400
n_{sample}	9482	9482	9482	240

Naturally, the estimates and their significance will appear relatively similar to the original Collins numbers, as the categorizations between the PUMS sample and Collins data are similar.

[23] We cannot control for the other covariates in the Collins et al model since they do not occur in our PUMS data. Hence, we resort to treating them independent from our four key socio-demographic covariates and assume the Intercept will subsume their effects.

A.4.2 Mixture Model

Alternatively, we can tease out the process that produces the zero-inflation into a separate model.

$$\begin{aligned}
\text{logit}[p(y_c = 1)] &= \alpha_0 + \alpha_1 x_{\text{Sex}} + \alpha_2 x_{\text{Age}} + \alpha_2 x_{\text{Educ}} + \alpha_3 x_{\text{Income}} \\
\log(y) &= \gamma_0 + \gamma_1 x_{\text{Sex}} + \gamma_2 x_{\text{Age}} + \gamma_2 x_{\text{Educ}} + \gamma_3 x_{\text{Income}} \\
\mu &= p(y_c = 1) \cdot \log(y) \\
p &= p(y_c = 1)(1 - e^{-\lambda}) \\
\mu &= \beta_0 + \beta_1 x_{\text{Sex}} + \beta_2 x_{\text{Age}} + \beta_2 x_{\text{Educ}} + \beta_3 x_{\text{Income}} \\
\hat{\boldsymbol{\beta}} &\sim \text{N}(\boldsymbol{\beta}_{-0}, \boldsymbol{\sigma}_{-0}{}^2)
\end{aligned}$$

We defer this analysis to future writings.

A.4.3 Intercept

Finally, in lieu of the intercept, β_0, we eventually fit the overall rate of 50% of error commission found by Collins et al. in the main likelihood model:

$$\text{logit}[\overline{p}(y > 0)] \sim \text{N}\big(\mu = \text{logit}[0.5], \sigma^2 = [\mu(1-\mu) n_{\text{C}} \kappa_{\text{C}}^n]^{-1}\big)$$

A.5 Wahlund (1992)

In his paper, (Wahlund, 1992) reports a structural equation model (in the form of path coefficients) that includes two covariates which directly predict tax evasion: 'opportunity' to evade taxes and 'attitudes to crime'. We offer some approaches towards isolating the correlation between these covariates and intentional error.

A.5.1 Evasion and GSS Obeying the Law

The author reports correlative effects among a battery of covariates in a (large) structural equation model (SEM). The 'attitudes to crime' is found to be directly linked to tax evasion with a correlation $\rho = -0.21$. Here, we employ the reported path coefficient as a direct correlation when, in fact, it is being controlled by other covariates; we will later attempt to infer the actual correlation, so for, now the results obtained in this section are slightly inaccurate and offered for the purposes of illustrating some methods we use throughout the paper. In incorporating this finding into our model, we make two assumptions:[24]

1. Correlations hold when both items are scaled as binary.

2. Attitudes to crime is sufficiently correlative to the GSS Obey Law item.

[24] We will address these assumptions later in this appendix.

In addition to the correlation, we also know the marginal probabilities: 14% for under-reporting income and 43% for absolute adherence to the law; the latter proportion is obtain from the GSS; note, we are not addressing general tax evasion, but a specific kind of non-compliance. We also estimate our sample size $n \approx 600$.[25] With this information we can construct and solve the following contingency table:

		Tax Evasion	
		0	1
Obey Law	0	$m_{00} = n \cdot p_{00}$	$m_{01} = n \cdot p_{01}$
	1	$m_{10} = n \cdot p_{10}$	$m_{11} = n \cdot p_{11}$

We know that:

$$p_{10} + p_{11} = 0.43$$

and

$$p_{01} + p_{11} = 0.14$$

We note the marginal sums: $n_1 = m_{10} + m_{11}$ and $m_1 = m_{01} + m_{11}$. Thus far, we have three unknowns and only two constraints/equations.[26] The third constraints comes in the form of the Pearson correlation:

$$\rho = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

for which we have all the pieces:[27]

$$\begin{aligned}
\mu_X &= n_1/n \\
\mu_Y &= m_1/n \\
\sigma_X^2 &= \frac{\sum_{i=0}^{1}(m_{i0} + m_{i1})(i - \mu_X)^2}{n} \\
\sigma_Y^2 &= \frac{\sum_{i=0}^{1}(m_{0i} + m_{1i})(i - \mu_Y)^2}{n} \\
E[\ldots] &= \sum_{i=0}^{1}\sum_{j=0}^{1} m_{ij}(i - \mu_X)(j - \mu_Y)
\end{aligned}$$

The constraints reduce to the following:[28]

$$m_{10} = n_1 - \frac{n_1 m_1}{n} - \left(n\rho \cdot \sqrt{\frac{m_1(n - m_1)}{n^2}} \cdot \sqrt{\frac{n_1(n - n_1)}{n^2}} \right)$$

[25]The SEM is based on data obtained from one of the four surveys reported in the paper, the size of which was not reported. Instead, the author reports the range of sample sizes, the mean of which was originally computed as 600. This estimate is incorrect and latter supplanted with 430. However, the use of the incorrect n does not affect the findings in this section.

[26]Recall that $p_{00} + p_{10} + p_{01} + p_{11} = 1$, and so $m_{11} = n - (m_{00} + m_{10} + m_{11})$.

[27]The Pearson correlation employs the population standard deviation rather than the sample standard deviation; hence, n instead of $n - 1$ appears in the denominator.

[28]We resort to Mathematica to perform the algebraic simplification for us.

and we obtain the following rounded contingency table:

		Tax Evasion 0	Tax Evasion 1
Obey Law	0	272	70
	1	244	14

which offers a $\rho = -0.2146$. Our proportions for tax evasion for each Obey Law category is $\frac{70}{272+70} = 0.205$ and $\frac{14}{244+14} = 0.054$, which can also be described by a logistic regression

$$\text{logit}[p(y > 0)] = \tau_0 + \tau_1 \cdot x_{\text{ObeyLaw}}$$

with the following parameters:

$$\tau_0 = -1.357, \sigma_{\tau_0} = 0.134$$
$$\tau_1 = -1.501, \sigma_{\tau_1} = 0.306$$

A.5.2 Multiple Types of Evasion

The author also offers the following data:

7% admitted having made illegal deductions.
14% admitted to not having declared some income.
The correlation between the two behaviors is 0.16.

We can employ the same deductive method as we used in the previous section to infer the contingency table for the two non-compliant behaviors and obtain:

		Income Evasion 0	Income Evasion 1
Illegal Deduction	0	1162	66
	1	165	34

We calculate the Pearson correlation for the above inferred table to be $\rho = 0.159$. The joint non-compliance rate is $\frac{34}{1427} = 0.024$, which is more than double the expected rate of $0.07 \times 0.14 = 0.010$, assuming independence.

A.5.3 Opportunity vs. Evasion

When we employ the path coefficient (stnadardized) between opportunity and evasion as a correlation, $\rho = 0.21$, and assume that those with no opportunity do not evade:

		Opportunities: None	Opportunities: Some
		0	1
No Evasion	0	a	b
Evasion	1	0	c

where $c = 0.37$ (the correct error rate for any kind of tax evasion), according to Wahlund's findings; hence $a + b = 1 - 0.37 = 0.63$. The algebraic solution is $a = 0.07$ and $b = 0.56$, giving us an evasion rate of $\frac{0.37}{0.37+0.56} \approx 0.40$. Accordingly, this solutions yields 93% of the population as having some opportunity of error. However, if we employ the inferred $\rho = 0.329$ (obtained later in this appendix), we instead obtain $a = 0.16$ and $b = 0.47$ which leads to an intentional error rate of $\frac{0.37}{0.37+0.47} = 0.44$, given some opportunity. The proportion of opportunity from this solution is a lower $0.37 + 0.47 = 0.84$ which is surprisingly concordant with the line-item-based opportunity rate from our PUMS data: 0.825.

A.5.4 Obey Law and Wahlund Crime

As mentioned earlier, Wahlund finds in his structural model that one of the five variables directly impacting tax evasion is "attitudes to crime" variable, which measures a respondent's leniency towards crime:

> Attitudes to crime: The more lenient attitudes towards crime, the lower the value.

We assume that this variable sufficiently mirrors the GSS' Obey Law(s) items and attempt to infer this proxy Obey Law's impact on tax evasion using Wahlund's variable. Wahlund reports, in his path diagram, that the impact of "attitudes to (c)rime" on tax (e)vasion is $\rho = -0.21$, controlling for four other covariates: "(o)pportunity", "(t)ax avoidance", "(p)erceived opportunity", and "perceived (r)isk"; hence, the effect of solely "crime" (or "obey law") on evasion ought to be noticeably stronger than -0.21. Wahlund's path diagram (Fig. 3) contains enough linkages around these five covariates for us to be able to construct a lower bound on the direct effect, but not enough for us to triangulate the exact, full correlation, at least not through analytic means.

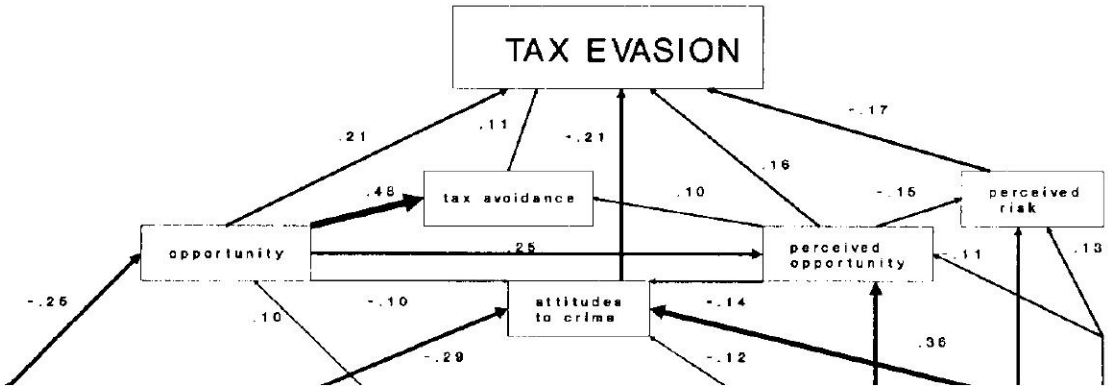


Figure 3: Path Diagram from Wahlund (1992). *The five directly effecting covariates are shown along with the significant paths. Reprinted without permisson.*

Inferring the lower bound on the actual "crime/tax evasion" correlation, as best as the data will allow, requires several steps. We use the following notation to enhance readability:

1. $(x, y|z)$ denotes the reported path prediction (i.e. the path coefficient shown in the diagram) from x to y controlling for z ($x \to y \leftarrow z$). We will sometimes use (x, y) as shorthand implying the path coefficient from x to y under the assumption that other covariates are controlled for. This is identical to standardized regression model $y = b_x \cdot x + b_z \cdot z$.
2. $\widehat{(x, y)}$ is our inferred total correlation between x and y.

A naïve lower bound considers only the paths from "crime" to "evasion":

$$\begin{aligned}\widehat{(c, e)} &= (c, e) + (o, c)(o, e) + (o, c)(o, t)(t, e) + (p, c)(p, e) + (p, c)(p, r)(r, e) \\ &= -0.21 + (-0.10)(0.21) + (-0.10)(0.48)(0.11) + (-0.14)(0.16) \\ &\quad + (-0.14)(-0.15)(-0.17) \\ &= -0.26225\end{aligned}$$

We can augment this approach by recognizing that the true correlations between "crime" and the other covariates, x, (i.e. $\widehat{(c, x)}$), reside near the path coefficient. Furthermore, we assume that these correlations do not diminish but instead likely increase within their respective valences, unless they are close to zero. For example, $(o, c) = -0.10$ implies that $\widehat{(o, c)} < -0.10$. We specify a mean and variance around each adjustment:

$$\begin{aligned}\widehat{(c, o)} &= (o, c) - \mathrm{N}(\mu, \sigma^2) \\ \widehat{(c, p)} &= (p, c) - \mathrm{N}(\mu, \sigma^2)\end{aligned}$$

Since both paths are negative, we assume the true correlation will increase in degree but not sign. For the other covariates lacking a path from c, we resort to using the indirect path/correlation:

$$\begin{aligned}\widehat{(c, t)} &= (c, o)(o, t) - \mathrm{N}(\mu, \sigma^2) \\ \widehat{(c, r)} &= (c, p)(p, r) + \mathrm{N}(\mu, \sigma^2)\end{aligned}$$

Finally, we can compute the total effect $\widehat{(c, e)}$:

$$\widehat{(c, e)} = (c, e) + \widehat{(c, o)}(o, e) + \widehat{(c, p)}(p, e) + \widehat{(c, t)}(t, e) + \widehat{(c, r)}(r, e)$$

The known values, obtained from Figure 3, are:

$$\begin{aligned}(o, c) &= -0.10 & (o, e) &= 0.21 \\ (p, c) &= -0.14 & (p, e) &= 0.16 \\ (o, t) &= 0.48 & (t, e) &= 0.11 \\ (p, r) &= -0.15 & (r, e) &= -0.17 \\ & & (c, e) &= -0.21\end{aligned}$$

When we employ $\mu = 0.05, \sigma = 0.04$, we obtain a distribution summarized as follows:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
−0.3368	−0.3036	−0.2946	−0.2950	−0.2857	−0.2494

Hence, our estimated, total correlation between 'attitudes to crime' (i.e. obey law) and tax evasion is the mean $\rho = -0.295$.

A.5.5 Inferring ρ Using Heuristic Optimization

Alternatively, we attempt to infer the direct association between "Attitudes to Crime" and "Tax Evasion" by searching for an assignment of values to the 22 covariates in Wahlund's path diagram using heuristic optimization (e.g. simulated annealing) such that a) the paths derived from our covariate assignment coincides with Wahlund's and b) 37% of our simulated population exhibits some tax evasion.

Since path coefficients are standardized linear regression coefficients, we need not be concerned with the actual values that the covariates take on, but rather the number of bins per covariate. Here, we take our best guess as to the number of bins Wahlund's covariates spanned, with the exception of those that are obviously indicator variables and "tax evasion", which Wahlund states takes on exactly three values: 0 = no evasion, 1 = evaded once, 2 = evaded more than once.

covariate	# of bins
age (a)	40
income (i)	10
working hours (wh)	50
student (s), retired (r), and self-employed (se)	2
opportunity (o)	3
attitudes to crime (atc)	3
tax evasion (te)	3
and all others	3

Hence, we assign each of the 430 simulated respondents with uniformly drawn responses taking on values defined by the bins above. The log-likelihood fit has several components. First, we fit our predicted path coefficients, β_{ij} to Wahlund's, $\beta_{ij}^{\text{Wahlund}}$, using the normal distribution:

$$\beta_{ij} \sim \mathrm{N}\big(\beta_{ij}^{\text{Wahlund}}, \sigma_{ij}^2\big)$$

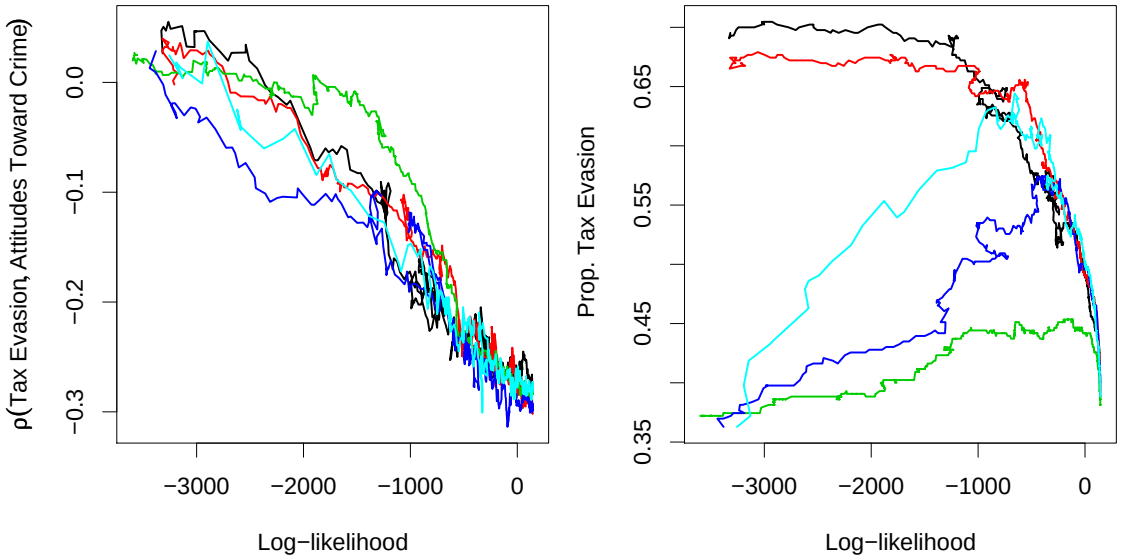
where i is the predicted covariate (recipient of a path) and j is the predicting covariate (source of a path), one of several in most cases, and arbitrary $\sigma_{ij} = 0.01$ for the main dependent variable and the directly predicting covariates, i.e. $i \in \{$tax evasion, opportunities, tax avoidance, attitudes to crime, perceived opportunities, perceived risk$\}$ and $\sigma_{ij} = 0.05$ for all other covariates.[29] These latter values are assigned arbitrarily giving more importance to "tax evasion" and those covariates that have direct paths to it. Furthermore, we need

[29]The rest of the path diagram may be found in Figure 10 of Appendix D.4.

to ensure that the rate of evasion hovers near 37%, employing the standard error around a proportion:

$$p(\text{“tax evasion”} > 0) \sim \mathrm{N}\big(p = 0.37, \sigma^2 = (1-p)p/(n = 430)\big)$$

The modal log-likelihood is $\hat{\mathcal{L}} = 140.9094 + 2.8411 = 143.7505$, each summand reflecting the above components respectively. We plot five annealing paths in Figure 4 and present the



(a) ρ v. Log-likelihood (b) Prop. v. Log-likelihood

Figure 4: Simulated Annealing Trajectories. *Three SA solution trajectories (every fifth data point) are shown. The left plot shows the triangulation of the Pearson correlation, ρ, between "tax evasion" and "attitudes to crime". The right plot shows the proportion of tax evaders, $p($"tax evasion"$ > 0)$; the green, blue, and cyan trajectories employ the "tax evasion" covariate initialized to the empirical proportion.*

final resting states and key measures, along with a weighted mean ρ, our new estimate, next to its weighted standard deviation:

$\mathcal{L}$	ρ	prop.	initialized prop.?
143.2233	−0.2762	0.3884	no
143.1073	−0.3017	0.3930	no
143.5790	−0.2807	0.3814	yes
142.8848	−0.2979	0.4000	yes
143.4131	−0.2815	0.3884	yes
weighted mean	−0.2861	(0.0095)	

Alternatively, we can use the annealing trajectory to predict the modal correlation, $\hat{\rho}$. In Figure 5, we show the trajectories for the higher log-likelihood ($\mathcal{L} > -400$). We employ a

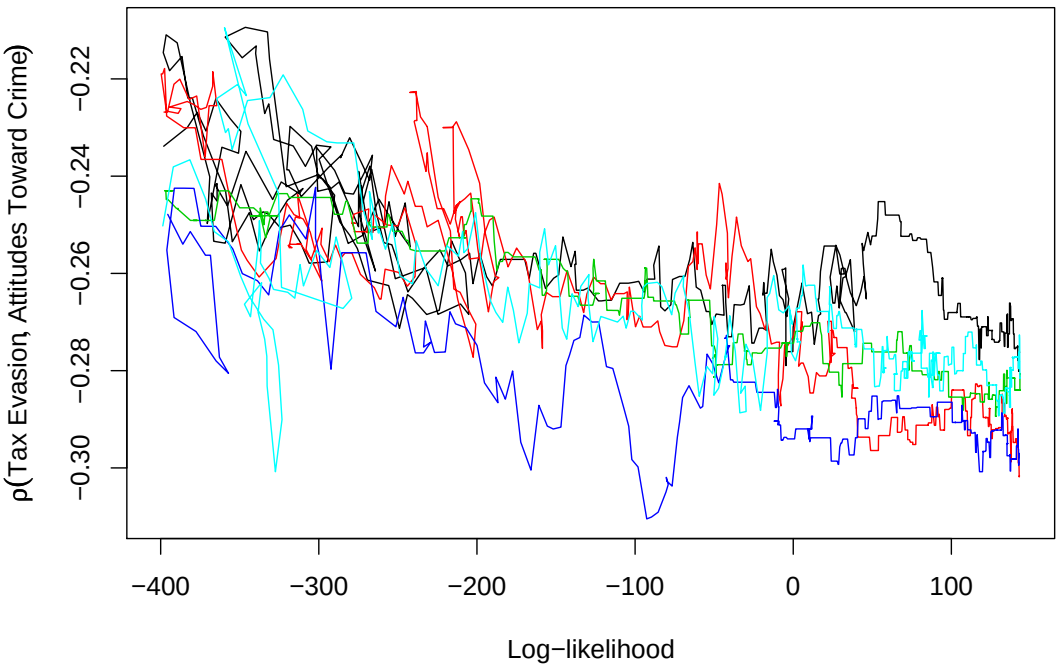


Figure 5: Simulated Annealing Trajectories. *We display the latter section of the annealing trajectories, i.e.* $\mathcal{L} > -400$*, for every fifth data point. The backtracking portions of the trajectories denote the annealer temporarily selecting less fit solutions.*

quadratic model (Adj-$R^2 = 0.6204$) to predict the modal $\hat{\rho}$:

$$\hat{\rho} = \widehat{(c,e)} = -0.2760 + \mathcal{L} \cdot -8.0281 \times 10^{-5} + \mathcal{L}^2 \cdot 4.9344 \times 10^{-8}$$

and, when we substitute $\mathcal{L} = \hat{\mathcal{L}} = 143.7505$, we obtain $\hat{\rho} = \widehat{(c,e)} = -0.2864$ (only four chains), which is very close to the weighted mean and reasonably close the quantity obtained earlier using Gaussian-based adjustments, i.e. the mean of draws from a set of normal distributions.

A.5.6 Enhanced Heuristic Inference

When we review the statistics for self-employment in the simulated data sets, we find some oddities:

$\mathcal{L}$	ρ_{atc}	ρ_{o}	ρ_{po}	ρ_{se}	$\mu_{\text{te}>0}$	μ_{se}	$\mu_{\text{te}>0}^{\text{se}=0}$	$\mu_{\text{te}>0}^{\text{se}=1}$
143.223	−0.276	0.325	0.274	−0.031	0.388	0.484	0.396	0.380
143.107	−0.302	0.337	0.296	0.109	0.393	0.479	0.344	0.447
143.579	−0.281	0.324	0.285	0.026	0.381	0.521	0.379	0.384
142.885	−0.298	0.342	0.283	−0.009	0.400	0.519	0.406	0.395
143.413	−0.281	0.325	0.286	0.039	0.388	0.544	0.367	0.406
weighted mean	−0.286	0.329	0.285	0.027	0.389	0.512	0.377	0.400

Specifically, we notice that the rate of self-employment (se) to be rather high, $\mu_{\text{se}} \approx 50\%$. Furthermore, the difference in tax evasion (te>0), between those who are self-employed and not, appears too low, $\sim 3\%$, leading us to consider including a fit towards self-employment rates in the likelihood. Fortunately, Vogel also surveyed a Swedish population, albeit in a different decade. We employ his marginal rates of both self-employment and tax evasion for each self-employment status, fitting to a Beta distribution, while imposing a scaling factor or multiplier, $\kappa_{\text{V}}^{\text{se}} = 0.5$, on Vogel's sample sizes, to acknowledge the uncertainty from the different decade as well as Vogel's unknown sampling procedures:[30]

$$\begin{aligned}
\mu_{\text{se}} &\sim \text{Beta}(\alpha_{\text{V}}^{\text{se}} + 1, \beta_{\text{V}}^{\text{se}} + 1) \\
\mu_{te>0}^{se=0} &\sim \text{Beta}(\alpha_{\text{te}>0}^{se=0} + 1, \beta_{te>0}^{se=0} + 1) \\
\mu_{te>0}^{se=1} &\sim \text{Beta}(\alpha_{\text{te}>0}^{se=1} + 1, \beta_{te>0}^{se=1} + 1)
\end{aligned}$$

where, occasionally omitting the (V)ogel subscript for readibility:

$$\begin{aligned}
\alpha_{\text{V}}^{\text{se}} &= (n_{\text{V}}^{se=1})(\kappa_{\text{V}}^{\text{se}}) = 106 \cdot 0.5 = 54 \\
\beta_{\text{V}}^{\text{se}} &= (n_{\text{V}}^{se=0})(\kappa_{\text{V}}^{\text{se}}) = 967 \cdot 0.5 = 483.5 \\
\alpha_{te>0}^{se=0} &= (n_{\text{V}}^{se=0})(\kappa_{\text{V}}^{\text{se}})(p_{te>0}^{se=0}) = 967 \cdot 0.5 \cdot 0.279 = 134.90 \\
\beta_{te>0}^{se=0} &= (n_{\text{V}}^{se=0})(\kappa_{\text{V}}^{\text{se}})(1 - p_{te>0}^{se=0}) = 967 \cdot 0.5 \cdot (1 - 0.279) = 348.60 \\
\alpha_{te>0}^{se=0} &= (n_{\text{V}}^{se=1})(\kappa_{\text{V}}^{\text{se}})(p_{te>0}^{se=0}) = 106 \cdot 0.5 \cdot 0.371 = 19.66 \\
\beta_{te>0}^{se=0} &= (n_{\text{V}}^{se=1})(\kappa_{\text{V}}^{\text{se}})(1 - p_{te>0}^{se=0}) = 106 \cdot 0.5 \cdot (1 - 0.371) = 33.34
\end{aligned}$$

Furthermore, we adjust some of the bin sizes:

covariate	# of bins	rationale
age (a)	72	reflects ages 18-89
opportunity (o)	8	aligns with our taxpayer categories
attitudes to crime (atc)	2	aligns with the GSS' 'obey law'

Furthermore, in the following sections, we incorporate additional fits to ensure reasonable distributions of age, retirement status, and working hours.

A.5.7 Fitting to PUMS Age

The distribution of ages ought to reflect some national sample; so, we employ the PUMS sample and account for the different nationalities by relaxing its restrictiveness (due to its sample size) and introducing a scaling factor/multiplier, $\kappa_{\text{W}}^{\alpha}$, which we apply to the α parameter of the multivariate Pólya (i.e. Dirichlet prior on a multinomial):

$$\boldsymbol{A} \sim \text{Pólya}(\alpha = \boldsymbol{\alpha}^{\text{PUMS}} + 1)$$

[30] Refer to Appendix A.3 for Vogel sample sizes.

where the ages are tabulated as follows are:

$$\begin{aligned}\boldsymbol{A} &= (n_{18}, \ldots, n_{89}) \\ \boldsymbol{A}^{\text{PUMS}} &= (n_{18}^{\text{PUMS}}, \ldots, n_{89}^{\text{PUMS}}) \\ \boldsymbol{\alpha}^{\text{PUMS}} &= \kappa_{\text{W}}^{\alpha} \cdot \boldsymbol{A}^{\text{PUMS}}\end{aligned}$$

and the count of ages in each age category, i, is:

$$n_i = \textstyle\sum_j \mathcal{I}(x_{\text{a}} = i) \text{ for } i \in \{18, \ldots, 89\} \text{ and } x_{\text{a}} \in \{x_{\text{age}}^{\text{W}}, x_{\text{age}}^{\text{PUMS}}\}$$

and where subscript 'a' denotes 'age', x^{W} is our simulated data (430×22), and x^{PUMS} is our 10K PUMS sub-sample. Furthermore, we will need to jointly fit the age distribution for both retirees and non-retirees as retirement is a function of age

A.5.8 Joint Age and Retirement

The joint age/retirement distribution will need to exhibit a mean retirement age of 62 and the 1st quartile should be near age 58 as reported by Gendell (1998); these statistics detail a normal distribution of mean of 62 and standard deviation of ∼5. As such, we ultimately seek a a function, $q_{\text{r}}(a)$ to predict the probability of retirement by a given age a, such that it reproduces the aforementioned statistics; 'r' denotes 'retired'. For now, we employ the U.S. age distribution to infer this function that tracks the cumulative probability of retirement for an individual at a given age, i.e. probability that retirement has occurred at this or a previous year. To start, we require a function that describes the probability of retirement at a particular age, $p_{\text{r}}(a)$, given that retirement has not yet already occured; this function needs to be monotonically increasing and perhaps increasing in slope as well given that the likelihood of retirement grows considerably with age. Hence, we employ a straighforward polynomial function, with some unknown exponent b:

$$\begin{aligned}p_{\text{r}}(a) &\propto a^b \\ &= a^b/89^b = (a - 89)^b\end{aligned}$$

We simply assume that the probability climbs to 1.0 at age 89.[31] So then, the probability of retirement occurring at or before age a would be:

$$\begin{aligned}F_p(a) &= 1 - \prod_{i=18}^{a} (1 - p_{\text{r}}(i)) \\ &= \Pr\{x_{\text{ra}} \leq a\}\end{aligned}$$

where x_{ra} is an unknown retirement age, past or future. Furthermore, this function is analogous to a cumulative distribution function (CDF), hence, we use that standard notation,

[31] We tested a model in which the probability climbs to some value < 1, employing a scaling factor; however the estimation gravitated towards the scaling factor equalling 1.

F. Since the probability of retirement function applies to only to the unretired portion of some sub-population, we compute the proportion of that sub-population, say a cohort, that exits/retires at each age up to and including the current age, using the following recurrence:

$$e_{\mathrm{r}}(a) = \begin{cases} p_{\mathrm{r}}(a) & \text{if } a = 18 \\ \left(1 - \sum_{i=18}^{a-1} e_i\right) \cdot p_{\mathrm{r}}(a) & \text{otherwise} \end{cases}$$

The logic here is that the proportion of new retirements at a particular age, a, applies only to the unretired proportion of the cohort. Hence, $e_{\mathrm{r}}(a)$ accumulates the proportion of a cohort, or age group, which has retired by that age, a. Furthermore, $e_{\mathrm{r}}(a)$ is now our actual probability distribution function, and its CDF (i.e. $F_e(a)$) equals $F_p(a)$.

Furthermore, we believe that $e_{\mathrm{r}}(a)$ is similiar to the probability distribution function (PDF) for the Gaussian/normal distribution so we introduce an alternative probability distribution function (and also its accompanying CDF) based simply on the normal distribution with unknown parameters μ and σ, subsequently normalized to deal with the truncated age range, 18–89:[32]

$$e_{\mathrm{r}}(a) = \frac{\mathrm{N}(a|\mu, \sigma^2)}{\sum_{i=18}^{89} \mathrm{N}(i|\mu, \sigma^2)}$$

$$\begin{aligned} F_e(a) &= \textstyle\sum_{i=18}^{a} e_{\mathrm{r}}(a) \\ &\approx \Phi(a|\mu, \sigma^2) \end{aligned}$$

We will test both the polynomial and normal approaches to $e_{\mathrm{r}}(a)$.

Now, we can calculate the approximate number of retirements at each age, $q_{\mathrm{r}}(a)$, by applying some population structure, in this case the PUMS distribution, A^{PUMS}, and we also calculate the corresponding normalized proportion, $\hat{q}_{\mathrm{r}}(a)$:

$$q_{\mathrm{r}}(a) = e_{\mathrm{r}}(a) \cdot \sum_{i=a}^{89} n_i^{\mathrm{PUMS}}$$

$$\hat{q}_{\mathrm{r}}(a) = \frac{q_{\mathrm{r}}(a)}{\sum_{i=18}^{89} q_{\mathrm{r}}(i)}$$

Essentially, the total number of retirements that occurred at a certain age, a, is some fraction of the sum of the all individuals who passed through that age (i.e. $\geq a$).

[32]The cumulative distribution function (CDF) for the Gaussian/normal distribution is:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-t^2/2}\, dt$$

Now, we find the mean retirement age, $\overline{q}_{\mathrm{r}}$, and the cumulative probability for $a = 58$ (i.e. the proportion of individuals who have retired by age 58), putatively the 1st quartile:

$$\overline{q}_{\mathrm{r}} = \sum_{i=18}^{89} i \cdot \hat{q}_{\mathrm{r}}(i)$$
$$F_q(a \leq 58) = \sum_{i=18}^{58} \hat{q}_{\mathrm{r}}(i)$$

And also, we compute the % of exits for which the 50–54 and 70+ age brackets are responsible:

$$\hat{q}_{\mathrm{r}}^{\text{50-54}} = \sum_{i=50}^{54} \hat{q}_{\mathrm{r}}^{A}(i)$$
$$\hat{q}_{\mathrm{r}}^{\text{70+}} = \sum_{i=70}^{89} \hat{q}_{\mathrm{r}}^{A}(i)$$

To find the exponent b (for the polynomial approach) as well as μ and σ (for the normal approach), we fit the following likelihoods using the Newton-Raphson algorithm:

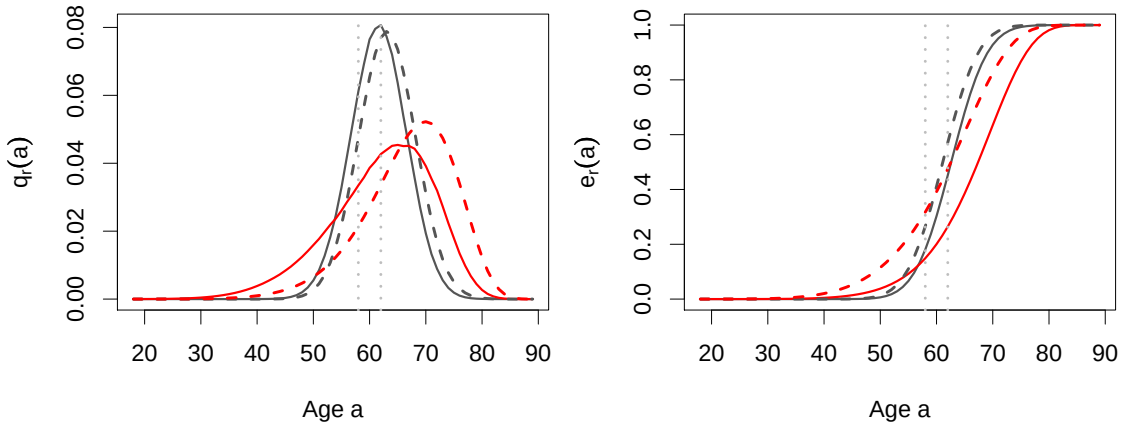
$$\overline{q}_{\mathrm{r}} \sim \mathrm{N}\left(\mu = 62, \sigma^2 = \frac{5^2}{n = 100{,}000}\right)$$
$$F_q(a \leq 58) \sim \mathrm{N}(\mu = \mathrm{logit}[0.25], \sigma^2 = [\mu(1-\mu)n]^{-1})$$
$$\hat{q}_{\mathrm{r}}^{\text{50-54}} \sim \mathrm{N}(\mu = \mathrm{logit}[0.099], \sigma^2 = [\mu(1-\mu)n]^{-1})$$
$$\hat{q}_{\mathrm{r}}^{\text{70+}} \sim \mathrm{N}(\mu = \mathrm{logit}[0.026], \sigma^2 = [\mu(1-\mu)n]^{-1})$$

where, here, we use a reduced $n = 100{,}000$ rather than 8.7 million, the population of Sweden in 1992, partly to reflect the uncertainty inherent in applying a U.S. population sample. Also, for the variance of the logit of the cumulative probability, we employ the approximation of the variance of the logit which was found, through simulation, to be an adequate.

We find the following modal solutions and overlay the accompanying distributions in Figure 6:

prob. function	parameters	$\overline{q}_{\mathrm{r}}$	$F_q(a \leq 58)$	% of Exits 50-54	70+	$\mathcal{L}$
$p_{\mathrm{r}}(a\|b)$	$b = 8.656$	61.99	0.317	9.9%	21.4%	−7857
$\mathrm{N}(a\|\mu, \sigma^2)$	$\mu = 63.11, \sigma = 5.06$	61.58	0.265	7.0%	5.4%	−1786
Gendell (1998)		62.00	0.250	9.9%	2.6%	

While both methods fit the mean age quite well, both also exhibit some noticeable shortcomings. Specifically, the polynomial approach fits just the % of Exits from 50–54 modestly well,



(a) e_{r} and q_{r} (PDFs) (b) CDFs of e_{r} and q_{r}

Figure 6: Probability of Retirement. *We display the PDFs (left graph) and CDFs (right graph) that describe the distribution functions for the polynomial approach (black, $p_r(a|b)$) and the normal distribution (red, $N(a|\mu,\sigma)$). In each graph, we depict the raw probability functions (solid lines, e_r) as well as the data-weighted probability functions (dashed lines, q_r). The dotted, vertical lines mark the the mean age of retirement 62 and the 1st quartile age 58.*

while grossly mismatching the other statistics. The normal approach however fits all three remaining statistics modestly well. These differences in approaches become apparent when we examine their respective distribution functions in Figure 6. Not surprisingly, the latter approach corresponds to a far superior likelihood ($\mathcal{L}$) and is our model of choice. We surmise that some of the lack of fit is due to our employing a sample of the 2000 U.S. population rather than the Swedish population age distribution in 1992.[33]

Finally, we need to extend our fit of simulated data (x^{W}) to the base age distribution, detailed in Section A.5.7, to reflect the distinct age distributions for retirees and non-retirees:

$$\begin{aligned}\boldsymbol{A}|x_{\mathrm{r}}=0 &\sim \text{Pólya}(\boldsymbol{\alpha}=\boldsymbol{A}^{\text{PUMS}}_{\text{r=0}}+1)\\ \boldsymbol{A}|x_{\mathrm{r}}=1 &\sim \text{Pólya}(\boldsymbol{\alpha}=\boldsymbol{A}^{\text{PUMS}}_{\text{r=1}}+1)\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{A}^{\text{PUMS}}_{\text{r=0}} &= \boldsymbol{A}^{\text{PUMS}}\cdot(1-\boldsymbol{F}_e)\\ \boldsymbol{A}^{\text{PUMS}}_{\text{r=1}} &= \boldsymbol{A}^{\text{PUMS}}\cdot\boldsymbol{F}_e\end{aligned}$$

and $\boldsymbol{F}_e = (F_e(18),\ldots,F_e(89))$. Basically, we partition the count of agents in each age category a (i.e. n_a^{PUMS}) to those who have not retired by that age (i.e. $1-F_e(a)$) and those who have (i.e. $F_e(a)$)

[33] We will employ the proper age distribution (Swedish poplulation) in later writings, time permitting.

A.5.9 Working Hours and Retirement

We also constrain the respondent working hours (wh) according to the retirement status (r), using normative assumptions that non-retirees work close to 40 hours per week, while retirees work almost zero hours:

$$x_{\text{wh}}|x_{\text{r}} = 0 \sim \text{N}(\mu = 40, \sigma^2 = 5^2)$$
$$x_{\text{wh}}|x_{\text{r}} = 1 \sim \text{Exp}(\lambda = 0.2)$$

where 'wh' denotes 'working hours' and 'Exp' denotes the exponentional distribution, considered the appropriate fit for the zero (or few) hours of work engaged by retirees. In Figure 7, we display both distributions.

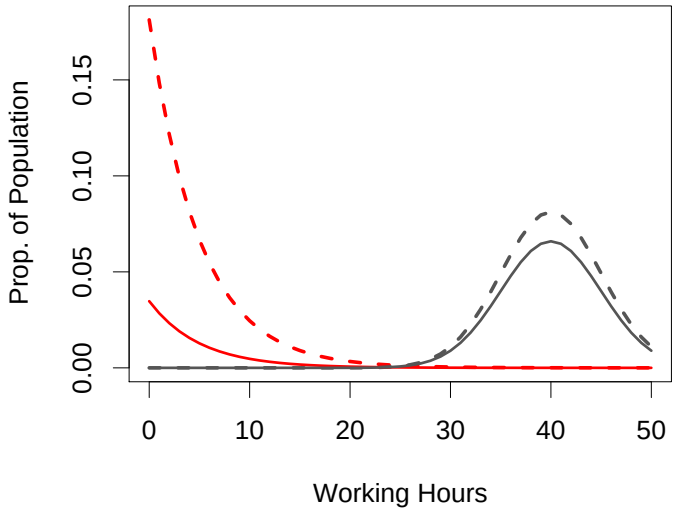


Figure 7: Working Hours. *Both the non-retiree (black, normal) and retiree (red, exponential) distributions are graphed. With each type, we also distinguish between the raw distribution (dashed line) and data-weighted distribution (solid line), in which retirees constitute only 17.4% of the population. Hence, the area under both solid line distributions should sum to unity, while the area under each dashed line sums to unity.*

A.5.10 Fitting Wahlund and Results

We now resort to a more straightforward optimization heuristic, expectation maximization (EM), due to the clumsiness of simulated annealing. The algorithm is similar to the stochastic maximation algorithm introduced in Section 6.3:

1. For each covariate across all respondents (i.e. $(n = 430) \times (m = 22) = 9460$ responses):

    (a) Calculate the fit (i.e. $\mathcal{L}$) for every possible new value; e.g. with binary outcomes, this comprises only one other $\mathcal{L}$ while for age, there would be 70 other bin values, for which we compute each $\mathcal{L}$.

(b) Choose the value which corresponds to the highest $\mathcal{L}$.

2. Repeat Step 1 until a pass over all responses is made without any quantities altered.

		Correlation Between Tax Evasion and ... Attitudes To Crime			Opportunity			
alg	fit	ρ_{atc}	σ_{atc}	Adj-R^2	ρ_{o}	σ_{o}	Adj-R^2	n
SA	lm	-0.2849	9.7×10^{-5}	0.60	–	–	–	22356
SA	lm	-0.2851	3.1×10^{-4}	0.57	–	–	–	2500
SA	μ	-0.2859	0.0110	–	0.3291	0.0076	–	5
SA	μ^{te2}	-0.2527	0.0150	–	0.3255	0.0290	–	5
EM	lm	-0.2901	6.8×10^{-4}	0.67	0.3448	4.0×10^{-4}	0.96	1000
EM	μ	-0.2918	0.0140	–	0.3464	0.0058	–	10
EM	lm^{te2}	-0.2680	0.0011	0.48	0.3308	6.3×10^{-4}	0.90	1000
EM	μ^{te2}	-0.2701	0.0180	–	0.3358	0.0140	–	10
EM	$\text{lm}^{\text{te2}}_{\text{o8}}$	-0.2693	0.0010	0.50	0.3317	5.3×10^{-4}	0.92	1000
EM	$\mu^{\text{te2}}_{\text{o8}}$	-0.2733	0.0170	–	0.3351	0.0110	–	10

Table 7: Predicted Correlations ρ_{atc} and ρ_{o}. *The optimization method is noted in the algorithm* (alg.) *column.*

In Table 7, we offer predicted correlations between 'tax evasion' and both 'attitudes to crime' (atc) and 'opportunity' (o) from both simulated annealing (SA) and stochastic expectation maximization (EM); the EM models include the latest fits (Vogel's self-employment, age, retirement, and working hours), while the SA models do not. We make predictions with either a quadratic linear model (lm) or weighted mean (μ) over the values corresponding to the the maximal likelihoods for each chain; the linear model uses a 100 point sub-sample of each chain with the condition of higher-likelihood: $\mathcal{L} > -400$. Sample sizes for each prediction is denoted by either the combined length of the ten chains or the number of chains, for weighted means. The standard deviations σ indicate the standard error around the linear fit, for 'lm', or the weighted standard deviation, for 'μ'. Furthermore, we tested the correlations with two sets of scales for each of the covariates: 'tax evasion', 'attitudes to crime', and 'opportunity'. Since the Wahlund's 'tax evasion' is on a 3-point scale, while our model seeks prediction for any intentional error, we assessed correlations to a similar, binary dependent variable (te2). The (SA) models employed a 3-point scale for 'attitudes to crime', while the (EM) models employ a 2-point/binary scale, to maintain similarity with the GSS' 'obey law' item. For 'opportunity' we test an 'opportunity' scale similar to our taxpayer categories (i.e. 0–7 taxpayer categories = 8 point scale); these predictions are denoted by (o8).

While there is a fair amount of concordance in the correlations between the various experimental and measurement conditions, we observe that correlations alter by a few points

depending on the categories/bins employed. Still, the slight variance reassures us that the precision in our error models will not be greatly affected by our choosing the wrong correlation.

A.6 The General Social Survey's 'Obey Law' Covariate

In this appendix, we examine the 'Obey Law' correlate, found in several survey years of the General Social Survey (GSS), along with several other pertinent variables to explore their applicability in predicting intentional error.

A.6.1 Obey Law in the GSS

The General Social Survey (GSS) asked several items pertaining to tax evasion as well as general adherence to the lawful behavior: Obey Law, Tax Cheat, Pay Taxes, and Obey Laws. We examine these for relevance.[34]

> 1158. **Obey Law**: In general, would you say that people should obey the law without exception, or are there exceptional occasions on which people should follow their consciences even if it means breaking the law? (CIRCLE ONE ANSWER)

	Obey Law	Follow Consciences	Can't Choose
count	2414	2079	190
prop.	0.515	0.444	0.041

> 1377. **Tax Cheat**: Do you feel it is wrong or not wrong if a taxpayer does not report all of his income in order to pay less income taxes.

	Not Wrong	A Bit Wrong	Wrong	Seriously Wrong
count	103	282	1333	769
prop.	0.041	0.113	0.536	0.309

> 1464. There are different opinions as to what it takes to be a good citizen. As far as you are concerned personally on a scale of 1 to 7, where 1 is not at all important and 7 is very important, how important is it to:
>
> B. **Pay Taxes**: Never to try to evade taxes.

	Not at All Important						Very Important
	1	2	3	4	5	6	7
count	28	9	22	64	107	165	1066
prop.	0.019	0.006	0.015	0.044	0.073	0.113	0.730

[34]For all these analyses, we employ the GSS' WTSSALL sample weighting variable.

C. **Obey Laws**: Always to obey laws and regulations.

	Not at All Important						Very Important
	1	2	3	4	5	6	7
count	5	2	18	52	141	263	986
prop.	0.003	0.001	0.012	0.035	0.096	0.179	0.672

Unfortunately, joint analysis is restricted since only two of these items were asked in the same survey year:

Year	Tax Cheat	Pay Taxes	Obey Law	Obey Laws
1985			✓	
1990			✓	
1991	✓			
1996			✓	
1998	✓			
2004		✓		✓
2005			✓	

As a comparison exercise, we fit a logit model to different partition thresholds employing the same covariates we do for the error model.[35] The dependent variable achieved through the partitioning predicts towards attitudes that would be consistent in individuals who commit intentional error, with increasingly precise anti-tax attitudes with the decreasing threshold. So first, we examine the first tax evasion item (Tax Cheat) by dichotomizing at different thresholds and employing a logistic regression:

	$y =$ Tax Cheat, $n = 2476$		
Variable	$y < 4$	$y < 3$	$y < 2$
Intercept	0.948^{***}	-1.405^{***}	-2.380^{***}
	(0.114)	(0.141)	(0.251)
Sex	$0.159^{\hat{}}$	0.319^{**}	0.455^{*}
	(0.092)	(0.117)	(0.211)
Age	-0.084	-0.425^{***}	-0.323^{*}
	(0.068)	(0.089)	(0.153)
Education	-0.195^{**}	$-0.154^{\hat{}}$	-0.716^{***}
	(0.061)	(0.083)	(0.169)
Income	0.053	0.065	0.002
	(0.032)	(0.042)	(0.079)
AIC	3503	2460	1005
μ	0.691	0.155	0.042

[35] That is, we re-categorize the GSS covariates and align them with the PUMS'.

We find that the direction of the significant predictors is consistent with those reported in most of the empirical studies. Specifically, sex (or being male) increases the likelihood of evasion while increasing age and education has a reducing effect. A cursory comparison between these three models and our inferred error models, earlier reported in Table 1, reveals that the 'tax cheat" cut-off best coinciding with actual intentional error resides somewhere between $y < 2$ and $y < 3$, which maps to "A Bit Wrong" and "Wrong". One facile implication is that a majority of evaders consider their actions to be at least a bit wrong. However, if we look at the proportion of the respondents in each parition, μ, we find that the region of interest lies between $y < 4$ and $y < 3$. That is, our presumptive rate of evasion of $\sim$25% lies in between the μ's corresponding to the two first two models. Next, we examine a similar covariate, "Pay Taxes":

	$y =$ Pay Taxes, $n = 1457$					
Variable	$y < 7$	$y < 6$	$y < 5$	$y < 4$	$y < 3$	$y < 2$
Intercept	-0.728^{***}	-1.304^{***}	-2.224^{***}	-2.868^{***}	-3.439^{***}	-3.357^{***}
	(0.153)	(0.182)	(0.244)	(0.355)	(0.456)	(0.498)
Sex	0.403^{***}	0.434^{**}	0.954^{***}	1.322^{***}	1.287^{***}	1.283^{**}
	(0.122)	(0.147)	(0.198)	(0.298)	(0.383)	(0.425)
Age	-0.367^{***}	-0.361^{**}	-0.321^{*}	-0.328	-0.289	-0.300
	(0.097)	(0.115)	(0.146)	(0.202)	(0.257)	(0.280)
Education	-0.077	-0.191^{*}	-0.166	-0.672^{***}	-0.638^{**}	-0.918^{**}
	(0.079)	(0.097)	(0.126)	(0.193)	(0.246)	(0.289)
Income	-0.013	0.015	-0.084	0.027	0.009	-0.014
	(0.038)	(0.046)	(0.061)	(0.084)	(0.108)	(0.125)
AIC	1510	1176	763	469	308	271
μ	0.269	0.163	0.087	0.044	0.026	0.021

Again, with a different GSS tax-related covariate, "Pay Taxes", we find the model predictions, valence-wise, coincide with the "Tax Cheat" as well as our own error models. Here, the models coincident to our error models are the first three $y < 7$, $y < 6$, and $y < 5$, which refer to a belief that it is at least moderately important for a good citizen to never evade taxes, while the first two models' μ's bound our predicted intentional error rate. We now look into some similar GSS covariates that deal with obedience to the law: "Obey Law" and

"Obey Laws".

	Obey Law, $n = 4659$	$y =$ Obey Laws, $n = 1462$				
Variable	$y > 1$	$y < 7$	$y < 6$	$y < 5$	$y < 4$	$y < 3$
Intercept	-0.267^{***}	-1.111^{***}	-2.112^{***}	-3.012^{***}	-4.215^{***}	-2.648^{***}
	(0.079)	(0.149)	(0.199)	(0.311)	(0.561)	(0.765)
Sex	0.256^{***}	0.586^{***}	0.473^{**}	0.769^{**}	$0.886^{\hat{}}$	-0.755
	(0.063)	(0.117)	(0.157)	(0.249)	(0.453)	(1.000)
Age	-0.233^{***}	-0.334^{***}	-0.284^{*}	-0.299	-0.271	-0.935
	(0.048)	(0.094)	(0.126)	(0.188)	(0.334)	(0.653)
Education	0.489^{***}	0.275^{***}	$0.184^{\hat{}}$	0.084	-0.095	-1.599^{*}
	(0.044)	(0.074)	(0.098)	(0.154)	(0.282)	(0.780)
Income	0.032	0.017	0.033	-0.097	-0.038	-0.719
	(0.021)	(0.036)	(0.047)	(0.076)	(0.135)	(0.586)
AIC	6632	1622	1025	523	204	64
μ	0.553	0.319	0.138	0.050	0.015	0.004

While sex and age predict disobedience to the law in similar directions as the tax items, education's prediction is opposite. This, however, is not entirely surprising since an expression of disobedience does not necessarily imply malfeasance or a lack of ethics but often civic conscientiousness, which some would argue is a product of higher education. One item of concern is that, at best, the partition of the newer "Obey Laws" covariate yields only a 30-70 split while the original "Obey Law" item maintains an almost 50-50 split. Since "Obey Laws" both coincides with Wahlund's "Attitude to Crime" and tax evasion, we examine the relationship (i.e. Pearson correlation):

Statistic	PayTaxes	ObeyLaws	PayTaxes < 7	ObeyLaws < 7
μ	6.39	6.47	0.269	0.319
σ_μ	1.26	0.94	0.012	0.012
ρ (σ_ρ)	0.373^{***}	(0.024)	0.393^{***}	(0.024)

The scaled (left) and dichotomized (right) correlations are similar. The dichotomizations occur where they best correspond to empirical rates of evasion and the alternate "Obey Law" proportions (i.e. $y < 7$ for both). We will use these correlations along with the estimated Wahlund correlation (next section) to infer a better correlation between obedience to the law and tax evasion, or intentional error.

A.6.2 Combining Obey Law Correlates from GSS and Wahlund

We can now combine both the GSS Obey Law findings with our inferred Wahlund correlation for "attitudes to crime" covariate. We employ $\rho_{\text{Wahlund}} = -0.2864$ and $\rho_{\text{GSS}} = -0.393$. First, we impose a constraint of Obey Law distribution using the GSS Obey Law on the Wahlund crime covariate:

GSS Obey Law	Follow Conscience	Obey Law
	0.5373	0.4627

We do this in order to obtain an exact solution for the 2×2 contingency table which yields $\rho_{\text{Wahlund}} = -0.2864$:

		Obey Law: Follow Conscience	Obey Law: Obey Law	
Evade Taxes	No	0.2439	0.3861	
	Yes	0.2934	0.0766	0.37
			0.4627	

We add the Wahlund and GSS contingency tables to reach a joined correlation, using two sample sizes for Wahlund, for one survey year ($n = 430$, i.e. phone interviews) and all four survey years ($n = 1427$, i.e. total follow-up surveys) and the one sample of the GSS ($n = 1461$):

	$n = 430$		$n = 1427$	
	Additive	MLE	Additive	MLE
ρ_{Both}	-0.3753	-0.3703	-0.3498	-0.3425
σ_{Both}	(0.0213)	(0.0214)	(0.0174)	(0.0175)

There remains some imprecision in the use of the GSS contingency table esp. that the GSS' 2004 "Obey Laws" dichotomy does not match the multi-year "Obey Law" outcomes. Furthermore, we assume there is perfect alignment between Wahlund's "Attitudes to Crime" and the GSS' "Obey Law" and have not taken into account any mismatch. One way to address this is to simply expand the standard error surrounding of correlation.

A.6.3 Obey Law Imputation

We can now estimate the impact of Obey Law on intentional error by several methods of imputation, primarily by estimating the Obey Law response of the PUMS agents using the relationship between the GSS' Obey Law and socio-demographic covariates. We first analyze a simple case of predicting intentional error with just one socio-demographic covariate, sex, along with Obey Law, in order to uncover any differences between numerical estimation of the model coefficients and imputation.

$$\mu(\text{logit}^{-1}[\boldsymbol{y}|x_k = \ell]) \sim \text{Beta}(\alpha = p_\ell^k n_\ell^k + 1, \beta = (1 - p_\ell^k) n_\ell^k + 1) \tag{9}$$

where $k \in \{\text{Sex}, \text{Obey}\}$; and $l \in \{0, 1\}$; and

$$\boldsymbol{y} = \beta_0 + \beta_{\text{Sex}} \cdot \boldsymbol{x}_{\text{Sex}} + \beta_{\text{Obey}} \cdot \boldsymbol{x}_{\text{Obey}}$$

and p^{Sex} and n^{Sex} is taken directly from Vogel's data:

	x_{Sex}		
stat	$0 = \text{Female}$	$1 = \text{Male}$	Description
p^{Sex}	0.217	0.323	probability of evasion
n^{Sex}	506	709	sample count from Vogel

and p^{Obey} and n^{Obey} is inferred from Wahlund's path coefficients and GSS' Obey Law, described in Appendix A.6.2, where $\rho_{\text{Both}}(\text{'Obey Law'},\text{'Tax Evasion'}) = 0.352$ yields:

	x_{Obey}		
stat	$0 = \text{Not Always}$	$1 = \text{Always}$	Description
p^{Obey}	0.5282	0.1874	probability of evasion
n^{Obey}	230.38	199.62	sample count for obey law category

Since we are primarily interested in intentionl error rates within each gender sub-group, we can arbitrarily assign a population size n which applies to each female and male sub-group. So now, we obtain the weighted mean probability of intentional error for each of the gender sub-groups (i.e female and male) and each of the obey law sub-groups (i.e. not always and always obey), contingent on some pre-defined partitioning of disobedience within each gender group, indicated by $\boldsymbol{m}$:

$$\hat{\mu}(\text{logit}^{-1}[y]|x_k = \ell, \boldsymbol{m}, \boldsymbol{\beta}) = \left(\frac{u}{n}\right) \text{logit}^{-1}[y|\boldsymbol{x} = (a, b)] + \left(\frac{v}{n}\right) \text{logit}^{-1}[y|\boldsymbol{x} = (c, d)] \tag{10}$$

where our unknown parameters are $\boldsymbol{\beta} = (\beta_0, \beta_{\text{Sex}}, \beta_{\text{Obey}})$. Our known parameters include $\boldsymbol{x} = (x_{\text{Sex}}, x_{\text{Obey}})$; $\ell \in \{0, 1\}$. The marginal y prediction is contingent on the covariate k as well as the choice of sub-group combination:

$$(u, v, a, b, c, d) = \begin{cases} (\quad m_\ell \quad , \quad n - m_\ell \quad , \ell, 0, \ell, 1) & \text{if } k = \text{Sex} \\ (m_0 + n\ell - 2m_0\ell, m_1 + n\ell - 2m_1\ell, 0, \ell, 1, \ell) & \text{if } k = \text{Obey} \end{cases}$$

and $\boldsymbol{m} = (m_0, m_1) \leq n$ is the count of "disobeyers" for each gender sub-group, each of size n; that is,

$$m_\ell = \sum_{i=1}^{n} \mathcal{I}(x_{i,\text{Obey}} = 0 | x_{i,\text{Sex}} = \ell)$$

Hence, $(n - \boldsymbol{m})$ gives us the counts of 'obeyers'. We define $L_k(\ell, \boldsymbol{m})$ as the likelihood from (9) substituting the $\mu(\text{logit}^{-1}[\boldsymbol{y}|x_k = l])$ with a $\hat{\mu}$ from (10). Now, we obtain the likelihood for a given partition of disobedience, $\boldsymbol{m}$, within each gender group:

$$L_\mu(\boldsymbol{m}) = \prod_{\substack{k \in \{\text{Sex}, \\ \text{Obey}\}}} \prod_{\ell \in \{0,1\}} L_k(\ell, \boldsymbol{m})$$

We parameterize the likelihood of $\boldsymbol{m}$ with the cross-tabulation of the GSS Obey Law and Sex covariates, in the matrix, $\boldsymbol{z}$:

$$\boldsymbol{z} = \begin{array}{cc|cc} & & \multicolumn{2}{c}{\text{Obey Law}} \\ & & 0 & 1 \\ \hline \text{Sex} & 0 & 1166 & 1199 \\ & 1 & 1248 & 880 \end{array}$$

which we index as $z_{s,o}$, e.g. $z_{0,0} = z_{\{\text{Sex}=0,\text{Obey}=0\}} = 1166$. The likelihood, for each gender group ℓ is then:[36]

$$\begin{aligned} m_\ell &\sim \int_{\theta=0}^{1} \text{Beta}(\theta|\alpha = z_{\ell,0} + 1, \beta = z_{\ell,1} + 1) \cdot \text{Binomial}(m_\ell|n, \theta)\, d\theta \\ &\sim \text{BetaBinomial}(n, \alpha = z_{\ell,0} + 1, \beta = z_{\ell,1} + 1) \end{aligned}$$

or, more succintly, for both $\boldsymbol{m}$:

$$L_m(\boldsymbol{m}) = \prod_{\ell \in \{0,1\}} \text{BetaBinomial}(m_\ell|n, \alpha = z_{\ell,0} + 1, \beta = z_{\ell,1} + 1)$$

We can now express the entire likelihood, summing across all combinations of sub-partitions:[37]

$$\mathcal{L} = \log \left[\sum_{i=0}^{n} \sum_{j=0}^{n} L_m(\boldsymbol{m} = (i,j)) \cdot L_\mu(\boldsymbol{m} = (i,j)) \right]$$

In short, we are ascertaining a weighted likelihood across all the possible ways two groups (i.e. female/male), each having n individuals, can be partitioned into two sub-groups (i.e. disobey and obey law). So now, for a pre-defined n, we can seek our unknown $\boldsymbol{\beta}$ parameters using Newton-Raphson. The alternative approach is to employ imputation/Monte Carlo simulation and randomly draw each m_ℓ (from the beta-binomial); we employed 100 sets of draws for $\boldsymbol{m}$ under each n condition. We present results from both approaches, while curtailing the n for the numerical/MLE analysis due to the increasing number of calculations

[36] The beta-binomial (i.e. binomial with a beta prior) is the natural distribution to employ here, given that both our known and unknown parameters are discrete quantities; the former defines a distribution of latent probabilities θ, while the latter specifies a discrete outcomes from all the θ.

[37] The normalization of L_m is unnecessary as it produces identical results.

required:[38]

	Numerical			Imputation		
n	β_0	β_{Sex}	β_{ObeyLaw}	β_0	β_{Sex}	β_{ObeyLaw}
10	-0.2021	0.3689	-1.5809	-0.3678	0.3260	-1.6011
25	-0.2878	0.4087	-1.6071	-0.4845	0.4255	-1.5359
50	-0.3565	0.4072	-1.5911	-0.4688	0.3963	-1.5317
100	-0.4130	0.4048	-1.5625	-0.4997	0.4172	-1.5148
500	-0.4717	0.4028	-1.5241	-0.4949	0.4047	-1.5087
1000	-0.4797	0.4026	-1.5184	-0.4920	0.4054	-1.5118
5000	$\mathbf{-0.4862}$	**0.4024**	$\mathbf{-1.5138}$	$\mathbf{-0.4913}$	**0.4030**	$\mathbf{-1.5107}$

Under both the numerical and imputation approaches, there is adequate convergence and similarity. Not surprisingly, the imputation approach requires far less computation time. We now inspect the inference of the standard errors around the estimated $\boldsymbol{\beta}$:[39]

	Numerical			Imputation		
n	σ_0	σ_{Sex}	σ_{ObeyLaw}	σ_0	σ_{Sex}	σ_{ObeyLaw}
10	0.2239	0.3337	0.2302	0.4057	0.5301	0.4097
25	0.1901	0.2862	0.2293	0.2026	0.2739	0.2625
50	0.1705	0.2272	0.2299	0.1675	0.2224	0.2406
100	0.1548	0.1900	0.2288	0.1493	0.1896	0.2313
500	0.1376	0.1577	0.2249	0.1345	0.1551	0.2245
1000	0.1352	0.1537	0.2241	0.1337	0.1526	0.2239
5000	**0.1332**	**0.1504**	**0.2235**	**0.1321**	**0.1485**	**0.2232**

The standard errors converge and coincide in the same manner as the estimates. In conclusion, there appears to be adequate parity between the imputation and numerical approaches, which permits us to employ either method. However, due to the computational expense of

[38] E.g. when $n = 100$, there $101 \times 101 = 10201$ possible partitions, each of which requires, from the Newton-Raphson differencing approach, $2m + 2m + 4m$ (where $m = \#$ of parameters $= 3$) $= 24$ likelihood calculations per step resulting in 244,824 total calculations.

[39] The total variance, T, of an estimate, say $\beta = \beta_{\text{Sex}}$, combines the between-sample and within-sample variances:

$$T = W + \frac{K+1}{K} B$$

where K is number of simulations (here, 100) and

$$B = \frac{1}{K-1} \sum_{k=1}^{K} \left(\beta_k - \overline{\beta}\right)$$

and

$$W = \frac{1}{K} \sum_{k=1}^{K} \sigma_k^2$$

the numerical approach, we will employ the imputation method of estimating 'obey law' for the PUMS agents.

In Figure 8, we display the logit of the mean of the empirical Obey Law response for each *uid*, alongside the linear model prediction (which employs all four socio-demographic covariates).[40] These point estimates are bounded by the appropriate standard deviation, which in the case of the empirical logit indicates the sub-sample size of each *uid*. This figure tells us that the linear fit, while modestly accurate in the mean rates of obey law, does not capture the uncertainty surrounding particular *uid*'s. So, for imputing the Obey Law response for our $n = 10{,}000$ PUMS agents, we can either 1) employ the empirical mean or 2) draw each agent's response from the beta-binomial distribution instead of numerically exploring the space of all possible *uid* partitions, of which there are 4.52×10^{222} combinations! Each Obey Law response draw depends on the agent's *uid*, and corresponding certainty afforded by the GSS:

$$x_{uid}^{\text{Obey}} \sim \text{BetaBinomial}(n = 1, \alpha = n_{uid}^{\text{Obey=1}} + 1, \beta = n_{uid}^{\text{Obey=0}} + 1)$$

where

$$n_{uid}^{\text{Obey=o}} = \sum_{i=1}^{n_{\text{GSS}}} \mathcal{I}(x_{i,\text{Obey}} = o, f_{\text{uid}}(\boldsymbol{x}_i) = uid)$$

Recall, we define the function $f_{\text{uid}} : \boldsymbol{x} \rightarrow uid$ in (7). For those *uid*s lacking corresponding types in the GSS (20 of them), we employ the linear prediction with uncertainty defined by one of a) an α and β pair that matches the fitted value's standard deviation, which assumes a *uid* sample size similar to the data, and b) a value of 1, to indicate the source is not directly empirical and exhibits high uncertainty. Yet, another alternative would be to simply use the GSS with covariates aligned to PUMS categories. This will be explored in further writings.

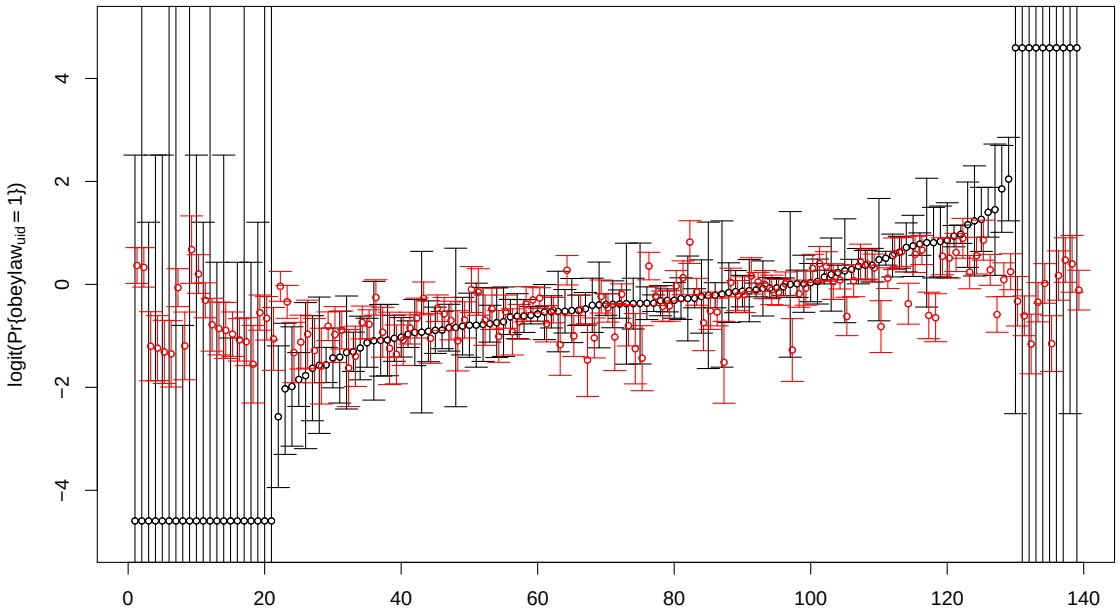
B Accounting for Social Influence Effects

Since the predictive models for line-item intentional error are designed to be used in a multi-agent information diffusion model, we attempt to partial out the peer influence effects from our models. The literature offers the following findings:

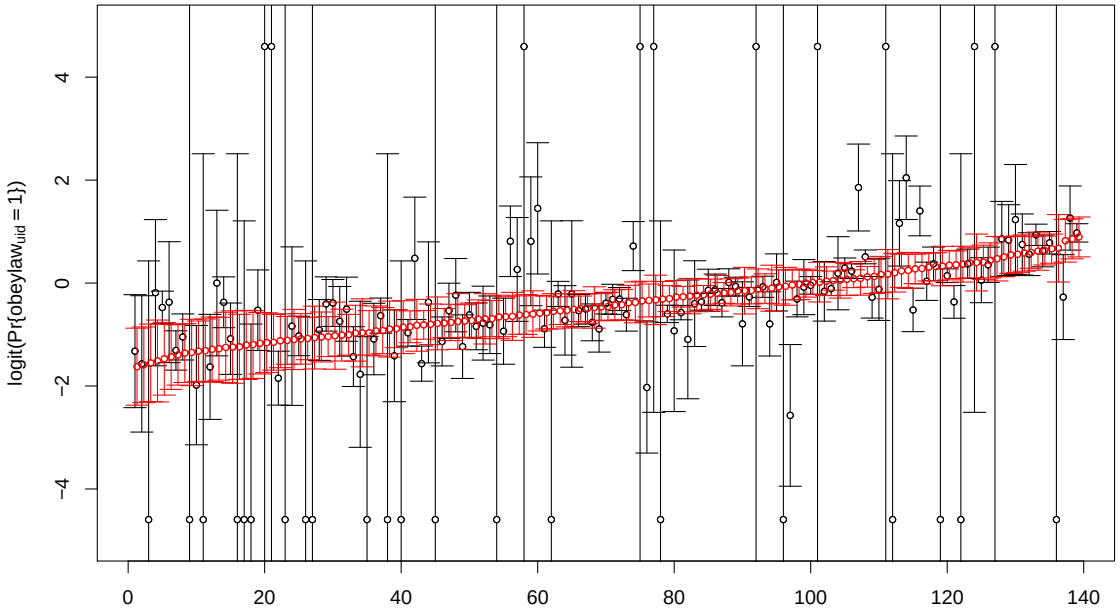
- Collins et al. (1992) report in their regression model of counts of non-compliance a significant coefficient of 0.163 on friends' non-compliance (3 point scale).

- Elffers et al. (1987) find a correlation $\rho = .22$ ($p < 0.01$) between the perceived prevalance of peer non-compliance (a 3-point scale) and 2-year self-report of evasion and the same correlation (i.e. identical ρ and p) perception of support (also a 3-point scale) and 2-year self-report. However, there is no significant correlation between these social covariates with documented non-compliance or amount of tax evaded.

[40] Refer to A.6.1 for the actual logistic regression model.



(a) Data Ordered



(b) Linear Model Ordered

Figure 8: Ordering of Obey Law Mean Probabilities. *The "obey law" mean probabilities (as logit) and s.d.'s are plotted alongside the linear model prediction for each of the 139 socio-demographic category, subscripted by a unique identifier (i.e. uid). In the upper plot, we order by the empirical means and, in the lower, by the linear model predictions. In neither plot does the x-axis map to the uid's.*

- Webley et al. (2001) find in their logit models that for Oslo (Norway) and Exeter (UK) respondents the perceived prevalence of friends' non-compliance (as a %) has a 0.38 ($p < 0.05$) effect on respondents' hypothetical non-compliance; for Paris and London respondents there is a similar 0.50 ($p < 0.05$) effect. Meanwhile, perceived peer support has a 0.76 ($p < 0.005$) effect on self-reported evasion for the former pair of sub-populations and a 0.65 ($p < 0.05$) effect on the hypothetical evasion for the latter pair.

- Vogel (1974) reports that 36.2% of respondents acquainted with non-compliant taxpayers also self-report non-compliance while only 21.8% of those who have no such acquaintances commit evasion: a 14.4% peer effect, not necessarily causal however.

While these findings are varied, they offer convincing evidence that peer effect has a prominent role in intention error commission. The integration of these findings into our error models will appear in future writings.

C Cauchy and Normal Priors on LR Coefficients

In the earlier line item models, we imposed a prior on the model coefficients in order to prevent them from obtaining degenerate values, say, $|\beta| > 5$. The first attempt employed a normal prior; however, Gelman et al. (2008) argues for Cauchy priors on logistic regression coefficients. In Figure 9, we compare our initial prior, $\mathrm{N}(\mu = 0, \sigma^2 = 9)$, to the Cauchy recommended by Gelman et al. We offer the intercept more leeway by using a slightly larger scale parameter, $\gamma = 3$, than the one recommended by Gelman, $\gamma = 2.5$, which we apply to the non-intercept coefficients. For further comparison, we display a normal with the equivalent variance ($\sigma^2 = 26$) of the latter Cauchy to demonstrate the Cauchy's narrower density.

D Background Work on Transforming Empirical Findings

D.1 Houston and Tran (2000)

The authors surveyed Australian adults and inferred from $n = 223$ of them that 16.5% of self-employed respondents have under-reported their income while only 3.6% of non-self-employed respondents have done so. They report the sizes of each group, $n_1 = 144$ for non-self-employed and 79 or self-employed, as well as the z-score to assess the significance of the difference in evasion rates: $z = 1.68$. Calculation of unbiased variance (σ^2) for each proportion, p, is straightforward:

$$\sigma^2_{\text{unbiased}} = p(1-p)/n$$

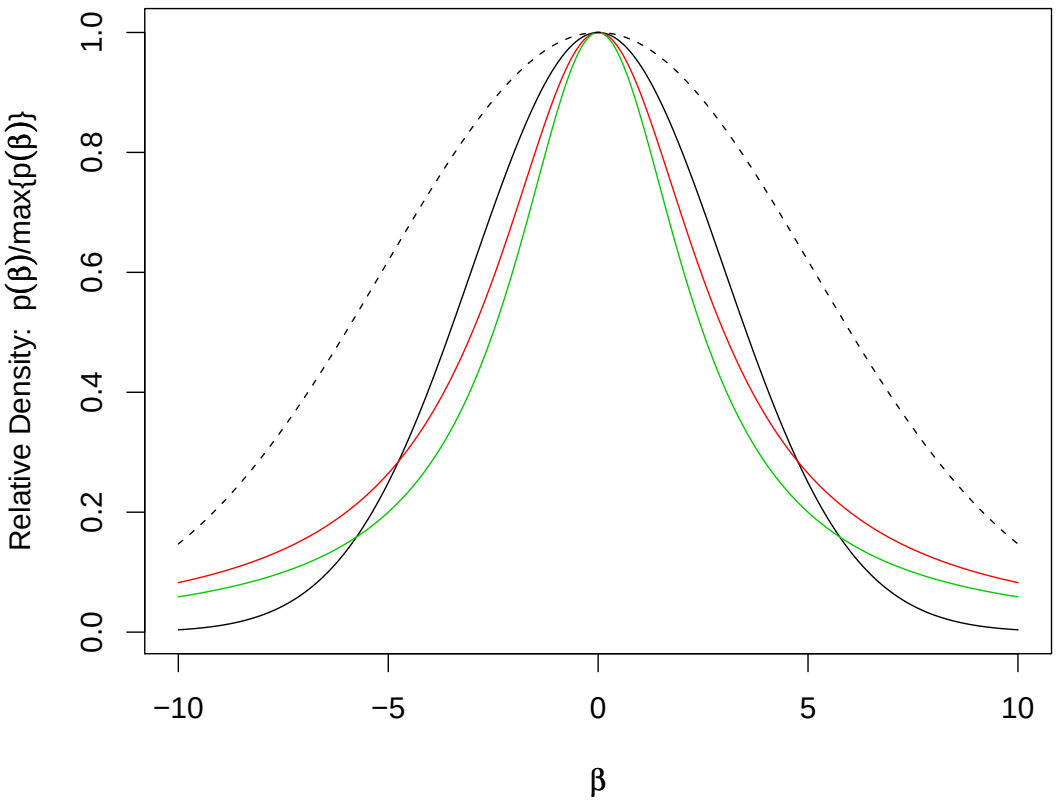


Figure 9: Cauchy and Normal Priors on β Coefficients. The black line denotes a normal distribution with standard deviation, $\sigma = 3$; the red and green lines denote Cauchy distributions with scale parameters, $\gamma = 3$ and 2.5, respectively; and the black dotted line denotes a normal distribution with the same variance as the latter Cauchy distribution.

The biased variance is not as straightforward due to their use of a randomized-response survey design for sensitive content. As such, the variance needs to account for the uncertainty in the chosen questions respondents answered:

$$\sigma^2_{\text{biased}} = \lambda(1-\lambda)/nq_s^2 \tag{11}$$

where $\lambda =$ the observed proportion of 'yes' responses and $q_s =$ the probability of a respondent answering the targeted sensitive question. Since

$$\lambda = pq_s + (1-q_s)q_{ns}$$

where $p =$ estimated proportion of evasion (from above) and $q_{ns} =$ the known proportion of 'yes' responses to the non-sensitive question, we obtain:

$$\lambda = p \cdot 0.7 + (1-0.7) \cdot \frac{1}{3} = 0.2155 \text{ (self-employed) and } 0.1252 \text{ (non-)}$$

and

$$\sigma_{\text{biased}} = 0.0661 \text{ and } 0.1252$$

Before inferring the regression coefficients, we transform the proportion with the logit function, $\log(x/1-x)$, which more accurately models the error surrounding a proportion. The uncertainty surrounding the proportions are also "logitized":

$$\sigma_{\text{biased,logit}} = \frac{\sigma_{\text{biased}}}{p\,(1-p)} = 0.480 \text{ and } 1.135$$

Since our covariates are simply 0 and 1, we can easily find α_0 and α_1:

$$\begin{aligned} \alpha_0 &= \text{logit}(p_0) \\ &= -3.288 \end{aligned} \qquad \begin{aligned} \alpha_1 &= \text{logit}(p_1) - \alpha_0 \\ &= \text{logit}(p_1) - \text{logit}(p_0) \\ &= 1.666 \end{aligned}$$

where p_0 and p_1 are the estimated proportions of evasion from non-self-employed and self-employed respondents, respectively.

Under normality, we have to assume constant variance; hence, we can obtain σ^2_α by appropriately weighting the variance surrounding the logits of p_0 and p_1:

$$\sigma^2_\alpha = \frac{n_0}{n}\sigma^2_0 + \frac{n_1}{n}\sigma^2_1$$

where σ_0 is the σ_{logit} for non-self-employed (either biased or unbiased) and σ_1 is for self-employed.[41] We obtain:

$$\begin{aligned} \sigma_{\alpha,\text{biased}} &= 0.9558 \\ \sigma_{\alpha,\text{unbiased}} &= 0.4021 \end{aligned}$$

[41] We can demonstrate the above numerically but have not managed to do so algebraically, as of yet.

D.2 Mason and Calvin (1978)

The authors employ a discriminant analysis to assess tax evasion behavior for 800 adults in Oregon. The respondents' ages were aggregated into six categories of values 1 to 6 and standardized. For their under-reporting of income model, the age effect is:

$$d_{\text{age}} = -0.83, F = 25.26$$

We estimate the error surrounding the discriminant coefficient using the F-statistic. The F-test is parameterized by two type of degrees of freedom. For a discriminant analysis they are calculated as:

$$\begin{aligned} df_1 &= m \cdot df_{\text{effect}} \\ df_2 &= s \cdot \left[df_{\text{error}} - \frac{m - df_{\text{effect}} + 1}{2} \right] - \left[\frac{m \cdot df_{\text{effect}} - 2}{2} \right] \end{aligned}$$

where $m = \#$ of predictor variables $= 6$ and $df_{\text{effect}} =$ (number of groups - 1). Since the prediction involves two groups, under-reporters vs. non-under-reporters, this is just $2-1 = 1$. $df_{\text{error}} =$ number of groups times $(n-1) = 2 \cdot 799 = 1598$. Also:

$$s = \sqrt{\frac{m^2 \cdot df_{\text{effect}}^2 - 4}{m^2 \cdot df_{\text{effect}}^2 - 5}}$$

We obtain $df_1 = 5$ and $df_2 = 1594$ which, when combined with the F-stat $= 25.62$, yield an extremely low p-value of 6.189×10^{-25}. The equivalent t-statistic is 10.49 corresponding to a standard error for d_{age}: $\sigma_{d_{age}} = 0.0791$.[42]

However, their analysis involved a standardized model which, for our purposes, needs to be "unstandardized". Without further details on the age categories or the quantities involved in the standardization, we are left to estimate these details with outside data. Specifically, we use the age distribution in the 1985 GSS to unstandarize the $d_{age} = -0.83$ effect.

$$\begin{aligned} \text{logit}(p_{underreport}) &= \hat{\beta}_0 + \frac{x_{\text{age}} - \mu_{\text{age}}}{\sigma_{\text{age}}} \cdot d_{\text{age}} \\ &= \left(\hat{\beta}_0 - \frac{\mu_{\text{age}}}{\sigma_{\text{age}}} \right) + \frac{x_{\text{age}} \cdot d_{\text{age}}}{\sigma_{\text{age}}} \\ &= \beta_0 + \frac{x_{\text{age}} \cdot d_{\text{age}}}{\sigma_{\text{age}}} \end{aligned}$$

where we use the GSS data to estimate the mean and standard deviation for the six age categories: $\mu_{age} = 2.11$ and $\sigma_{age} = 1.714$. We set:

$$\beta_1 = \frac{d_{\text{age}}}{\sigma_{\text{age}}} = \frac{-0.83}{1.714} = -0.484$$

[42] We maintain some reservations that this is the correct way to estimate the standard error, but it seems reasonable for now.

and

$$\sigma_{\beta_1} = \left|\frac{\beta_1}{t}\right| = \left|\frac{-0.484}{10.49}\right| = 0.0461$$

The intercept, β_0, is unknown, so we use the known information to model it:

$$\beta_0 \sim \mathrm{N}\left(\mathrm{logit}(p) - \beta_1 \cdot x_{\mathrm{age}}, \sigma_p^2 + \sigma_{d_{\mathrm{age}}}^2\right)$$

For the Newton-Raphson fit, we use the six age categories, coded in the GSS as 0 through 5, for x_{age} and compare the aggregated probability of under-reporting for each age group to the population rate of under-reporting, $p = 0.145$, reported by the authors in their paper; we employ the standard variance for a binomial: $\sigma_p = \sqrt{p(1-p)/n} = 0.01244$. The logitized quantities are $\mathrm{logit}(p) = -1.774$ and $\sigma_{\mathrm{logit}(p)} = 0.1004$.

We obtain $\beta_0 = -0.977$ and $\sigma_{\beta_0} = 0.1177$. With these, we confirm the population rate of under-reporting with our model:

$$\begin{aligned} 0.145 &= \sum_{i=0}^{5} \mathrm{logit}^{-1}\left(\beta_0 + \beta_1 \cdot i\right) \cdot q_i \\ &= \sum_{i=0}^{5} \mathrm{logit}^{-1}\left(-0.977 + -0.484 \cdot i\right) \cdot q_i \\ &= 0.1450057 \end{aligned}$$

where $\mathrm{logit}^{-1}(x)$ is the inverse logit function (i.e. $\exp(x)/(1+\exp(x))$) and q represents the GSS distribution of the age categories: $q \in \{0.222, 0.143, 0.146, 0.134, 0.126\}$.

D.3 Mason/Calvin Revisited

It turns out that a direct logistic regression treatment of discriminant analysis coefficients can only be appropriate if the structure of the permits it; and this appears to be the case

fortunately.

	LDA Model	LR Models	
Predictor	Empirical	Simulated #1	Simulated #2
Intercept		6.393***	6.339***
		(0.647)	(0.730)
Fear of Appreh.	−0.400***	−0.397***	−0.381***
	(0.084)	(0.049)	(0.058)
Income	−0.140	−0.135***	−0.134***
	(0.291)	(0.025)	(0.032)
Age	−0.830***	−0.803***	−0.801***
	(0.080)	(0.077)	(0.078)
Sex	−0.330**	−0.338ˆ	−0.344
	(0.117)	(0.194)	(0.234)
Occup. Prestige	−0.030	−0.029***	−0.029***
	(0.638)	(0.005)	(0.007)

We compute standard errors from the reported LDA coefficients and F-statistics using the method described in Appendix D.2. Also, we can directly predicted group membership (i.e. non-evader vs. evader) probabilities. We generate two kinds of simulated datasets: one that samples uniformly (#1) and another that samples from the GSS (#2), in order to maintain some empirical relationship between the predictors as well as achieve a distribution whose group means resemble the Mason/Calvin data.[43] The Mason/Calvin LDA coefficients tells us the predicted Bernoulli probabilities for group membership, which we simulate alongside the data set generation.

The resulting logistic regression (LR) coefficients surprisingly coincide with the LDA coefficients, which can occur for certain data structures.[44] However, the simulated significance differs greatly for some predictors; this is not a huge concern as long as we employ the empirical standard errors in our meta-analysis/imputation process.

D.4 Background Work on Wahlund's Correlation Inference

In this section, we attempted to continue the analytical inference of Wahlund's correlation between 'attitudes to crime' and 'tax evasion'. However, it became increasingly clear that the complexities inherent in this approach would forbid us from reaching a final estimate. Still, we report the analysis for illustrative purposes. In Figure 10, we display the lower portion of Wahlund's path diagram, segments of which we refer to in this analysis.

In the first stage, we want the full correlation between (o)*pportunity* and (p)*erceived opportunity*. However, the reported partial coefficient controls for the effects from (s)*elf-employment* and $(a)ge$; notationally, this is $(o, p|s, a)$. Hence, we will need to reintegrate the

[43]The GSS contains no information on Fear of Apprehension; hence we are unable to properly model this covariate.

[44]We explored different example coefficients and the closest explanation we have found is that a decent spread in the predicted probabilities (i.e. covering 0 to 1) results in similar LDA and LR coefficients.

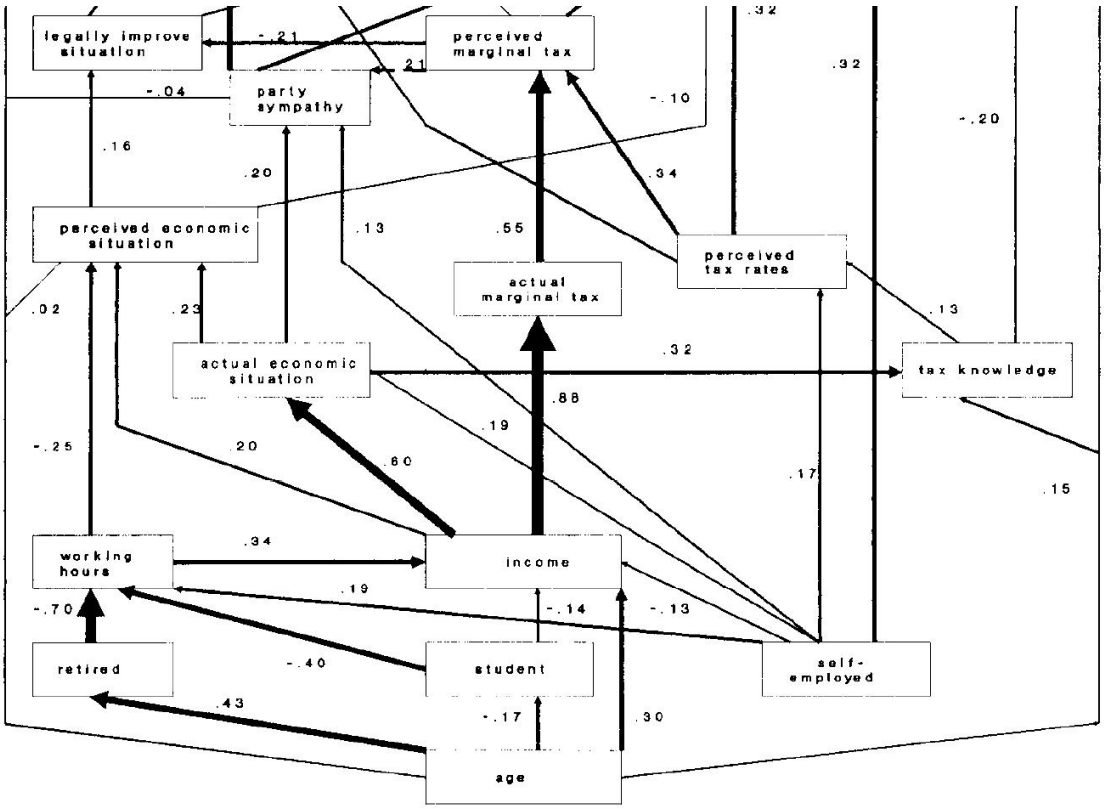


Figure 10: Bottom of Path Diagram from Wahlund (1992). *The "perceived marginal tax" atop points to "opportunity" in the earlier figure. The 0.32 coefficient from "Self-employed" is received by "perceived opportunity". The path on the far left leads from "age" to "opportunity". Reprinted without permisson.*

effect of $o \rightarrow p$ explained by a and s. We do this by backtracking the paths of $a \rightarrow o$ and $a \rightarrow p$. However, we are missing a direct path between s and o (i.e. $s \rightarrow o$ or $o \rightarrow s$), which in turn will account for some of the variance in $a \rightarrow s$, and we account for the missing effect with a placeholder, $\varepsilon_{o,s}$. These latter paths and others which we will address can be found in Fig. 10 appearing the end of this section.

$$\widehat{(o,p)} = (o,p|a,s) + (a,p|s) \cdot \widehat{(a,o)} + (s,p) \cdot \left[\widehat{(a,o)}(a,s) + \varepsilon_{o,s})\right]$$

where

$$\varepsilon_{o,s} = [(o,a)(a,s|o) + (o,s|a)] - (o,a)(a,s)$$

and $(a,s|o)$ and $(o,s|a)$ are hypothetical predictions of s from both o and a. Hence, the minuend represents the true effect while the subtrahend is what is reported by Wahlund. Since Wahlund does not report a direct self-employment/age correlation, (a,s), we employ the GSS' correlation of 0.075. Finally, $\widehat{(a,o)}$ is the total effect between *age* and *opportunity*, an unknown at this point.

We now solve for *age*'s full effect on *opportunity*, $\widehat{(a,o)}$. There is a set of lengthy paths from age to the "perceived marginal tax" (pmt) which is the source of the other, 0.10 partial correlation path into opportunity; however, each of these paths is narrow, meaning we can calculate the correlation between age and "perceived marginal tax" with a minimal number of unknowns. The intervening covariates of one path are student (st), income (in), "actual marginal tax" (amt). A second path includes "self-employed", "perceived tax rates" (ptr). Both these predict pmt. A third path takes us from "self-employed" through "tax knowledge" (tk) and into ptr.

We first solve for the total age effect on income:

$$\begin{aligned}\widehat{(a,in)} &= (a,in) + (a,st) \cdot (st,in|a) + (a,s) \cdot (s,in|a) \\ &= 0.30 + -0.17 \cdot -0.14 + 0.075 \cdot -0.13 \\ &= 0.314\end{aligned}$$

Next we, collapse "self-employed" and "tax knowledge" for a single age correlation to "perceived tax rates":

$$\begin{aligned}\widehat{(a,ptr)} &= (a,s) \cdot (s,ptr|tk) + (a,tk) \cdot (tk,ptr|s) + \varepsilon_{a,ptr} \\ &= 0.075 \cdot 0.17 + 0.15 \cdot 0.13 \\ &= 0.03225 + \varepsilon_{a,ptr}\end{aligned}$$

where $\varepsilon_{a,ptr}$ is variance/correlation left unexplained due to the absence of the (a,ptr) path. Again, this placeholder accounts for the difference between the reported effects and the path coefficients if there existed also a direct path between a and ptr. Given there is only one

path between income and "actual marginal tax", our estimate between age and *amt* also contains an unknown:

$$\begin{aligned}\widehat{(a, amt)} &= 0.88 \cdot \widehat{(a, in)} + \varepsilon_{a,amt} \\ &= 0.88 \cdot 0.314 + \varepsilon_{a,amt} \\ &= 0.27632 + \varepsilon_{a,amt}\end{aligned}$$

Now we can compute the correlation between age and "perceived marginal tax":

$$\begin{aligned}\widehat{(a, pmt)} &= \widehat{(a, amt)} \cdot (amt, pmt|ptr) + \widehat{(a, ptr)} \cdot (ptr, pmt|amt) + \varepsilon_{a,pmt} \\ &= (0.27632 + \varepsilon_{a,amt}) \cdot 0.55 + (0.03225 + \varepsilon_{a,ptr}) \cdot 0.34 + \varepsilon_{a,pmt} \\ &= 0.151976 + 0.010965 + 0.55\varepsilon_{a,amt} + 0.34\varepsilon_{a,ptr} + \varepsilon_{a,pmt} \\ &= 0.162941 + 0.55 \cdot \varepsilon_{a,amt} + 0.34 \cdot \varepsilon_{a,ptr} + \varepsilon_{a,pmt} \\ &= 0.162941 + \varepsilon_{\widehat{a,pmt}}\end{aligned}$$

where

$$\varepsilon_{\widehat{a,pmt}} = 0.55 \cdot \varepsilon_{a,amt} + 0.34 \cdot \varepsilon_{a,ptr} + \varepsilon_{a,pmt}$$

and $\varepsilon_{a,pmt}$ accounts for the lack of a direct path between a and pmt. Next, we compute the total age effect on opportunity:

$$\begin{aligned}\widehat{(a, o)} &= (a, o) + \widehat{(a, pmt)} \cdot (pmt, o) \\ &= -0.25 + \widehat{(a, pmt)} \cdot 0.10 \\ &= -0.25 + (0.162941 + \varepsilon_{\widehat{a,pmt}}) \cdot 0.10 \\ &= -0.2662941 + \varepsilon_{\widehat{a,pmt}} \cdot 0.10\end{aligned}$$

We can now insert $\widehat{(a, o)}$ into the expression for the total correlation between "opportunity" and "perceived opportunity", $\widehat{(o, p)}$:

$$\begin{aligned}\widehat{(o, p)} &= (o, p|a, s) + (a, p|s) \cdot \widehat{(a, o)} + (s, p) \cdot \left[\widehat{(a, o)}(a, s) + \varepsilon_{o,s})\right] \\ &= 0.25 + -0.11 \cdot (-0.2662941 + (0.10)\varepsilon_{\widehat{a,pmt}}) \\ &\quad + 0.23 \cdot \left[(-0.2662941 + (0.10)\varepsilon_{\widehat{a,pmt}}) \cdot 0.075 + \varepsilon_{o,s}\right] \\ &= 0.2746988 + [(-0.11)(0.10) + (0.23)(0.10)(0.075)]\varepsilon_{\widehat{a,pmt}} + (0.23)\varepsilon_{o,s} \\ &= 0.2746988 + (-0.009275)\varepsilon_{\widehat{a,pmt}} + (0.23)\varepsilon_{o,s}\end{aligned}$$

E Background Meta-Analysis

We combine findings from multiple papers which show tax evasion to be tied to attitudinal, behavioral, and socio-demographic covariates. While many of these studies focused primarily

on only a single type of predictor, we aim to construct a predictive model of tax evasion incorporating all three types of covariates. In order to achieve this goal, we conduct both a meta-analysis of the findings from these studies and imputation in applying these findings to our focal data set, the GSS.

We first offer a simplified model for continuous data to outline the approach as well as its complications. In this section, we consider three distinct data sources, two of which contains the dependent variable, tax evasion, and only one type of covariate, while the third contains just the two types of covariates found in the first two sources. The linear model equations here refer to the first two hypothetical data sources:

$$x = \alpha_0 + \alpha_1\theta + \varepsilon_\alpha \tag{12}$$
$$y = \beta_0 + \beta_1\phi + \varepsilon_\beta \tag{13}$$

In (12) and (13), x and y denote tax evasion behavior from the first two distinct data sources; θ is a significant attitudinal predictor and ϕ is a significant socio-demographic predictor. These models reflect, say, the works by Elffers et al (as well as Webley et al) and the summary of findings in Houston et al, respectively. The third data source lacks a tax evasion variable but contains both the attitudinal and socio-demographic covariates, θ and ϕ, much like the GSS. Accordingly, we can predict one of these covariates with the other to some, unknown, degree of significance:

$$\theta = \gamma_0 + \gamma_1\phi + \varepsilon_\gamma \tag{14}$$
$$\phi = \psi_0 + \psi_1\theta + \varepsilon_\psi \tag{15}$$

However, what we require is a model that predicts tax evasion from both covariates:

$$z = \delta_0 + \delta_1\theta + \delta_2\phi + \varepsilon_\delta \tag{16}$$

So now, we want to use (12) - (15) to determine the unknown parameters δ_0, δ_1, and δ_2, allowing us to predict z from θ and ϕ. First, ignoring the error terms (ε's) for the moment, we derive the coefficients in (12) and (13) in terms of the unknown δ's and the model of just the covariates (i.e. (14) and (15)). For example, when we substitute (14) into (16), we obtain:

$$z = \delta_0 + \delta_1(\gamma_0 + \gamma_1\phi) + \delta_2\phi \tag{17}$$

and after grouping the terms

$$z = (\delta_0 + \delta_1\gamma_0) + (\delta_2 + \delta_1\gamma_1)\phi$$

Essentially, we have derived a predictive model for just ϕ on tax evasion behavior, which is identical to (13), with the first term referring to the intercept β_0 and the second the slope coefficient, β_1. These substitutions yield the following equalities (again ignoring error for

now):

$$\alpha_0 = \delta_0 + \delta_2\psi_0 \tag{18}$$
$$\alpha_1 = \delta_1 + \delta_2\psi_1 \tag{19}$$
$$\beta_0 = \delta_0 + \delta_1\gamma_0 \tag{20}$$
$$\beta_1 = \delta_2 + \delta_1\gamma_1 \tag{21}$$

We have four equations to solve three unknowns δ's, hence we only require three of these equations and the fourth is superfluous. For example, using (19), we set δ_1 in terms of δ_2:

$$\delta_1 = \alpha_1 - \delta_2\psi_1$$

We obtain:

$$\begin{aligned}
\delta_2 &= \beta_1 - (\alpha_1 - \delta_2\psi_1)\gamma_1 \\
&= \beta_1 - \alpha_1\gamma_1 + \delta_2\psi_1\gamma_1 \\
\delta_2(1 - \psi_1\gamma_1) &= \beta_1 - \alpha_1\gamma_1 \\
\delta_2 &= \frac{(\beta_1 - \alpha_1\gamma_1)}{(1 - \psi_1\gamma_1)}
\end{aligned}$$

We can now obtain δ_0 and δ_1 using eqns. 18 and 19.

In order to obtain the standard errors for the δ's, we will first derive the distribution of residuals, ε_δ, by inserting the error terms into (17):

$$\begin{aligned}
z &= \delta_0 + \delta_1(\gamma_0 + \gamma_1\phi + \varepsilon_\gamma) + \delta_2\phi + \varepsilon_\delta \\
&= (\delta_0 + \delta_1\gamma_0) + (\delta_2 + \delta_1\gamma_1)\phi + \delta_1\varepsilon_\gamma + \varepsilon_\delta
\end{aligned} \tag{22}$$

Again, this equation resembles the prediction of tax evasion using only ϕ:

$$y = \beta_0 + \beta_1\phi + \varepsilon_\beta$$

We assume all the the residuals to be distributed normally, that is:

$$\varepsilon_\beta \sim \mathrm{N}(0, \sigma_\beta^2)$$
$$\varepsilon_\gamma \sim \mathrm{N}(0, \sigma_\gamma^2)$$
$$\varepsilon_\delta \sim \mathrm{N}(0, \sigma_\delta^2)$$

So, the variance of the residuals in (22) (i.e. $\delta_1\varepsilon_\gamma + \varepsilon_\delta$) is equivalent to the variance for the residuals ε_β (i.e. σ_β^2). And, since we can derive the variance for the sum of two independent normal distributions of known variance, we obtain:[45]

$$\sigma_\beta^2 = \delta_1^2\sigma_\gamma^2 + \sigma_\delta^2$$

[45]Proof can be demonstrated through convolution of two normal densities. Additionally, we consider α to be independent on ψ; that is, how θ relates to x is ex-ante independent of how θ relates to ϕ.

Similarly, we can also substitute ϕ with (15) into (16) and obtain:

$$\sigma_\alpha^2 = \delta_2^2 \sigma_\psi^2 + \sigma_\delta^2$$

giving us two equivalent forms for σ_δ^2:

$$\sigma_\delta^2 = \sigma_\alpha^2 - \delta_2^2 \sigma_\psi^2 = \sigma_\beta^2 - \delta_1^2 \sigma_\gamma^2$$

We can know compute the standard errors for the δ coefficients, by first computing the Fisher information matrix (showing only the diagonal and upper triangle since the matrix is implicitly symmetric):

$$\mathcal{I}(\boldsymbol{\delta}) = -\begin{bmatrix} \frac{n}{\sigma_\delta^2} & \frac{\sum \theta}{\sigma_\delta^2} & \frac{\sum \phi}{\sigma_\delta^2} \\ & \frac{\sum \theta^2}{\sigma_\delta^2} & \frac{\sum \theta\phi}{\sigma_\delta^2} \\ & & \frac{\sum \phi^2}{\sigma_\delta^2} \end{bmatrix}$$

where n is the number of data points and each summation occurs over the appropriate data (e.g. $\sum \theta = \sum_{i=1}^{n} \theta_i$). The inverse of $-\mathcal{I}(\boldsymbol{\delta})$, solved numerically, gives us the covariance matrix for the δ coefficients:

$$\Sigma_{\boldsymbol{\delta}} = \begin{bmatrix} \sigma_{\delta_0}^2 & \mathrm{Cov}(\delta_0, \delta_1) & \mathrm{Cov}(\delta_0, \delta_2) \\ & \sigma_{\delta_1}^2 & \mathrm{Cov}(\delta_1, \delta_2) \\ & & \sigma_{\delta_2}^2 \end{bmatrix}$$

Since our information sources comprises reported findings, we will often not have enough information to calculate the summations (e.g. $\sum \theta$ or $\sum \theta\phi$) in which case we will estimate them from alternate data source such as the General Social Survey.

E.1 Meta-Analysis with Generalized Linear Models

E.1.1 Two Covariates: Self-Employment and Age

Houston and Tran (2000) find self-employment to be a significant predictor of tax evasion, specifically under-reporting of income. Using their reported findings, we construct a logistic regression model to predict under-reporting contingent on a single covariate.[46]

			Unbiased	Biased
$\alpha_0 =$	$-3.288,$	$\sigma_{\alpha_0} =$	0.447	1.135
$\alpha_1 =$	$1.666,$	$\sigma_{\alpha_1} =$	0.540	1.232

The authors report that the variances around the proportions of tax evasion, for both the self-employed and non-self-employed groups, are inflated due to a mistake in their randomized response survey design. We report both the expected unbiased standard error (σ) as well as the biased one, based on the variance calculated from their paper. With the unbiased errors,

[46] Refer to Appendix D.1 for details.

both the intercept and coefficient are significant, while the biased errors render the effect of the covariate insignificant (t-statistic $= \frac{1.666}{1.232} = 1.352$).

Mason and Calvin (1978) conduct a discriminant analysis to assess the relationship between socio-demographic covariates and under-reporting of income. For this example, we focus on their standardized age variable, to inform our model. We determine the following unstandardized logistic model:[47]

$$\begin{aligned}\beta_0 &= -0.977, \quad \sigma_{\beta_0} = 0.1177 \\ \beta_1 &= -0.484, \quad \sigma_{\beta_1} = 0.0461\end{aligned}$$

Since self-employment (θ) is a binary variable, we need to model it appropriately:

$$\begin{aligned}\text{logit}(\theta) &= \gamma_0 + \gamma_1\phi \\ \theta &= \text{logit}^{-1}(\gamma_0 + \gamma_1\phi)\end{aligned}$$

or

$$\theta = \frac{e^{\gamma_0+\gamma_1\phi}}{1 + e^{\gamma_0+\gamma_1\phi}}$$

giving us:

$$\gamma_0 = -2.058, \quad \gamma_1 = 0.0831$$

Inference of the δ's is now complicated, since the equation:

$$z = \delta_0 + \delta_1 \cdot \text{logit}^{-1}(\gamma_0 + \gamma_1\phi) + \delta_2\phi$$

cannot be algebraically reduced to the form: $z = A + B\phi$. Similarly, the age categories constitute an ordinal variable, which we define with θ using an ordinal logistic regression, also known as a proportional odds model.

$$p(\phi <= i) = \text{logit}^{-1}(\psi_i^\alpha - \psi^\beta\theta)$$

So, instead, we numerically solve the δ's using both θ and ϕ predictions in the likelihood:

$$\delta_0 + \delta_1\theta + \delta_2\phi \sim \text{N}(\alpha_0 + \alpha_1\theta, \sigma^2_{\theta,\phi}) \cdot \text{N}(\beta_0 + \beta_1\phi, \sigma^2_{\phi,\theta})$$

Actually the likelihood is this:

$$\begin{aligned}L(\delta_0, \delta_1, \delta_2 | \alpha_0, \alpha_1, \beta_0, \beta_1) = \\ \prod_{\theta=0}^{1} \prod_{\phi=0}^{5} p(\delta_0 + \delta_1\theta + \delta_2\phi | \alpha_0 + \alpha_1\theta, \sigma^2_\phi) \cdot p(\delta_0 + \delta_1\theta + \delta_2\phi | \beta_0 + \beta_1\phi, \sigma^2_\theta)\end{aligned}$$

[47] Refer to Appendix D.2 for details.

where

$$\sigma_\phi^2 = 1/p_\theta(1-p_\theta)n_{\theta,\phi}^H$$
$$\sigma_\theta^2 = 1/p_\phi(1-p_\phi)n_{\phi,\theta}^C$$

and

$$n_{\theta,\phi} = n_\theta^H \cdot q_{\phi,\theta} \quad p_\theta = n_\theta/N_\theta^H$$
$$n_{\phi,\theta} = n_\phi^C \cdot q_{\theta,\phi} \quad p_\phi = n_\phi^C/N_\phi^C$$

and

$$q_{\phi,\theta} = \begin{cases} \text{logit}^{-1}(\psi_\phi^\alpha + \psi^\beta\theta) - 0 & \text{if } \phi = 0 \\ \text{logit}^{-1}(\psi_\phi^\alpha + \psi^\beta\theta) - \text{logit}^{-1}(\psi_{\phi-1}^\alpha + \psi^\beta\theta) & \text{if } 0 < \phi < 5 \\ 1 - \text{logit}^{-1}(\psi_\phi^\alpha + \psi^\beta\theta) & \text{if } \phi = 5 \end{cases}$$

and

$$q_{\theta,\phi} = \text{logit}^{-1}(\gamma_0 + \gamma_1\phi)$$

We can easily do this due to the discrete nature of θ and ϕ and easily derivable error around the logistic dependent variable:

$$\begin{aligned}\sigma_\phi &= \frac{1}{\sqrt{p(1-p)n}} \\ &= \{0.397, 0.202, 0.501, 0.249, 0.518, 0.269\}\end{aligned}$$

where

$$p = \text{logit}^{-1}(\beta_0 + \beta_1\phi)$$

and $\phi \in \{0,1,2,3,4,5\}$ (i.e. age categories) and $n \in \{183, 178, 115, 117, 107, 101\}$ (i.e. the GSS age category tabulation normalized by the number of respondents in the Mason/Calvin study ($n = 800$)). That is, each ϕ is associated with a unique logit variance σ_ϕ^2. We compute σ_θ similarly and obtain $\{0.447, 0.303\}$.

We now have sufficient information to infer our logistic regression model of under-reporting incorporating the findings from the Houston and Tran (2000) and Mason and Calvin (1978) papers, θ and ϕ, respectively. We report results using both the unbiased and biased variances from the Houston and Tran study.

	δ Coefficients for		
Predictor	Unbiased	Biased	Sampled
Intercept	-1.161^{***}	-1.039^{***}	-1.156^{***}
	(0.137)	(0.142)	(0.137)
$\theta_{\text{S.E.}}$	0.256	0.125	0.257
	(0.240)	(0.276)	(0.240)
ϕ_{Age}	-0.399^{***}	-0.444^{***}	-0.402^{***}
	(0.067)	(0.073)	(0.067)
$\mathcal{L}$	-22.249	-23.980	-22.096
n	1023	1023	1023
p_{GSS}	0.1404	0.1455	0.1404
	(0.0109)	(0.0110)	(0.0190)

Houston and Tran's findings imply 8.2% of the population commit under-reporting while Mason and Calvin's paper report 14.5%; some of this difference may be due to different nationalities of the respondents (Australian vs. American) and/or the year of the study (2000 vs. 1978). Also, according to the GSS, self-employment and age category are almost independent; the correlation is quite low ($\rho = 0.048$) yet mildly significant ($p < 0.10$). Hence, we see that only one of the coefficients resemble its original value, namely ϕ and β_1.

Fitting the model to the GSS data (p_{GSS}), we obtain the projected proportions of under-reporting in the GSS sample, 14.04% and 14.55%.[48] These are similar to the proportion of under-reporting reported by Mason and Calvin; as a result of a large sample size, their parameters exhibit higher confidence. Another reason for the high proportion is that there are relatively fewer self-employed respondents in the GSS compared to the Houston/Tran data. Also, we note that the unbiased model exacts tighter errors bounds for Houston data resulting in the overall proportion of under-reporting drifting down towards the proportion reported in the Houston paper.

E.1.2 Sampled Results

Alternatively, we can estimate the δ coefficients via Monte Carlo sampling. We first draw sample of age categories, ϕ, for the Houston data set as well as the under-reporting committed by its respondents, x, using the known paramters. Then, we draw a sample of the self-employment covariate, θ, for the Mason/Calvin data as well as the under-reporting for its respondents, y. We then apply a logistic regression on the combination data set. The procedure conforms to the following generalized linear models:

$$\begin{aligned}
\text{logit}(x)|\theta &\sim \text{N}(\alpha_0 + \alpha_1\theta, [\sigma_\theta^H]^2) \\
\phi^{\text{H}}|\theta &\sim \text{Multinom}(\boldsymbol{q}_{.,\theta}, \boldsymbol{n}_{\theta,.}) \\
\text{logit}(y)|\phi &\sim \text{N}(\beta_0 + \beta_1\phi, [\sigma_\phi^C]^2) \\
\text{logit}(\theta^{\text{C}})|\phi &\sim \text{N}(\gamma_0 + \gamma_1\phi, [\sigma_\phi^C]^2)
\end{aligned}$$

Also, we can substitute draws from the data reflecting empirical proportions, rather than the GLMs, to obtain our ϕ^{H} and θ^{C} since all covariates are discrete. We display the $\boldsymbol{\delta}$ estimated through this additional layer of sampling alongside our findings from the GLM specification. We perform 1000 of draws of each of the procedures to estimate $\boldsymbol{\delta}$ and its accompanying

[48] These are all unweighted results; we will look at weighted data later.

standard errors, $\sigma_{\boldsymbol{\delta}}$:

Predictor	GLM	Bootstrapped ϕ^H, θ^C
Intercept	-1.270^{***} (0.138)	-1.261^{***} (0.140)
$\theta_{\text{S.E}}$	0.259 (0.246)	0.258 (0.246)
ϕ_{Age}	-0.397^{***} (0.068)	-0.401^{***} (0.066)
n	1023	1023
p_{GSS}	0.1285 (0.0105)	0.1289 (0.0105)

The coefficients between the analytical model and the sampled/bootstrapped models differ due to low probabilities, both predicted and empirical, associated with the occurrence some θ, ϕ combinations.

E.1.3 Three Covariates: Obey Law

We introduce a third covariate that reflects respondents' attitudes towards the law. Wahlund finds respondents' attitudes to crime to be correlated to tax evasion behavior; specifically, a more lax attitude is associated with higher likelihood of evading taxes. The General Social Survey captures this attitude in its Obey Law item in which respondents are asked:

> "In general, would you say that people should obey the law without exception, or are there exceptional occasions on which people should follow their consciences even if it means breaking the law?"

In Appendix A.5, we infer parameters for predicting tax evasion, which we can now incorporate into our predictive model. Each of the predictive GLMs are updated to include this new covariate, ω:

$$\begin{aligned} \theta &= \text{logit}^{-1}(\gamma_0 + \gamma_1\phi + \gamma_2\omega) \\ p(\phi <= i) &= \text{logit}^{-1}(\psi_i^{\alpha} - \psi_1^{\beta}\theta - \psi_2^{\beta}\omega) \\ \omega &= \text{logit}^{-1}(\lambda_0 + \lambda_1\theta + \lambda_2\phi) \end{aligned}$$

The compound model has the following distribution:

$$\delta_0 + \delta_1\theta + \delta_2\phi + \delta_3\omega \sim \text{N}(\alpha_0 + \alpha_1\theta, \sigma^2_{\theta,\{\phi,\omega\}}) \cdot \text{N}(\beta_0 + \beta_1\phi, \sigma^2_{\phi,\{\theta,\omega\}}) \cdot \text{N}(\tau_0 + \tau_1\omega, \sigma^2_{\omega,\{\theta,\phi\}})$$

We present the analytical and bootsrapped results:

Predictor	MLE	Bootstrapped
Intercept	-1.190^{***} (0.124)	-1.257^{***} (0.124)
$\theta_{\text{Self-Employed}}$	0.125 (0.193)	0.165 (0.188)
ϕ_{Age}	-0.213^{***} (0.049)	-0.237^{***} (0.049)
ω_{ObeyLaw}	-0.489^{**} (0.165)	-0.534^{**} (0.168)
$\mathcal{L}$	-83.284	NA
n	1623	1623
p_{GSS}	0.1450 (0.0144)	0.1314 (0.0139)

Due to the collinearity between ϕ and ω, $\rho = 0.181$ ($p < 0.001$), their effects are diminished from original models.

F Earlier Meta Analysis Results

	$n_{\text{Collins}} = 9482$		$n_{\text{Collins}} = 240$	
Predictor	Model #1a	Model #1b	Model #2a	Model #2b
Intercept	-0.465^{***} (0.093)	-0.403^{***} (0.100)	-1.070^{***} (0.136)	-1.021^{***} (0.140)
Sex	0.287^{***} (0.080)	0.313^{***} (0.081)	0.419^{**} (0.137)	0.412^{***} (0.112)
Age	-0.532^{***} (0.074)	-0.544^{***} (0.074)	-0.599^{***} (0.111)	-0.604^{***} (0.105)
Education	-0.219^{***} (0.055)		0.030 (0.133)	
|Education − 1|		-0.356^{***} (0.104)		-0.121 (0.110)
Income	-0.047 (0.037)	-0.089^{**} (0.034)	0.098 (0.140)	$0.127^{\hat{}}$ (0.068)
$\mathcal{L}$	-160	-162	-102	-101
n	10000	10000	10000	10000

	Summary Statistics for $p(y<0)$					
	Min.	25%	Median	Mean	75%	Max.
Model #1a	0.0819	0.1814	0.2288	0.2501	0.3151	0.4557
Model #1b	0.0662	0.1791	0.2452	0.2547	0.3075	0.4774
Model #2a	0.0882	0.1764	0.2457	0.2435	0.3006	0.4990
Model #2b	0.0780	0.1828	0.2420	0.2444	0.3035	0.5386
Previous	0.0000	0.0200	0.0962	0.2507	0.4231	0.9796

G Background Line Item Analysis

Before we offer our likelihood model, we need to estimate some empirical rate of error incurred through the 'Misc' category of line-items, that is some line-item not covered by the seven explicit taxpayer categories. In the left equation below, we estimate 'Misc' error commission using non-categorical agents in our subsample, and in the right, we estimate, using the free parameter α_0^{Misc}, the predicted rate of 'Misc' error commission among categorical agents:

$$\overline{p}^{\text{Misc}} = \frac{\sum_i (p_i \cdot \mathcal{I}_i^p)}{\sum_i \mathcal{I}_i^p} \quad \text{and} \quad \overline{q}^{\text{Misc}} = \frac{\sum_i \left(q_i^{\text{Misc}} \cdot \mathcal{I}_i^q\right)}{\sum_i \mathcal{I}_i^q}$$

where we employ indicators, $\mathcal{I}_i^p$ and $\mathcal{I}_i^q$, to sequester non-categorical and categorical taxpayers, respectively. In short, the sum of the former indictaor yields the count of non-categorical taxpayers while the sum of the latter yields the count of categorical taxpayers:

$$\mathcal{I}_i^p = \begin{cases} 0 & \text{if } \sum_{tp} x_{i,tp} > 0 \\ 1 & \text{if } \sum_{tp} x_{i,tp} = 0 \end{cases} \quad \text{and} \quad \mathcal{I}_i^q = \begin{cases} 0 & \text{if } \sum_{tp} x_{i,tp} = 0 \\ 1 & \text{if } \sum_{tp} x_{i,tp} > 0 \end{cases}$$

We compute $\overline{p}^{\text{Misc}}$ from our sampled data and estimate it to be 0.114. Next, we notationally assign $\boldsymbol{P} = (\ldots, p_i, \ldots,)$ and $\boldsymbol{Q} = (\ldots, q_i, \ldots)$, where $i \in \arg(\mathcal{I}_i^q = 1)$, restricting our analysis to only categorical taxpayers.[49] There are two components of the likelihood model. The first piece,

$$\text{logit}[\boldsymbol{Q}] \sim \text{N}\left(\text{logit}[\boldsymbol{P}], \frac{1}{\boldsymbol{P}(1-\boldsymbol{P})}\right) \tag{23}$$

fits the combined rate of taxpayer categorical errors, q_i, as determined by all the free α parameters to the predicted error rates, p_i, from the intentional model.[50] Concurrently, the second component:

$$\overline{q}^{\text{Misc}} \sim \text{N}\left(\overline{p}^{\text{Misc}}, \frac{\overline{p}^{\text{Misc}} \cdot (1-\overline{p}^{\text{Misc}})}{\sum_i \mathcal{I}_i^p}\right) \tag{24}$$

fits the mean of the 'Misc' error rate of the categorical taxpayers to that of non-categorical taxpayers, while employing the standard deviation of a typical binomial. Below, we offer the

[49] This notation for collecting the appropriate i's is my own shorthand and will require amendment.

[50] We employ the canonical variance for the logit of a probability, p: $\frac{1}{p(1-p)n}$. However, since we are fitting each data point individually, our n becomes 1.

maximum-likehood parameter fits for the first model, using (23). Again, these parameters correspond to the α coefficients from (5) and our model fits to the data for agents who fall under at least one of the taxpayer categories; hence the sample size of these models ($n = 8{,}251$) is less than the size of our original sample ($n = 10{,}000$):

	Taxpayer/Line-Item Categories						
α/Predictor	Tips	SEmp	EIC	SLns	Cap	Frm	SSB
Intercept	−2.320	−1.369	0.679***	−0.544	−1.769**	−1.450	−1.775
	(1.767)	(1.139)	(0.148)	(0.834)	(0.613)	(2.459)	(1.142)
Sex	1.006	1.611ˆ	1.164***	1.780***	1.449**	1.351	1.396***
	(1.482)	(0.936)	(0.128)	(0.436)	(0.523)	(2.352)	(0.303)
Age	0.072	−0.187	−0.348***	−0.387	−0.250	−0.354	0.382
	(1.044)	(0.571)	(0.083)	(0.360)	(0.350)	(1.515)	(0.558)
Education	−1.877	−1.886***	−1.827***	−2.293***	−1.716***	−1.766	−1.842***
	(1.423)	(0.555)	(0.112)	(0.667)	(0.374)	(1.743)	(0.274)
Income	0.060	−0.369	−0.817***	−0.431*	−0.134	−0.225	−0.377*
	(0.761)	(0.274)	(0.081)	(0.172)	(0.198)	(0.962)	(0.176)
α_0^{Misc}	−0.0213	$\overline{q}^{\text{Misc}}$	0.133	$\mathcal{L}$	−20294		
	(0.112)		(0.210)	n	8251		

Alternatively, we can express the first piece of the likelihood using only the taxpayer categories, requiring us to adjust the predicted probability of any error to include the probability of a 'Misc' error:[51]

$$\text{logit}\left[\boldsymbol{Q}^{\text{-Misc}}\right] \sim \text{N}\left(\text{logit}\left[\frac{1-\boldsymbol{P}}{1-\boldsymbol{Q}^{\text{Misc}}}\right], \frac{(1-\boldsymbol{Q}^{\text{Misc}})}{(1-\boldsymbol{P})(\boldsymbol{P}-\boldsymbol{Q}^{\text{Misc}})}\right) \tag{25}$$

[51] Initially, we implemented this alternative model because a) it worked and the other did not and b) initially, we did not parameterize the 'Misc' intercept α_0^{Misc}, and instead provided a fixed estimate which arguably should be included on the right side of the equation. However, since now both approaches work and give slightly different results, we need more compelling arguments for selecting one over the other.

Here are the results when we employ the alternative model, using (25), in lieu of (23):

	Taxpayer/Line-Item Categories						
α/Predictor	Tips	SEmp	EIC	SLns	Cap	Frm	SSB
Intercept	−2.098*	−1.346ˆ	0.630***	−0.467	−1.875***	−1.375	−1.493*
	(1.027)	(0.770)	(0.086)	(0.639)	(0.415)	(1.527)	(0.636)
Sex	0.691	1.488*	1.056***	1.765***	1.336***	1.076	1.359***
	(0.936)	(0.622)	(0.087)	(0.318)	(0.353)	(1.431)	(0.223)
Age	0.193	−0.193	−0.324***	−0.388	−0.217	−0.438	0.263
	(0.687)	(0.411)	(0.060)	(0.272)	(0.240)	(1.093)	(0.308)
Education	−2.110*	−1.814***	−1.755***	−2.244***	−1.613***	−1.761	−1.825***
	(0.996)	(0.392)	(0.080)	(0.513)	(0.259)	(1.151)	(0.212)
Income	0.132	−0.285	−0.744***	−0.438**	−0.002	−0.044	−0.348*
	(0.539)	(0.209)	(0.058)	(0.139)	(0.146)	(0.683)	(0.144)
α_0^{Misc}	−0.306**	$\overline{q}^{\text{Misc}}$	0.114	$\mathcal{L}$	−20397		
	(0.117)		(0.190)	n	8251		

With the exception of several insignificant coefficients, most of them differ by only a few percentage points across the two models. Still, while the first model based on (23) yields a higher likelihood, the second, alternative model, based on (25), not only offers more significant coefficients but the mean probability of commission in 'Misc' line-items (i.e. $\overline{q}^{\text{Misc}}$) falls closely to the rate inferred for non-categorical agents (i.e. $\overline{p}^{\text{Misc}} = 0.114$). While it is tempting to recommend the second model over the first, the straightforward nature of the first model also makes it a compelling candidate. At this point, the only explanation we have for the more tightly bound coefficients in the alternative model is that the probabilities to be fitted are such that the median variance is smaller than the primary model.

G.1 Line Item Model Revisited

Here, we offer an earlier line item model with Tips that assumed a restrictive marginal error difference:

$$\overline{\Delta p}_{tp,uid} \sim \text{N}(\mu = 0.05, \sigma^2 = 0.02^2)$$

where $\Delta p_{tp,uid} = p_{tp,uid} - p_{-tp,uid}$

	MLE Results		SA Results	
Variable	Tips	Misc	Tips	Misc
Intercept	−2.623***	−1.092***	−1.602***	−1.088***
	(0.145)	(0.039)	(0.313)	(0.053)
Sex	0.108	0.427***	0.641*	0.434***
	(0.120)	(0.033)	(0.296)	(0.048)
Age	−0.142ˆ	−0.598***	−0.397ˆ	−0.580***
	(0.081)	(0.022)	(0.218)	(0.037)
Educ	0.008	0.034ˆ	0.475ˆ	0.062*
	(0.073)	(0.018)	(0.244)	(0.031)
Inc	0.024	0.099***	0.180	0.151***
	(0.046)	(0.014)	(0.156)	(0.019)
n	10000	10000	10000	10000

G.2 Using Opportunities

Houston and Tran and Vogel both predict an approximately 10% increase in tax evasion for self-employed individuals. One simple approach is to apply this difference across all opportunity categories (i.e. taxpayer categories). However, Wahlund's correlation of 0.21 confers up to a possible 37% increase in evasion due to self-employment, when we incorporate the 8.5% rate of self-employment in our PUMS sample; when we employ the inferred correlation, via simulation, of $\rho = 0.329$, the increase in error reduces 20%, under Vogel's self-employment rate of 9.9%. However, when we assume that no opportunity strictly yields no error, we find the error increase to be either 40% or 47% depending on which ρ we employ.[52] The following models differ in the uncertainty surrounding this 10% difference, as controlled by the n portion of the standard deviation around a proportion, denoted as n_{LI}:

1. $n_{\text{LI}} = 1$

Variable	Tips	SEmp	EIC	SLns	Cap	Frm	SSB	Misc
Intercept	−1.48	−0.32	−3.22**	−2.41*	−1.95ˆ	−0.53	−1.01	−1.38***
	(1.48)	(1.16)	(0.90)	(0.94)	(1.14)	(1.75)	(1.70)	(0.16)
Sex	−0.19	−0.07	−0.01	0.12	0.55	0.12	−0.26	0.44***
	(1.84)	(1.19)	(1.41)	(0.76)	(1.06)	(1.85)	(1.53)	(0.08)
Age	−0.82	−0.81	−0.73	−0.50	−1.11	−0.74	−1.54	−0.53***
	(1.54)	(0.91)	(1.32)	(0.67)	(1.26)	(1.46)	(1.02)	(0.10)
Education	−0.52	−0.49	−0.72	−0.18	−0.96	−0.10	−0.07	0.05
	(1.62)	(0.68)	(1.37)	(0.58)	(1.20)	(1.52)	(1.04)	(0.05)
Income	−0.76	−0.13	−0.83	−0.06	−0.17	−0.03	−0.31	0.12***
	(1.62)	(0.36)	(1.67)	(0.29)	(0.45)	(1.09)	(0.82)	(0.03)

[52] Refer to Appendix A.5.3 for the analysis.

2. $n_{\text{LI}} = 2$

Variable	Tips	SEmp	EIC	SLns	Cap	Frm	SSB	Misc
Intercept	−1.46	−0.30	−3.23**	−2.41*	−1.94ˆ	−0.58	−1.00	−1.39***
	(1.47)	(1.15)	(0.89)	(0.94)	(1.13)	(1.71)	(1.70)	(0.15)
Sex	−0.18	−0.08	−0.02	0.11	0.56	0.10	−0.25	0.44***
	(1.84)	(1.18)	(1.41)	(0.76)	(1.05)	(1.87)	(1.52)	(0.08)
Age	−0.81	−0.81	−0.72	−0.49	−1.13	−0.73	−1.53	−0.53***
	(1.53)	(0.90)	(1.31)	(0.66)	(1.26)	(1.48)	(1.01)	(0.10)
Education	−0.51	−0.49	−0.71	−0.18	−0.95	−0.16	−0.06	0.05
	(1.61)	(0.67)	(1.37)	(0.58)	(1.19)	(1.52)	(1.01)	(0.05)
Income	−0.73	−0.13	−0.86	−0.06	−0.18	−0.08	−0.31	0.12***
	(1.61)	(0.36)	(1.67)	(0.28)	(0.45)	(1.13)	(0.80)	(0.03)

3. $n_{\text{LI}} = 15$

Variable	Tips	SEmp	EIC	SLns	Cap	Frm	SSB	Misc
Intercept	−1.24	−0.26	−3.22**	−2.45*	−1.83ˆ	−0.65	−0.79	−1.42***
	(1.42)	(1.15)	(0.88)	(0.90)	(1.05)	(1.65)	(1.67)	(0.14)
Sex	−0.03	−0.12	−0.10	0.08	0.67	0.06	−0.11	0.45***
	(1.86)	(1.20)	(1.34)	(0.71)	(0.91)	(1.88)	(1.35)	(0.09)
Age	−0.81	−0.82	−0.61	−0.41	−1.18	−0.73	−1.54	−0.52***
	(1.45)	(0.89)	(1.15)	(0.59)	(1.19)	(1.50)	(0.96)	(0.10)
Education	−0.31	−0.50	−0.61	−0.11	−0.90	−0.24	0.04	0.05
	(1.44)	(0.67)	(1.30)	(0.49)	(1.07)	(1.52)	(0.75)	(0.05)
Income	−0.47	−0.12	−1.08	−0.04	−0.19	−0.14	−0.30	0.12***
	(1.18)	(0.35)	(1.56)	(0.25)	(0.40)	(1.11)	(0.61)	(0.03)

4. $n_{\text{LI}} = 900$

Variable	Tips	SEmp	EIC	SLns	Cap	Frm	SSB	Misc
Intercept	0.39	0.94	−3.50***	−2.21***	0.54	−0.58	4.15***	−2.56***
	(1.06)	(1.31)	(0.34)	(0.38)	(0.44)	(1.75)	(0.88)	(0.16)
Sex	−0.51	−1.02	0.92***	0.28	0.26	−0.38	0.47ˆ	0.41***
	(1.63)	(1.03)	(0.23)	(0.21)	(0.60)	(1.86)	(0.23)	(0.10)
Age	−2.01ˆ	−1.93ˆ	−1.10***	−0.68**	0.20	−0.74	−2.66***	−0.31**
	(1.16)	(0.94)	(0.25)	(0.24)	(0.42)	(1.53)	(0.38)	(0.11)
Education	−0.27	−0.32	−0.07	0.15	−0.86**	−0.02	0.11	0.07
	(1.05)	(0.61)	(0.14)	(0.14)	(0.31)	(1.50)	(0.11)	(0.05)
Income	−0.93	−0.06	1.22***	0.05	−4.08***	−0.31	−0.64**	0.26***
	(0.67)	(0.31)	(0.16)	(0.10)	(0.95)	(1.12)	(0.19)	(0.04)

$n_{\text{LI}} = 150$

Variable	Tips	SEmp	EIC	SLns	Cap	Frm	SSB	Misc
Intercept	−0.51	−0.13	−3.75***	−2.45***	−0.34	−0.60	1.13	−1.81***
	(1.26)	(1.32)	(0.58)	(0.59)	(0.62)	(1.70)	(1.15)	(0.17)
Sex	−0.13	−0.46	0.78ˆ	0.24	0.09	−0.14	0.46	0.42***
	(1.76)	(1.13)	(0.39)	(0.38)	(0.71)	(1.87)	(0.43)	(0.10)
Age	−1.24	−1.09	−0.95*	−0.60	−0.06	−0.71	−1.87**	−0.46***
	(1.25)	(0.95)	(0.39)	(0.39)	(0.61)	(1.52)	(0.53)	(0.11)
Education	−0.05	−0.34	−0.14	0.07	−0.75	−0.22	0.13	0.05
	(1.11)	(0.69)	(0.27)	(0.24)	(0.47)	(1.56)	(0.22)	(0.06)
Income	−0.60	−0.05	0.97***	0.02	−2.66*	−0.21	−0.38	0.17***
	(0.79)	(0.34)	(0.24)	(0.15)	(1.04)	(1.10)	(0.26)	(0.03)

H Schemes and Credits

We briefly explored specific schemes and/or credits that taxpayers have been known to be involved in. The presumptive sociodemographic constraints for each of the schemes and credits (*as outlined in Brian's slides*) are as follows:

Scheme/Credit	< $30K	College+	Black	Children	Middle-Aged–
Generic Scheme					
Slavery Reparation	+		+		
Home-Based Office	–	+			
Generic Credit					
EITC	+			+	
Education Credit		+			+

The operationalization of each of these constraints as an indicator variable/flag requires a heavy negative coefficient for the antithesis of the constraint . For example, a respondent who is not black will not be involved with the 'slavery reparation' scheme; hence, the reverse of the indicator (being non-black) will have a high negative coefficient, say −40, so that the logistic regression will render the probability nil. Being black is tantamount to no negative contribution from the (reversed) indicator, so that the rest of the coefficients can determine the probability of involvement.

Enumerating each of these constraints into separate flags will allow for coding of a single model that expects coefficients for the general set of sociodemographic coefficients plus all the flags. Otherwise, we would require either several models or coding of additional 'if-then' conditions for each of the schemes/credits. The coefficients, particularly the ones for the intercept and the flags, will vary per scheme/credit prediction. Hence the flags should be as follows:

Income	Education	Race	# Children	Age
< $30K, > $30K	< College	Not Black	No Children	> Middle Age

The β coefficients on these flags will be some large negative value, like -40. Note that we require two flags for income; the first one is relevant to the 'home-based office' scheme, while the second is revelant to both both 'slavery reparation' and EITC. If we allow all acts of non-compliance to fall under one of the six scheme/credit categories, the total probability of at least one scheme/credit occurring should approach our population estimate of 25%:

$$p_{\text{evade}}^{\text{pop}} \approx 1 - \prod_{i=1}^{6} \left(1 - p_i^{\text{scheme/credit}}\right)$$

H.1 Two Variable Example

This section underscores the basics of the inference process using just the covariates for sex, income, and race to infer models for involvement in one of the schemes and one of the credits: the 'slavery reparation' scheme and EITC credit abuse. We first start with a truncated version of the overall intentional error model, which predicts the logit of the probability of one or more acts of non-compliance:

$$\beta_0 = -1;\ \beta_{\text{male}} = 1.6;\ \beta_{\text{income}} = -0.96$$

where $x_{\text{income}} \in \{0, 1, 2, 3, 4\}$ which corresponds to the income levels:

0	1	2	3	4
\$0 – \$15K	\$15 – \$30K	\$30 – \$50K	\$50 – \$80K	> \$80K

The above coefficients yield a population intentional error rate of 25.8%.[53]

The predictive models for 'slavery reparation' (SR) and EITC credit are as follows:

$$\begin{aligned}
\text{logit}(p_{\text{SR}}) &= \alpha_0 + \alpha_1 \cdot x_{\text{male}} + \alpha_2 \cdot x_{\text{income}} + \alpha_3 \cdot \mathcal{I}(x_{\text{race}} \neq \text{Black}) + \alpha_4 \cdot \mathcal{I}(x_{\text{income}} > 1) \\
\text{logit}(p_{\text{EITC}}) &= \gamma_0 + \gamma_1 \cdot x_{\text{male}} + \gamma_2 \cdot x_{\text{income}} + \gamma_3 \cdot \mathcal{I}(x_{\text{race}} \neq \text{Black}) + \gamma_4 \cdot \mathcal{I}(x_{\text{income}} > 1)
\end{aligned}$$

in which we set the flag coefficients $\alpha_3 = \alpha_4 = \gamma_4 = -40$ and $\gamma_3 = 0$, since race matters only to slavery reparation while income matters to both. We now set up the likelihood model to solve for the α's and the γ's:

$$\begin{aligned}
\text{logit}(p_{\text{SR}}) &\sim \text{N}(\text{logit}(\mu_{\text{SR}} = 1/1290 \approx 0.001), \sigma^2 = 1001) \\
\text{logit}(p_{\text{EITC}}) &\sim \text{N}(\text{logit}(\mu_{\text{EITC}} = 4.5/129 = 0.0349), \sigma^2 = 29.7)
\end{aligned}$$

We take a moment to explain our mean parameters μ_{SR} and μ_{EITC}. In 2000, the IRS processed 129 million individual returns. And, according to one source, the 2001 estimate of the number of returns containing attempts to obtain 'slavery reparation' refunds is 100,000; we round the proportion to the nearest thousandth.[54] Furthermore, in 1999, there were an estimated

[53] Here, we employ the 1985 General Social Survey as our source of covariance.

[54] We do this, in part, because we can find a solution with 0.001; so far, the Newton-Raphson does not converge for the actual estimate.

9 million returns containing EITC overclaims; we apply earlier findings from literature that about half the errors (4.5 million) are intentional and the other half, inadvertent. The high variance for each distribution is warranted considering our likelihood model fits each data point, rather than a summary statistic such as average population commission of either scheme/credit.

Finally, these two acts of non-compliance constitute only 3.6% of all returns or 13.9% of all non-compliant returns, assuming independence of non-compliance conditional on the covariance structure:

$$\overline{p}_{\{\text{SR and/or EITC}\}} = \mu_{\text{SR}} + \mu_{\text{EITC}} - \mu_{\text{SR}} \cdot \mu_{\text{EITC}}$$
$$0.036 = \frac{1}{1000} + \frac{4.5}{129} - \frac{1}{1000} \cdot \frac{4.5}{129}$$

This leads us to the final component of the likelihood model:

$$\text{logit}(p_{\{\text{SR and/or EITC}\}}) \sim \text{N}(\text{logit}(0.036), \sigma^2 = 0.241)$$

The maximum likelihood fit, using both the 1985 and 2004 survey years of the GSS, yields the following coefficients:[55]

			General Social Survey Year		
	Parameter	Predictor	1985	2004	Both
Slavery Reparation	α_0	Intercept	−6.76 (40.00)	−6.52 (28.65)	−6.59 (23.16)
	α_1	Male	2.78 (41.43)	2.80 (27.97)	2.77 (23.06)
	α_2	Income	−1.20 (36.78)	−0.86 (18.01)	−0.98 (16.71)
	$\overline{p}_{\text{SR}}$		0.0003	0.0005	0.0004
	n_{SR}		63	150	213
EITC	γ_0	Intercept	−3.24*** (0.86)	−3.11*** (0.71)	−3.16*** (0.55)
	γ_1	Male	0.67 (1.06)	0.79 (0.84)	0.74 (0.66)
	γ_2	Income	−0.36 (1.07)	−0.44 (0.85)	−0.41 (0.66)
	$\overline{p}_{\text{EITC}}$		0.0277	0.0248	0.0259
	n_{EITC}		610	833	1443
	$\mathcal{L}$		109	165	274
	$n_{\text{SR}} + n_{\text{EITC}}$		985	1688	2646

[55] We have used these two survey years in this paper for convenience; they have primarily informed our social network inference. For studying non-compliance, we only require robust covariance between our socio-demographic (non-network) variables, so their use here is arbitrary. Thus, it would behoove us to look at other survey years, particularly 2000.

While the coefficients are more-or-less similar across the survey years, indicating that the covariance structure is consistent, we lose much of the predictive power, as evidenced by the lack of significance, resulting in the predicted probability for 'slavery reparation', $\overline{p}_{\mathrm{SR}}$, falling short of the empirical estimate; these concerns will receive attention in the subsequent analyses, in which we estimate models for the rest of the schemes and credits. One glaring limitation is the lack of data points for 'slavery reparation'; we might look to including other GSS years to enhance the covariance structure.

References

Alm, James, Gary H. McClelland, and William D. Schulze. 1992. "Why Do People Pay Taxes?" *Journal of Public Economics* 48:21–38.

Andreoni, James, Brian Erard, and Jonathan Feinstein. 1998. "Tax Compliance." *Journal of Economic Literature* 36:818–860.

Antonides, Gerrit and Henry S.J. Robben. 1995. "True positives and false alarms in the detection of tax evasion." *Journal of Economic Psychology* 16:617–640.

Antunes, Luis, João Balsa, Paulo Urbano, Luis Moniz, and Catarina Roseta-Palma. 2006. "Tax Compliance in a Simulated Heterogeneous Multi-agent Society." In *Lecture Notes in Artificial Intelligence 3891: MABS 2005*, edited by J.S. Sichman and L. Antunes, pp. 147–161. Springer Verlag.

Bloomquist, Kim M. 2004. "Multi-Agent Based Simulation of the Deterrent Effects of Taxpayer Audits." Paper presented at the 97th Annual Conference of the National Tax Association, Minneapolis, MN November 11-13, 2004.

Bloomquist, Kim M. 2006. "A Comparison of Agent-Based Models of Income Tax Evasion." *Social Science Computer Review* 24:411–425.

Burton, Hughlene, Stewart S. Karlinsky, and Cindy Blanthorne. 2005. "Perception Of A White-Collar Crime: Tax Evasion." *American Taxation Association's Journal of Legal Tax Research* 3.

Clotfelter, Charles T. 1983. "Tax Evasion and Tax Rates: An Analysis of Individual Returns." *The Review of Economics and Statistics* 65:363–373.

Collins, Julie H., Valerie C. Milliron, and Daniel R. Toy. 1992. "Determinants of Tax Compliance: A Contingency Approach." *Journal of the American Taxation Association* 14:1–29.

Coricelli, Giorgio, Mateus Joffily, Claude Montmarquette, and Marie-Claire Villeval. 2007. "Tax Evasion: Cheating Rationally or Deciding Emotionally?" Discussion Paper Series IZA DP No. 3103, Institute for the Study of Labor (IZA), Bonn, Germany.

Cowell, F. A. 1992. "Tax Evasion and Inequity." *Journal of Economic Psychology* 13:521–543.

Devos, Ken. 2008. "Tax Evasion Behavior and Demographic Factors: An Exploratory Study in Australia." *Revenue Law Journal* 18.

Dubin, Jeffrey A., Michael J. Graetz, and Louis L. Wilde. 1990. "The Effect of Audit Rates on the Federal Individual Income Tax, 1977-1986." *National Tax Journal* 43:395–409.

Elffers, Henk, Henry S. J. Robben, and Dick J. Hessing. 1992. "On measuring tax evasion." *Journal of Economic Psychology* 13:545–567.

Elffers, Henk, Russell H. Weigel, and Dick J. Hessing. 1987. "The Consequences Of Different Strategies For Measuring Tax Evasion Behavior." *Journal of Economic Psychology* 8:311–337.

Erard, Brian. 1997. "Self-Selection With Measurement Errors: A Microeconometric To Seek Tax Assistance And Its Implications For Tax Compliance." *Journal of Econometrics* 81:319–356.

Fortin, Bernard, Bernard Fortin, and Marie-Claire Villeval. 2007. "Tax Evasion And Social Interactions." *Journal of Public Economics* 91:2089–2112.

Frey, Bruno S. and Lars P. Feld. 2002. "Deterrence and Morale in Taxation: An Empirical Analysis." CESifo Working Paper Series No. 760, Center for Economic Studies and Ifo Institute for Economic Research, Ludwig-Maximilians-Universitaet and the Ifo Institute for Economic Research, Munich, Germany.

Gelman, Andrew, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models." *The Annals of Applied Statistics* 2:1360–1383.

Gendell, Murray. 1998. "Trends In Retirement Age In Four Countries, 1965-95." *Monthly Labor Review* 121:20–30.

Goleman, Daniel. 1988. "The Tax Cheats: Selfish to the Bottom Line." *The New York Times* .

Grasmick, Harold G., Jr. Robert J. Bursik, and John K. Cochran. 1991. "Render unto Caesar What Is Caesar's: Religiosity and Taxpayers' Inclinations to Cheat." *The Sociological Quarterly* 32:251–266.

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162:1243 – 1248.

Hirschi, Travis and Michael Gottfredson. 1987. "Causes of White-Collar Crime." *Criminology* 25:949–974.

Houston, Jodie and Alfred Tran. 2000. "A Survey of Tax Evasion Using the Randomized Response Technique." In *Contemporary Issues in Taxation Research*, edited by the Tax Research Network (David Salter), chapter 4, pp. 45–68. U.K.: Ashgate Publishing.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis Of Decision Under Risk." *Econometrica* 47:263–291.

Klepper, Steven and Daniel Nagin. 1989. "Tax Compliance and Perceptions of the Risks of Detection and Criminal Prosecution." *Law and Society Review* 23:209–240.

Knight, Ray A. and Lee G. Knight. 1992. "Criminal Tax Fraud: An Analytical Review." *Missouri Law Review* 57:175–222.

Little, Craig B. 1996. "Whither White-Collar Crime." *Teaching Sociology* 24:333–337.

Mason, Robert and Lyle D. Calvin. 1978. "A Study of Admitted Income Tax Evasion." *Law and Society Review* 13:73–89.

Porcano, Thomas M. 1988. "Correlates of Tax Evasion." *Journal of Economic Psychology* 9.

Robben, Henry S.J., Paul Webley, Russell H. Weigel, Karl-Erik Wärneryd, Karyl A. Kinsey, Dick J. Hessing, Francisco Alvira Martin, Henk Elffers, Richard Wahlund, Luk Van Langenhove, Susan B. Long, and John T. Scholz. 1990. "Decision Frame and Opportunity as Determinants of Tax Cheating." *Journal of Economic Psychology* 11:341–364.

Stack, Steven and Augustine Kposowa. 2006. "The Effect of Religiosity on Tax Fraud Acceptability: A Cross-National Analysis." *Journal for the Scientific Study of Religion* 45:325–351.

Torgler, Benno. 2006. "The Importance of Faith: Tax Morale and Religiosity." *Journal of Economic Behavior* 61:81–109.

United States Census Bureau. 2003. "Public Use Microdata Sample: 2000 Census of Population and Housing."

Vogel, Joachim. 1974. "Taxation And Public Opinion In Sweden: An Interpretation Of Recent Survey Data." *National Tax Journal* 27.

Wahlund, Richard. 1992. "Tax changes and economic behavior: The case for tax evasion." *Journal of Economic Psychology* 13:657–677.

Walters, Glenn D. and Matthew D. Geyer. 2004. "Criminal Thinking and Identity in Male White-Collar Offenders." *Criminal Justice and Behavior* 31:263–281.

Webley, Paul, Michaela Cole, and Ole-Petter Eidjar. 2001. "The prediction of self-reported and hypothetical tax-evasion: Evidence from England, France, and Norway." *Journal of Economic Psychology* 22:141–155.

Weigel, Russell H., Dick J. Hessing, and Henk Elffers. 1987. "Tax Evasion Research: A Critical Appraisal And Theoretical Model." *Journal of Economic Psychology* 8:215–235.

Weigel, Russell H., Dick J. Hessing, and Henk Elffers. 1999. "Egoism: Concept, measurement and implications for deviance." *Psychology, Crime, and Law* 5:349–378.

Wentworth, Diane Keyser and Annette Urso Rickel. 1985. "Determinants of Tax Evasion and Compliance." *Behavioral Sciences and the Law* 3:455–466.