NBER WORKING PAPER SERIES

MEASURING "SCHMEDULING"

Alex Rees-Jones
Dmitry Taubinsky

Working Paper 22884
http://www.nber.org/papers/w22884

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2016, Revised March 2019

An earlier version of this paper circulated as "Heuristic Perceptions of the Income Tax: Evidence and Implications for Debiasing." We thank Doug Bernheim, Stefano DellaVigna, Emmanuel Farhi, Xavier Gabaix, Alex Gelber, Jacob Goldin, Alex Imas, Damon Jones, Erzo F.P. Luttmer, Kyle Rozema, Josh Schwartzstein, Richard Zeckhauser, Jonathan Zinman and many seminar participants for helpful comments and advice. We thank Andrew Joung, Sargent Shriver, and Linda Yao for excellent research assistance. We gratefully acknowledge research funding from the Pension Research Council/Boettner Center for Pension and Retirement Research at the Wharton School of the University of Pennsylvania, and from the Urban Institute. The opinions expressed in this paper are solely the authors', and do not necessarily reflect the views of any individual or institution listed above, nor the views of the National Bureau of Economic Research.

Measuring “Schmeduling”
Alex Rees-Jones and Dmitry Taubinsky
NBER Working Paper No. 22884
December 2016, Revised March 2019
JEL No. D03,D04,D1,H0,K34

ABSTRACT

What mental models do individuals use to approximate their tax schedule? Using incentivized forecasts of the U.S. Federal income tax schedule, we estimate the prevalence of the “schmeduling” heuristics for constructing mental representations of nonlinear incentive schemes. We find evidence of widespread reliance on the “ironing” heuristic, which linearizes the tax schedule using one's average tax rate. In our preferred specification, 43% of the population irons. We find no evidence of reliance on the “spotlighting” heuristic, which linearizes the tax schedule using one's marginal tax rate. We show that the presence of ironing rationalizes a number of empirical patterns in individuals' perceptions of tax liability across the income distribution. Furthermore, while our empirical framework accommodates a rich class of other misperceptions, we find that a simple model including only ironers and correct forecasters accurately predicts average underestimation of marginal tax rates. We replicate our finding of prevalent ironing, and a lack of other systematic misperceptions, in a controlled experiment that studies real-stakes decisions across exogenously varied tax schedules. To illustrate the policy relevance of the ironing heuristic, we show that it augments the benefits of progressive taxation in a standard model of earnings choice. We quantify these benefits in a calibrated model of the U.S. tax system.

Alex Rees-Jones
The Wharton School
University of Pennsylvania
3730 Walnut Street
Philadelphia, PA 19104
and NBER
alre@wharton.upenn.edu

Dmitry Taubinsky
University of California, Berkeley
Department of Economics
530 Evans Hall #3880
Berkeley, CA 94720-3880
and NBER
dmitry.taubinsky@berkeley.edu

1 Introduction

Financial incentives often feature nonlinearities, leading to complex decision environments. Economic models of fully optimizing and fully informed decision-makers offer a rationale for this complexity: in the presence of information asymmetries, standard results in mechanism design show that the optimal incentive scheme is often nonlinear. In practice, however, understanding these incentives appears to be difficult for many decision-makers. For example, the challenges of optimizing with nonlinear incentive schemes are starkly apparent in insurance plan choice (Bhargava et al., 2017), cell-phone usage (Grubb and Osborne, 2015), and water and energy consumption (Carter and Milon, 2005; Ito, 2014). In the context of taxation, a growing literature documents behavior inconsistent with full optimization with respect to tax credits (Miller and Mumford, 2015; Feldman et al., 2016; Chetty and Saez, 2013; Chetty et al., 2013), and surveys of taxpayers document misunderstanding of features of the income tax schedule (reviewed in section 2.2).[1]

As recently shown by Farhi and Gabaix (2018) and others,[2] the canonical mechanism-design frameworks for optimal taxation can be modified to account for these types of mistakes, with the modifications leading to new and quantitatively important implications for public policy. However, a reasonably precise understanding of the mistakes, and how they vary with policy parameters, is necessary to implement the "behavioral sufficient statistics" approach of these frameworks. From an empirical and practical standpoint, this necessarily complicates policy analysis because suboptimal responses can arise from many widely varied heuristics. However, in cases where it can be established that people rely on a small set of parsimoniously-modeled heuristics, it is possible to measure and integrate them into standard mechanism design.

In this paper, we undertake this task for two potentially focal approaches to simplifying a nonlinear incentive schedule: the models of "ironing" and "spotlighting" discussed and popularized in Liebman and Zeckhauser's "Schmeduling" (2004). These models capture individuals' tendency to approximate complex, non-linear schedules with less complex linear ones. When applying the spotlighting heuristic, the individual assumes that the slope of the linearized schedule is equal to his marginal incentive. When applying the ironing heuristic, the individual assumes that the slope of the linearized schedule is equal to his average incentive. These heuristics—and in particular ironing—capture the pervasive difficulty of "thinking on the margin."[3] Due in part to their intuitive appeal, the schmeduling heuristics are often suggested as potential mechanisms underlying misoptimized responses to tax incentives (see, e.g., Congdon et al., 2009; Finkelstein, 2009; Chetty et al., 2009; Miller and Mumford, 2015; Rees-Jones and Taubinsky, 2018). But despite the frequent

[1] Supporting the notion that understanding the tax schedule is challenging, a recent literature documents large cognitive costs from complex tax filing (see, e.g., Benzarti, 2017; Aghion et al., 2017).

[2] See, e.g., Gerritsen (2016), Blomquist and Micheletto (2006), Kanbur et al. (2008), Kanbur et al. (2006), as well as Bernheim and Taubinsky (2018) for a review.

[3] The difficulty of thinking on the margin should be apparent to anyone who has taught undergraduate microeconomics classes. Beyond such anecdotes, this difficulty has been documented in high-stakes managerial decisions about product pricing (see, e.g., Altomonte et al., 2015).

discussion of these heuristics in the behavioral public economics literature, little existing evidence informs their importance as an account of tax misperceptions.

We report on two complementary experimental studies that we designed to identify individuals' propensity to rely on the schmeduling heuristics. Study 1 relies on detailed survey elicitation of individuals' perceptions of the U.S. Federal income tax schedule. The key innovation of the approach in this study arises from the observation that direct elicitation of individuals' perceptions of broader regions of the tax schedule is extremely useful for separately identifying the schmeduling heuristics from closely related candidate forms of misunderstanding. Such elicitation makes it possible to study how variation in respondents' average and marginal tax rates relates to their tax forecasts for a given level of income. In Study 2, individuals make incentivized choices between receiving additional taxable or non-taxable income, and face randomly assigned tax schedules. This design allows us to infer tax perceptions from choice behavior, and to study how tax perceptions respond to exogenous variation in average and marginal tax rates.

In section 2 we derive a series of distinguishing predictions for schmedulers' beliefs about different regions of the tax schedule. These predictions describe the biases that arise when forecasting one's own marginal tax rate, the steepness of the tax schedule over a broader region of incomes, the tax liabilities of the relatively rich and the relatively poor and, importantly, how all of these forecasts evolve with the respondent's own income. For example, the ironing heuristic leads individuals to overestimate the taxes paid by low income earners and underestimate the taxes paid by high income earners, though the overall perception of the tax burden on both the poor and the rich is increasing in individuals' incomes. Conversely, the spotlighting heuristic leads to average underestimation of all tax burdens, with perceptions of the taxes paid by the poor decreasing in respondents' incomes. We show that measurements of these kinds of perceptions allow the relative propensity of each schmeduling heuristic to be separately identified from a rich class of other candidate misunderstandings.

In section 3, we describe the design and results of Study 1. We deployed an incentivized tax-forecasting task to an approximately representative sample of 4,197 U.S. taxpayers. Respondents forecast the tax due by an example taxpayer who faces an income different from their own, but who is otherwise constructed to be very similar to themselves on tax-relevant dimensions. Forecasts are repeatedly elicited for different potential income amounts for the example taxpayer, facilitating inference on the structure of the schedule that the respondents believe is in place.

Using these data, we examine the basic structure of perceptions of the U.S. income tax schedule. Consistent with prior work, we find that taxpayers underestimate the marginal tax rates that apply in their own tax brackets. Examining perceptions beyond the respondent's own bracket, we find a systematic tendency to overestimate the taxes paid by the comparatively poor, to a degree that becomes more severe as respondents' own income increases. Conversely, we find a systematic tendency to underestimate the taxes paid by the comparatively rich, to a degree that becomes less severe as respondents' own income increases. Taken together, these results are largely consistent with the predictions of ironing, are inconsistent with the predictions of spotlighting, and are inconsistent with models in which misperceptions are invariant to respondents' incomes.

Motivated by these reduced-form results, we estimate a structural model of tax perceptions that embeds populations of ironers, spotlighters, correct forecasters, and a rich class of other candidate mistakes. We find that tax perceptions are best explained by a model with approximately 43% of filers adopting the ironing heuristic and no filers adopting the spotlighting heuristic. The remaining filers appear to overestimate taxes on most of the population by a relatively constant amount, and underestimate taxes applicable to the top 5 percent of the population. On average, these remaining filers have perceptions of marginal tax rates not far off from the truth. These results are qualitatively unchanged in robustness checks that consider various sample restrictions or possible misunderstandings of the survey prompt.

Despite the many different ways in which taxpayers could misperceive taxes, we find that a model accounting for only our estimated rate of ironing is able to precisely explain the systematic underestimation of marginal tax rates in one's own tax bracket. This result holds not only when ironing propensity is estimated using our full sample of elicited perceptions, but also in an out-of-sample fit assessment that estimates schmeduling propensity while excluding forecasts from the respondent's own bracket. This validates our use of individuals' perceptions of broader regions of the tax schedule to help estimate the heuristic processes that govern more local perceptions. Furthermore, this suggests that a parsimonious two-type model of ironers and correct forecasters provides a satisfactory account of systematic marginal tax rate misperceptions.

Of course, predicting field behaviors based on our survey study requires caution (see Bernheim and Taubinsky, 2018 for a critical discussion of the use of belief elicitations in behavioral welfare analyses). Individuals may not necessarily act on the beliefs that they state, or they may be more prone to think about marginal tax rates when considering an incremental increase in taxable income.[4] Furthermore, misperceptions could be endogenous to the tax schedule; for example, individuals might be less likely to rely on the ironing heuristic if the tax schedule were simplified.

To explore these issues, and to examine the robustness of our results beyond our particular elicitation context, we conducted an incentivized online experiment that reveals the rate of ironing through individuals' choices. In Study 2, we presented 3,130 participants with a choice between an incremental contribution to a taxable account or an incremental contribution to an untaxed account. Participants were randomized into one of 54 possible tax schedules, which generated exogenous variation in both marginal and average tax rates. The tax schedules also varied in complexity: half of the participants were assigned to tax schedules with two tax brackets, while half were assigned to tax schedules with five tax brackets. Our design closely relates to work by de Bartolome (1995), which uses a single tax schedule and shows that individuals underreact to its marginal tax rates.

[4] A survey study eliciting marginal tax rates could potentially mitigate this concern, but would face other severe challenges to identification. See the discussion in Section 2.3.

By contrast, we exploit exogenous variation in both average and marginal tax rates to quantify how much of that underreaction may be attributed to ironing.

We report the results of Study 2 in section 4. Consistent with our findings from Study 1, we find substantial reliance on the ironing heuristic: our primary estimates show that 9% of respondents correctly use the marginal tax rate, 35% of respondents incorrectly use their average tax rate as if it were their marginal tax rate, and the remainder are unresponsive to variation in either the average or marginal tax rate. Beyond the participants who are unresponsive (consistent with prior findings of inattention to taxes, such as Chetty et al., 2009; Taubinsky and Rees-Jones, 2018), we find no evidence of other systematic misperceptions. We find that a simple three-type model in which individuals either ignore variation in tax schedules, iron, or react correctly explains over 87.5% of the variation in tax perceptions arising from variation in schedules. Furthermore, we find no evidence that the propensity to rely on the ironing heuristic is endogenous to the complexity of the tax schedule.

In section 5 we illustrate the implications of our findings for tax policy. We embed a population of ironers in an otherwise standard model of income taxation building on, e.g., Mirrlees (1971) and Saez (2001). In this model, ironing increases the welfare attained under a convex tax schedule, since reliance on this heuristic generates revenue in a progressive fashion. For the U.S. income tax system, our preferred estimates imply that the presence of ironing increases welfare by an amount equivalent to a 2.3% windfall in available government revenue—a quantitatively large effect both in absolute and relative terms.[5] Moving to analysis of policy reform, we find that ironing increases the welfare costs of moving toward a flat tax. Because ironers do not underestimate marginal tax rates in a flat tax system, moving to this system generates the additional cost of eliminating the welfare-enhancing mistakes created by ironing. Our preferred estimates suggest that ironing increases the welfare costs of moving from the U.S. tax schedule to a revenue-equivalent flat tax by 14%. Moving to analyses in the tradition of Saez (2001) (and their behavioral extensions in Farhi and Gabaix, 2018), we find that our estimated prevalence of ironing substantially increases the optimal top marginal tax rate.

In section 6, we conclude by highlighting what we view as the primary considerations of external validity when evaluating the implications of our results for policy analysis. We highlight the elements of our analysis that we view as most, and least, susceptible to external validity concerns, summarize the generalizable lessons for tax policy, and highlight productive paths for further research.

While economists traditionally favor revealed-preference analysis, recent works have argued that direct belief elicitation serves a key function in economics: reported beliefs can often discriminate between models that are indistinguishable from observed behavior alone (Manski, 2004; Gennaioli et al., 2016). This logic is particularly relevant when studying misperceptions of income tax systems. Existing quasi-experimental studies (e.g., Feldman et al., 2016) are qualitatively consistent with

[5]To illustrate, 2.3% of individual income tax return revenue in our year of study amounts to 32 billion dollars—approximately half of the cost of programs such as the Earned Income Tax Credit ($68 billion in 2014) or the child tax credit ($57 billion in 2014); for details and other useful benchmarks, see Congressional Budget Office (2013).

the presence of ironing, but cannot identify ironing and spotlighting propensities, or establish the degree to which these heuristics offer a complete characterization of systematic misperceptions.

We contribute to two distinct lines of research applying survey methods to questions in public economics. First, our paper contributes to the existing literature on direct elicitation of tax perceptions. Surveys of tax knowledge (e.g., Lewis, 1978; Fujii and Hawley, 1988) suggest that a key prediction of ironing—underestimation of marginal tax rates—is common. We directly improve on this literature by providing better powered estimates of marginal tax rate underestimation using improved elicitation methods. But more importantly, we leverage elicited perceptions of taxes on incomes outside of one's bracket to identify the mechanisms driving tax rate misunderstanding. Tests of tax understanding at different regions of the tax schedule, as in Blaufus et al. (2015) and Gideon (2017), move in the direction of our empirical design, and suggest misperceptions of taxes on the comparatively rich and poor that could arise from ironing. Yet as we show in Section 2, none of the existing survey evidence is able to firmly distinguish ironing from other candidate models (and indeed, this was not the goal of this previous work).

Additionally, we relate to a recent literature that elicits individual beliefs about, and preferences over, inequality (see, e.g., Cruces et al., 2013; Weinzierl, 2014; Kuziemko et al., 2015; Weinzierl, 2017; Alesina et al., 2018). This literature studies individuals' preferences for redistribution and their determinants. In contrast, we study individuals' understanding of the economic incentives generated by the tax systems in place—a necessary component for the integration of misperceptions into optimal tax policy analysis.[6] To the extent that beliefs about the taxes paid by others are of direct interest (as they pertain to questions of policy support and political economy), our study provides an unusually comprehensive study of the beliefs held by U.S. taxpayers.

2 "Schmeduling" and its Predictions

2.1 Definitions

Liebman and Zeckhauser (2004) propose two compelling heuristics for decision-making with a non-linear incentive schedule like the income tax. In summarizing these heuristics, we denote perceived income tax schedules by $\tilde{T}(z|z^*)$, and the true income tax schedules by $T(z)$. These report the tax due as a function of income (z), given the individual's chosen income (z^*).

Definition 1. The ironing heuristic arises when an individual uses the average price at the point where he consumes to forecast prices at other consumption levels. In the tax context, this corresponds to an individual knowing his average tax rate and applying that rate to any amount of income. Formally, $\tilde{T}_I(z|z^*) = \frac{T(z^*)}{z^*} \cdot z$.

[6] Kuziemko et al. (2015) have one question pertaining to perceptions of tax rates: whether the top income tax rates today are higher or lower than what they were in the 1950s and 1960s.

As illustrated in Figure 1, use of the ironing heuristic corresponds to approximating the non-linear tax schedule with a secant line drawn through one's own position on the schedule and the origin. As noted in Liebman and Zeckhauser, the reliance on this heuristic can be rationalized given the manner in which tax information is often conveyed to taxpayers. Paystubs typically present one's gross earnings and the subtracted tax withholdings, making the average tax salient and leaving the marginal tax rate unknown. Furthermore, at the time of completion of annual tax returns, filers must calculate both their total annual taxable income as well as the total annual tax due, again inviting assessment of the average tax rate. Taxpayers who find these numbers more accessible could conceivably adopt this forecasting rule as a time-saving heuristic, allowing them to approximate their own position on the schedule while accessing only a single, salient number. Despite its convenience, reliance on this heuristic would result in misoptimized labor supply choices on the intensive margin.

Definition 2. The spotlighting heuristic arises when an individual uses the local slope of his price schedule to forecast prices at non-local levels. In the tax context, this corresponds to an individual acting as if his own tax bracket extends to other regions of the tax schedule. Formally, $\tilde{T}_S(z|z^*) = T(z^*) + T'(z^*) \cdot (z - z^*)$.

As illustrated in Figure 1, use of the spotlighting heuristic corresponds to approximating the non-linear tax schedule with a tangent line drawn through one's own position on the schedule. Use of this heuristic can be rationalized by noticing that some individuals might take the time to learn their own local region of the tax schedule—for example, by noting their past-years' tax burden and by looking up their statutory marginal tax rate—but might mistakenly forecast these local parameters beyond the narrow region to which they apply. Compared to the ironing heuristic, the usage of the spotlighting heuristic is somewhat more cognitively demanding, as it is requires knowledge of two idiosyncratic parameters (one's tax burden and marginal tax rate) rather than just one (the average tax rate, in the case of ironing). Although this heuristic leads to (approximately) optimized labor supply choices on the intensive margin, it would lead to incorrect labor supply choices on the extensive margin.[7]

Additionally, if taken literally, the use of this heuristic can result in the extreme prediction that tax burdens on very low incomes are negative, whereas the ironing heuristic results in weakly positive tax forecasts. An alternative model of this heuristic would replace negative tax forecasts with forecasts of zero tax. As we will discuss in section 3.3.2, using this alternative definition has minimal impact on our results.

Several possible psychological frameworks could underlie these heuristics. One is the psychological literature that likens human judgments to those of a "naive intuitive statistician" (Fiedler and Juslin, 2006). This literature argues that decision makers are often able to form reasonably

[7] See, e.g., Saez (2002b) for a theory and quantification of the importance of extensive margin elasticities.

accurate forecasts of simple sample properties, such as frequencies or averages (for early examples, see Spencer, 1961, 1963). However, they often fail to account for sampling biases or constraints in such judgments. As summarized in Juslin et al. (2007), "people tend spontaneously to assume that the samples they encounter are representative of the relevant populations." Both schmeduling heuristics may be considered specific examples of this general decision error. The decision-maker correctly assess the average tax rate over either all dollars earned that year (in the case of ironing) or over small changes to his own income (in the case of spotlighting), but then incorrectly applies this average as the range is changed in a manner rendering their previously considered sample non-representative.[8]

Another possible foundation for these heuristics is the cognitive economy of linear approximations. Linearization is a common computational technique in economics and other disciplines; it is possible that non-academics uses similar techniques more heuristically and less deliberately in their day-to-day decisions. Other possible manifestations of this psychology include exponential growth bias (Stango and Zinman, 2009; Levvy and Tasoff, 2016). The bounded rationality model of Gabaix (2014) features some of this psychology by assuming the individuals linearize the marginal benefits of information acquisition or attention allocation; extensions of the model could potentially be used to formalize linearization as a cognitive shortcut more generally.

In general, we note that ironing and spotlighting are unlikely to be "primitive biases." Consequently, until the details of the underlying cognitive processes are fully understood, an analyst must consider that reliance on these heuristic may be endogenous to features of the tax system. We will directly test for this possibility in Study 2.

2.2 Predictions of Schmeduling

We now formalize a series of predictions aimed at illustrating the features of tax perceptions that do, or do not, distinguish between these heuristics. In principle, the use of these heuristics can be examined at the individual level (related analyses will be presented in section 3.3.4). However, in practice, common imperfections in survey responses—such as measurement error or rounding heuristics—can substantially confound individual-level analyses. These issues lead us to formulate our predictions with respect to the average tax schedule perceived by a population of heuristic forecasters.

We first analyze predictions about the levels of misperceptions, and then analyze predictions about the slope of the misperceptions. We document these predictions for the empirically-relevant case of progressive—i.e., convex—tax schedules. We also note that in addition to the predictions spelled out here, both models make the assumption that the forecasted tax schedule is linear and intersects the true one at the forecaster's own income. We examine these assumptions in Section

[8] A potentially related psychology is narrow-bracketing, which is the tendency of people to treat different dimensions of a decision separably, without considering how they interact with each other.

3.3.4.

Prediction 1. Perceptions of taxes on low- and high-income filers.

1-I: Ironers overestimate the taxes paid by low-income filers and underestimate the taxes paid by high-income filers.

1-S: Spotlighters underestimate the taxes paid by both low- and high-income filers.

The reasoning behind Prediction 1 is apparent in Figure 1. At the individual-level, ironers overestimate the taxes due for individuals with lower earnings than their own and underestimate the taxes due for individuals with higher earnings than their own. When averaging the perceived schedules of a population of ironers, this results in Prediction 1-I, with low- and high-income evaluated relative to an appropriately weighted average of the incomes of the tax forecasters. In contrast, the forecast corresponding to the spotlighting prediction is always below the true taxes due at the individual level, resulting in Prediction 1-S.

Prediction 2. Perceptions of taxes on low- and high-income filers, by own income.

2-I: Higher-income ironers exhibit more overestimation of the taxes paid by low-income filers and less underestimation of taxes paid by high-income filers.

2-S: Higher-income spotlighters exhibit more underestimation of the taxes paid by low-income filers and less underestimation of taxes paid by high-income filers.

The reasoning behind Prediction 2 is again apparent in Figure 1. As the ironer's income is increased, the slope of the secant line increases, directly leading to Prediction 2-I. As the spotlighter's income is increased, the tangent line rotates upward, leading to Prediction 2-S.

We now consider how the schmeduling heuristics influence perceptions of the steepness of the tax schedule, both locally and more globally.

Prediction 3. Perceptions of marginal tax rates (MTRs).

3-I: Ironers underestimate their own MTR.

3-S: Spotlighters correctly estimate their own MTR.

Claim 3-S follows immediately from the definition of spotlighting: by assumption, spotlighters know and apply their MTR. Turning to Prediction 3-I, we note that the definition of ironing does not fundamentally require underestimation of MTRs, but that this underestimation arises for progressive tax schedules, in which average tax rates (ATRs) are always lower than MTRs.

As a means of characterizing perceptions of the "steepness" of a broader region of the tax schedule, it is convenient to define the *perceived slope* over income range $[z_1, z_2]$ to be $\frac{\tilde{T}(z_2|z^*)-\tilde{T}(z_1|z^*)}{z_2-z_1}$, and the *actual slope* to be $\frac{T(z_2)-T(z_1)}{z_2-z_1}$.

Prediction 4. Perceived slope of tax schedule.

4-I: Consider income range $Z = [\underline{z}, \bar{z}]$ which satisfies $z_i^* \in Z$ for all individuals i. Evaluated over Z, ironers' perceived slope underestimates the actual slope.

4-S: Consider income range $Z = [\underline{z}, \bar{z}]$ which satisfies $z_i^* \in Z$ for all individuals i. There exists a threshold $z^\dagger \in Z$ such that spotlighters earning $z_i^* \leq z^\dagger$ underestimate the actual slope over Z, whereas spotlighters earning $z_i^* > z^\dagger$ overestimate the actual slope.

As illustrated in Figure 1, the use of the ironing heuristic results in a linear approximation to the schedule that ultimately is shallower than the schedule itself. Since the schedule is convex, treating one's average tax rate as the relevant slope for higher income values will result in an approximate schedule that underestimates the highest taxes due, and concurrently is "flatter" than the true schedule. In contrast, spotlighters' perceptions of the average slope of the full tax schedule depend more critically on their own income. For lower-income individuals, one's own marginal tax rate can be lower than the average slope of the entire schedule. However, for sufficiently high-income individuals, this effect reverses, leading spotlighters to overestimate the average slope. This contrast is displayed in the two examples of Figure 1.

Moreover, because both marginal and average tax rates are rising with one's own income in a progressive tax system, the perceived slope over an income range will be increasing in the schmeduler's income:

Prediction 5. Perceived slope of tax schedule, by own income.

5-I: Ironers' perceived slope over income range Z is increasing in their earnings, z^*.

5-S: Spotlighters' perceived slope over income range Z is increasing in their earnings, z^*.

While Prediction 5 does not distinguish between the use of the two heuristics, it does clearly differentiate their use from correct forecasting, or from forecasting based on misperceptions of the schedule that are not tied to one's earnings.

Summary of Existing Survey Evidence: Perceptions of the income tax have long been of interest in public finance, with a literature on the survey measurement of these perceptions extending back at least to the 1960s.[9] While this literature has asked questions in some ways related to our own, the different objectives of these papers lead them to collect data of limited use in identifying the schmeduling heuristics. For example, early papers in this literature (Enrick, 1963, 1964; Wagstaff, 1965) assess if taxpayers know the size of their own tax bill—a feature of tax perceptions for which ironing, spotlighting, and correct forecasting all make the same prediction.[10]

[9]Appendix Table A1 provides a summary the most relevant survey literature and indicates the schmeduling predictions that this these papers inform.

[10]For recent work further examining perceptions of average tax rates, see Ballard and Gupta (2018).

Later papers became more interested in assessing taxpayers' understanding of their marginal tax rate, which permits an assessment of Prediction 3. The early works of Gensemer et al. (1965) and Brown (1969) document substantial misperceptions of marginal tax rates, with Brown documenting a tendency towards marginal tax rate overestimation. Subsequent studies with substantially larger samples have found average underestimation (Lewis, 1978; Fujii and Hawley, 1988), consistent with Prediction 3-I and rejecting Prediction 3-S.[11] However, viewed in isolation, such findings do not clearly establish ironing as the mechanism driving this underestimation.

Fewer results are available to assist in assessing Predictions 1, 2, 4, or 5. Blaufus et al. (2015) ask a panel of German respondents about their perceptions of the taxes paid by those earning 10,000, 40,000, 300,000, or 2 million euros a year. They find evidence that the tax burden is overestimated at the bottom of the scale and underestimated at the top of the scale, consistent with Prediction I-1. Gideon (2017) finds that survey respondents tend to underestimate the top marginal tax rate, a feature that could lead to underestimation of taxes paid by the rich (as in 1-I) and could arise from an underestimation of slope of the full tax schedule (as in 4-I). Beyond results such as these, this literature's focus on perceptions of taxes at a smaller number of very specific, typically local points results in a general inability to test the nuanced predictions that firmly separate the schmeduling heuristics from other potential mechanisms. This motivates the design of our experiment, eliciting the key class of data absent from this work: perceptions of taxes over a broader variety of incomes, sampled with continuous support.

2.3 Identifying Schmeduling Propensities

Formally, we characterize the average perceived tax schedule as a mixture model of forecasting types, given by

$$E[\tilde{T}(z|z^*)|z, z^*] = \gamma_I \tilde{T}_I(z|z^*) + \gamma_S \tilde{T}_S(z|z^*) + \sum_k \omega_k \tilde{T}_k(z) + (1 - \gamma_I - \gamma_s - \sum_k \omega_k)T(z). \tag{1}$$

In this model, γ_I denotes the fraction of individuals using the ironing heuristic and γ_S denotes the fraction of individuals using the spotlighting heuristic. Non-schmeduling taxpayers may hold alternative perceptions of the schedule (captured by the term $\sum_k \gamma_k \tilde{T}_k(z)$), with the remainder of taxpayers adopting the correct tax schedule.

We aim to make minimal assumptions about the structure of the alternative misperceptions beyond explicitly specifying that they are not a function of earnings choice, and thus are not a function of either the individual's average or marginal tax rate. Misperceptions of this sort could correspond to, e.g., average overestimation of the tax burden or a general tendency to underestimate

[11] In more nuanced recent work, Gideon (2017) examines the accuracy of marginal tax rate perceptions across the income distribution. He finds average underestimation of marginal tax rates for respondents with gross income exceeding \$50,000, and average overestimation for those of lower income.

marginal tax rates. For example, $\tilde{T}_k$ could correspond to underestimation of all MTRs by 50%; that is, $\tilde{T}_k = t_0 + \frac{1}{2}T$, for some constant $t_0 \in \mathbb{R}$. Note that although we think it most psychologically natural to consider γ_I and γ_S as probabilities of pure ironing or spotlighting types, we could instead interpret these parameters as weights in a representative agent model of partial ironing and spotlighting.

To obtain intuition for the data requirements for identifying equation (1), consider first a dataset in which only perceptions of respondents' own MTR under a fixed schedule are measured (i.e., $\frac{d}{dz}E[\tilde{T}(z|z^*)|z,z^*]|_{z=z^*}$). Such observations are plainly insufficient to separately identify γ_I, γ_S, and $\sum_k \omega_k \tilde{T}_k(z)$. To illustrate, note that correct average perceptions of MTRs could arise from correct tax forecasting, or from spotlighting, or from some individuals underestimating their own MTRs due to ironing while others over-estimate MTRs in a perfectly offsetting manner. As another example, consider a group of individuals facing an average tax rate of 10% and an MTR of 20%, but perceiving their MTR to be 15%. Such perceptions could be rationalized by a 50-50 mixture of correct forecasters and ironers. However, they could also be rationalized by a 50-50 mixture of correct forecasters and individuals who simply underestimate *all* MTRs by half; or by a 3:1 mixture of correct forecasters and individuals who underestimate *all* MTRs by 75%.

The "ideal" variation for identifying γ_I and γ_S is exogenous variation in the marginal and average tax rates at z^*. This would generate exogenous variation in $\tilde{T}_I(z|z^*\ )$ and $\tilde{T}(z|z^*)$, which would lead to clear identification of γ_I and γ_S. We employ this kind of exogenous variation in Study 2.

In Study 1, we work with the fixed U.S. income tax schedule, and thus cannot utilize this ideal variation. Instead, we utilize variation in marginal and average tax rates that is obtained from respondents' own earned income z^*. Under the assumption that residual misperceptions ($\sum_k \omega_k \tilde{T}_k(z)$) do not covary with z^*, the variation in z^* provides the needed variation in average and marginal tax rates.

Formally, we elicit beliefs about $\tilde{T}(z|z^*)$ across a joint distribution of (z, z^*) spanning the relevant range of each variable's support. With such data, the analyst has access to the empirical moments $\frac{d}{dz^*}E[\tilde{T}(z|z^*)|z,z^*]$. Both models of schmeduling offer different, full accounts of the structure of this derivative. For example, predictions 1 and 2 show that the models predict different *signs* for this derivative for $z < z^*$. Moreover, since either correct forecasting or our position-independent alternatives require that this derivative be zero at all points, this is sufficient to identify both γ_I and γ_S. Conditional on such estimates, the average of all other perceived schedules is identified by the "residual" structure of $E[\tilde{T}(z|z^*)|z,z^*]$ unexplained by the estimated schmeduling propensity. This illustrates the identifying power of tax perceptions elicited over a broader support of (z, z^*), and motivates our design of an experiment capturing such perceptions.

Focusing on forecasts of absolute tax paid, rather than on forecasts of marginal tax rates, is an important feature of our empirical strategy. Note that a dataset that elicits perceptions of marginal tax rates at different points on the tax schedule would primarily be identified through the heuristics' differing predictions about $\frac{d^2}{dzdz^*}E[\tilde{T}(z|z^*)|z, z^*]$. As shown in prediction 5, both models predict that perceptions of MTRs increase with z^*, resulting in greater colinearity of predictors since ATR is an increasing function of the MTR in progressive tax schedules. Such an empirical framework would be less well-powered to differentiate between the two heuristics of interest, and would also be less robust to assumptions about functional form.[12] These concerns guided our decision to collect a different class of data than has been pursued in many prior works.

3 Study 1: Perceptions of the U.S. Federal Income Tax

3.1 Experimental Design

We administered Study 1 during the tax season of 2015. From March 15th through May 17th, respondents were recruited for a brief[13] web survey hosted on the Qualtrics platform, with recruitment targeting similar sample sizes in all weeks of this sampling window. Subject recruitment was managed by ClearVoice Research, a market research company that maintains a large, national population of respondents willing to take brief online surveys.[14] Respondents were recruited based on demographic data previously provided to ClearVoice, allowing us to generate a sample with demographics that approximate the national age, income, and gender distribution found in the U.S. census records (for tabulations of demographics in our sample and the census, see Appendix Table A2).

3.1.1 Experimental Protocol

The Qualtrics survey featured four modules. Screenshots of the full experiment are available in the Survey Appendix; we summarize the contents here.

Introductory Module: The first module elicited basic information about respondents' tax filing behavior, allowing us to construct a similar hypothetical tax filer in the forecasting module. Respondents were asked if they had already filed their tax return; who completed (or would complete) that tax return; their filing status; their exemptions claimed; if they claimed the standard or itemized deduction; their total income; if they filed each of schedule B through F; if they used TurboTax or similar software; if they or their spouse were born before January 2, 1950; and if they claimed the Earned Income Tax Credit.[15] Additionally, respondents were asked their degree of confidence in the key parameters determining their tax: their filing status, their exemptions, their

[12]For example, consider nonlinear transformations of the schmeduling predictions, as in the alternative definition of spotlighting that we discuss in Section 2.1.

[13]Median completion time: 16 minutes. Interquartile range: 11-25 minutes.

[14]For other economic research making use of the ClearVoice panel, see Benjamin et al. (2014) or Taubinsky and Rees-Jones (2018).

[15]Respondents who claimed the Earned Income Tax Credit completed an additional brief battery of questions regarding their understanding of this tax provision.

deduction status, and their income. Confidence in these parameters was high. Given ratings options of "very confident," "somewhat confident," and "not confident at all," 96% of respondents were "very confident" in their filing status; 89% of respondents were "very confident" in their number of exemptions; 90% of respondents were "very confident" in their deduction status; 71% of respondents were "very confident" that their total income reported was within $1,000 of being correct.

Forecasting Module: The key questions for our empirical analysis were contained in the forecasting module. Respondents were presented with a variant of the following prompt, describing a hypothetical taxpayer whose filing behavior was very similar to their own:[16]

> This next group of questions is about Fred, a hypothetical taxpayer who is very similar to you. Fred is your age, and has a lifestyle similar to yours. Fred filed his 2014 Federal Tax Return claiming [own exemptions] exemption(s) and [own status] filing status, like you did. Fred also claimed the standard deduction, like you did.[17] However, Fred's tax computation is particularly simple, since all of his taxable income comes from his annual salary. He has no other sources of taxable income, and is not claiming additional credits or deductions.
>
> For the following questions, we will ask you to estimate the total federal income tax Fred would have to pay for different levels of total income. To help motivate careful thought about these questions, we are providing a monetary reward for correct answers. At the end of the survey, one of these questions will be chosen at random. If your answer to that question is within $100 of the correct answer, $1 will be added to your survey compensation.

Following this preamble, respondents made 16 forecasts of taxes due under different amounts of income, given the following prompt:

> If Fred's total income for the year were $[X], the total federal income tax that he has to pay would be:

The amounts of income substituted into the prompt above were drawn according to three sampling schemes. Ten forecasts were drawn from the "mid-range sampling distribution." This is a range of income values spanning all but the top of the national income distribution, sampling uniformly from $0 up to a point partially through the fourth tax bracket. This sampling pattern differs by filing status, leading us to present estimates separately by filing status in some of our analysis. Four forecasts were drawn uniformly from the "high-income sampling distribution," starting from the top of the mid-range income distribution and ranging to approximately $500,000. We call the sample

[16] For respondents who had not yet completed their tax return, the verb tense was changed from past to future as appropriate.

[17] For filers not claiming the standard deduction, this sentence read: "Unlike you, Fred claimed the standard deduction."

of all 14 points the "full sampling distribution." Finally, two draws were included that guarantee the presence of some forecasts "close" to the respondent's own income. One draw substituted the respondent's own reported income for X above, while the second applied that income plus a random perturbation taking a value between 0 and 1,000.[18] When assessing respondents' knowledge of the tax schedule local to their own income, we will restrict data to the "local distribution" consisting of these two forecasts as well as any of the random forecasts that happen to fall in the respondent's own tax bracket. However, when we are not assessing questions about local tax perceptions, we exclude these two forecasts to preserve a random sampling structure.

Miscellaneous Questions: After the forecasting task, respondents faced a brief battery of miscellaneous questions. These included an elicitation of the salience of their income tax, assessments of their health and savings behaviors, an elicitation of their elasticity of charitable giving, the "big three" financial literacy questions of Lusardi and Mitchell (2014), and a test of knowledge of their sales tax rate. These also included an attention check, in which the text of the question instructed the respondent not to selected any of the multiple-choice options below the text. This assists in screening out respondents providing random answers without reading instructions.[19]

Incentives: On the final screen, one of the respondent's 16 tax forecasts was randomly selected for incentivization. They were told the correct answer, reminded of their own answer, and awarded the bonus payment if their response was within $100 of the truth.

3.2 Sample for Analysis and Dataset Preparation

Over the course of our sampling period, 4,828 individuals completed the full survey. We exclude respondents according to several criteria. First, we exclude 5 respondents with missing data on one or more of the tax forecasts. Second, we exclude 73 respondents forecasting either 0 tax or 100% tax for all forecasts, as we believe this reporting pattern indicates either misunderstanding of the prompt or represents an attempt to quickly click through the survey without meaningfully responding to questions. Third, we restrict our sample torespondents reporting income ranging from zero to $250,000, excluding 117 respondents whose self-reported incomes are outside the typical range of the panel. Finally, we exclude 436 respondents who failed the attention check included in the miscellaneous questions module. To limit the influence of extreme tax forecasts, we conduct a rolling Winsorization of tax forecasts to the 1st and 99th percentile values in each $10,000 income bin.

This set of restrictions results in a final sample of 4,197 respondents, and a total of 58,758

[18] This value fell within the respondent's own tax bracket for all but 89 respondents.

[19] The text of the attention check was as follows: "Sometimes participants who take online surveys don't read all of the instructions and click through the questions more quickly than they should. We want to make sure that participants taking this survey are paying attention to the instructions. On this screen, your instructions are simply to click on the continue button at the bottom of the screen, and not to fill in any of the answer below. Please click continue now." Below this prompt, multiple choice options ranging from "Always" to "Never" were present. These were meant to appear like standard likert-scale responses to a respondent who had not read instructions.

forecasts of tax liability for randomly drawn incomes. In section 3.3.4, we discuss the robustness of our empirical results to these dataset construction decisions.

3.3 An Empirical Assessment of Tax Misperceptions

3.3.1 Reduced-Form Tests of Schmeduling

Systematic Over- and Under-Estimation (Schmeduling Predictions 1 and 2): To present an initial, non-parametric summary of income tax perceptions, Figure 2 plots a kernel-smoothed estimate of average perceived tax schedules. Since schedules are filing-status specific, we present estimates separately for the two largest filing-status groups: single and married filing jointly. The top panels present estimates restricted to the data from the ten income draws of the mid-range sampling distribution. The lower plots extend the support to include the four income draws from the high-income sampling distribution.

The top two panels illustrate two important features of perceived tax schedules. First, on average, the perceived tax schedule is qualitatively similar to the true tax schedule, though it displays some systematic error. For example, over the mid-range sampling range, respondents overestimate the tax burden by \$679 (clustered s.e.: \$185) on average, or 3.2 percentage points (clustered s.e.: 0.003pp) in effective tax rates. Standard errors are clustered at the respondent level.

Second, and perhaps more importantly, these plots also demonstrate that the sign of the average misperception depends on the amount of income that is being taxed. In both plots, the average perceived tax schedule appears more linear than the true schedule, with a tendency towards overestimation of the tax burden for low amounts of income and underestimation of the tax burden for high amounts of income. In the lower two plots of Figure 2, which present analyses including the high-income sampling distribution, this underestimation of taxes on high incomes becomes even more pronounced. This pattern indicates a general underappreciation of the degree of progressivity in the current U.S. tax code. The qualitative patterns of these results provide a test of Prediction 1, formally rejecting the predictions made by the spotlighting model (1-S) while remaining consistent with the predictions made by the ironing model (1-I).

Turning to Prediction 2, we next explore the differences in perceived schedules as a function of respondents' own income. Figure 3 summarizes the average bias in forecasts as a function of the true tax due, and plots this bias conditional on respondent's own income quartile. We present fitted values from the regression model:

$$(\tilde{T} - T)_{i,f} = \sum_{b,q}\sum \alpha_{b,q} * I(\text{income}_f \in \text{bin}_b) * I(\text{income}_i \in \text{quartile}_q) + \epsilon_{i,f}.$$

In this regression, we predict the difference between the perceived tax ($\tilde{T}$) and the true tax (T) for person i's assessment of Fred scenario f. We estimate the average of this forecast error in \$5,000 bins (denoted b), estimated separately by the income quartile of the respondent (denoted q).[20] As seen in this figure, the primary pattern described above—overestimation of low tax burdens and underestimation of high tax burdens—persists across all four income groups. Despite this consistent pattern, tax perceptions are significantly different across income quartiles. Wald tests reject joint equality in all pairwise comparisons of the income-quartile-specific estimates of α (all p-values $<$0.014). An important feature revealed in the plots is that the "crossing point" where overestimation turns to underestimation occurs at higher income values for higher income respondents. These patterns are consistent with the predictions of ironing (2-I), and are broadly inconsistent with the predictions of spotlighting (2-S).

Perceived Slope of Tax Schedules (Schmeduling Predictions 3-5): Figures 2 and 3 suggest underestimation of the slope of the tax schedule. This is visually apparent in the "flattening" of the estimated schedules in Figure 2, and in the negative slope of the bias functions in Figure 3. We now turn to a statistical assessment of this underestimation, and use our results to assess Predictions 3-5.

To formally test for underestimation of the slope of the tax schedule, we estimate fixed-effect OLS regression models of the form $\tilde{T}_{i,f} = \beta T_{i,f} + \nu_i + \epsilon_{i,f}$. The object of interest in this analysis is β, which measures the scaling of the tax schedule implicit in the respondents' forecasts. By including respondent-specific fixed effects (ν_i) we identify β from the effective slope of the tax schedule reported within-subject. We test the null hypothesis of $\beta = 1$, the value that would be estimated if respondents indicated a rate-of-increase of taxes consistent with the true tax schedule. An estimated value over 1 would indicate an implicit steepening of the schedule, and an estimated value under 1 would indicate an implicit flattening. Results of this analysis are presented in Table 1.

In the first panel of Table 1, we restrict the estimation sample to the "local distribution," consisting only of the two locally sampled income draws and any of the randomly sampled income values that happen to fall in the respondent's own tax bracket. Beliefs in this region of the tax schedule directly reveal whether taxpayers correctly perceive the marginal tax rates that they face, and thus allow the assessment of Prediction 3. We find that our respondents underestimate the marginal tax rates in their own tax-bracket ($\beta = 0.81$, clustered s.e.=0.043), consistent with the prediction of ironing (3-I) and inconsistent with the prediction of spotlighting (3-S). The 2nd-5th columns of the table provide estimates of this same parameter when the sample is restricted to respondents in each of the four income quartiles. Across these estimates, we find that underestimation remains statistically detectable for respondents in the top two income quartiles. For respondents in the bottom two income quartiles the standard errors of these estimates are sufficiently large that we can reject neither correct perception, nor substantial underestimation (or overestimation) of their MTRs.

[20] Note that the estimation sample is restricted to cases with tax burdens in the range [0,55000).

The second panel of Table 1 presents estimates of β derived from the 10 random draws of the mid-range sampling distributions. As in the local analysis, we find substantial and statistically significant underestimation of the steepness of the tax schedule ($\beta = 0.82$, clustered s.e.=0.013). In contrast to the local analysis, we find that this underestimation persists among all four income quartiles, with the null hypothesis of $\beta = 1$ rejected at at least the 1% level for each estimate. The degree of underestimation is most severe among the lowest-income respondents: estimates range from 0.70 (clustered s.e.=0.029) for the lowest-income respondents to 0.94 (clustered s.e.=0.023) for the highest-income respondents. These results provide an investigation of Predictions 4 and 5, and again refute the predictions of the spotlighting model while supporting the predictions of the ironing model.

The third panel estimates β using data from the full sampling distribution. Results are congruent with those generated from only the mid-range sampling distribution. Underestimation of the slope of the tax schedule persists, and indeed is more severe ($\beta = 0.62$, clustered s.e.=0.010). As in the second panel, we find that all four income quartiles underestimate the average slope of the tax schedule, with the difference again increasing in the respondent's own income.

Summary: Assessing the five predictions articulated in section 2, we find no support for the predictions of the spotlighting model and consistent support for the predictions of the ironing model.

3.3.2 Disentangling Heuristic Use

While the reduced-form results support the presence of ironing and provide little evidence of spotlighting, these analyses do not address three questions of primary interest. First, how much of the apparent under-appreciation of progressivity can be attributed to ironing? Second, even if they are not prevalent, are there any spotlighters in the population? Third, are there features of average tax perceptions that schmeduling fails to capture? We address these questions with a mixture-model approach to estimating the propensity of each heuristic.

Estimating Heuristic Propensity: To provide quantitative estimates of the propensity of heuristic use, we estimate the structural model in equation (1). We present results from two estimating equations:

$$\tilde{T}_{f,i} = (1 - \gamma_I - \gamma_S)T(z_{f,i}) + \gamma_I \tilde{T}_I(z_{f,i}|z_i^*) + \gamma_S \tilde{T}_S(z_{f,i}|z_i^*) + \epsilon_{f,i} \tag{2}$$

$$\tilde{T}_{f,i} = (1 - \gamma_I - \gamma_S)(T(z_{f,i}) + r(T(z_{f,i}))) + \gamma_I \tilde{T}_I(z_{f,i}|z_i^*) + \gamma_S \tilde{T}_S(z_{f,i}|z_i^*) + \epsilon_{f,i} \tag{3}$$

In these equations, $\tilde{T}_{f,i}$ denotes the forecasts of the taxes due by the hypothetical taxpayer. Individual respondents are indexed by i, and iterations of the hypothetical taxpayer question are indexed by f. We model tax forecasts as a convex combination of three possible models of tax perceptions. We include the ironing and spotlighting forecasts as defined in section 2.1. We generate the ironing forecast ($\tilde{T}_I(z_{f,i}|z_i^*) = \frac{T(z_i^*)}{z_i^*} \cdot z_{f,i}$) by calculating the individual's federal tax due based on all reported information, dividing that by their reported income to generate an average tax rate, and multiplying that average tax rate by the hypothetical income assigned to Fred ($z_{f,i}$). We generate the spotlighting forecast ($\tilde{T}_S(z_{f,i}|z_i^*) = T(z_i^*) + T'(z_i^*) \cdot (z_{f,i} - z_i^*)$) by calculating the individual's federal tax due based on all reported information, and then adding to that the product of their statutory marginal tax rate and the difference between Fred's income and their own. In equation (2), we estimate a model in which aggregate tax forecasts are formed by a mixture of these two heuristics and the true tax liability ($T(z_{f,i})$). In equation (3), this latter term is augmented to $T(z_{f,i}) + r(T(z_{f,i}))$, denoting the true tax due plus a *residual misperception function.* By including this term and estimating it with a flexible functional form, we can separately identify our candidate heuristics from general misperceptions of the tax schedule not attributed to the models defined above. In the estimates we present below, we model the residual misperception function as a fifth order polynomial.

Table 2 presents non-linear least squares estimates of these models. Columns 1 and 3 present estimates of equation 2. Columns 2 and 4 present estimates of equation 3, integrating the residual misperception function into the analysis. Standard errors are clustered at the respondent level. Column 4 is our preferred specification because it utilizes all randomly sampled data and includes the residual misperception function.

The first two columns, which estimate the model from the mid-range sampling distribution, show a substantial weight on the ironing heuristic. In column 1, the point estimate implies 21% weight on the ironing heuristic. However, the point estimate on the spotlighting forecast is negative 9%—outside the range of valid probability values, and marginally significantly so. We view the estimation of invalid probabilities for heuristic propensity as evidence of model mispecification, and a demonstration of the difficulty of inference in this setting when non-income-dependent misperceptions are not accommodated.[21] Illustrating that point, when the residual misperception function is included in this estimation in column 2, weight on the spotlighting heuristics becomes statistically indistinguishable from zero, while weight on the ironing heuristic increases to 29%. The contrast of columns 1 and 2 demonstrates the importance of allowing for residual misperceptions when estimating the propensity of these heuristics: since these heuristics can change the level of aggregate tax forecasts, their identification can be confounded with level effects when residual misperception is not accommodated.

In columns 3 and 4, we repeat the estimation exercise of columns 1 and 2 using the full sampling distribution. In these specifications, we again find substantial weight on the ironing forecast and effectively zero weight on the spotlighting forecast. In the fourth column, which includes a

[21] To rationalize this result and illustrate the confound introduced by excluding controls for residual misperceptions, recall that we found systematic overestimation of the taxes due across the mid-range income distribution. Quantitatively, this overestimation cannot be generated by ironing. But because the spotlighting heuristic generates underestimation of taxes due outside of one's own bracket, placing negative weight on this heuristic is a simple way for the model to approximate our finding of systematic overestimation of tax levels.

residual misperception function and utilizes all randomly-sampled forecasts, we estimate an ironing propensity of 43% (clustered s.e.=9.5%) and a spotlighting propensity indistinguishable from zero (-2%, clustered s.e.=7.6%). This estimated ironing propensity is somewhat larger than, but statistically indistinguishable from, the estimate of 29% derived from mid-range forecasts. In sum, this analysis strongly supports the presence of a large subpopulation of ironers and provides no evidence supporting the presence of spotlighters.[22]

Interpreting Residual Misperceptions: The estimates of our mixture model suggest that reliance on the ironing heuristic is common, but not universal. We now turn to characterizing the residual misperceptions held by those who do not iron.

Figure 4 plots the estimates of the residual misperception function generated in the regressions of Table 2. The top panel presents the estimates generated from the mid-range sampling distribution. Residual misperceptions over the mid-range sampling distribution are characterized by a reasonably uniform overestimation of taxes due. The bottom panel presents the estimates incorporating the high-income sampling distribution. We again observe relatively stable overestimation over comparatively low tax bills. However this eventually transitions to (statistically insignificant) underestimation of higher tax burdens. Taken at face value, the initial upward slope of residual misperceptions implies overestimation of the MTRs associated with comparatively small tax bills. The gradual downward slope that follows implies modest underestimation of the MTRs associated with large tax bills.

It is noteworthy that this simple representation of residual misperceptions arises despite its specification as a fifth-order polynomial. Our estimation procedure is capable of detecting substantial nonlinearity in residual misperceptions if they are present, but these analyses suggest that the misperceptions that remain are comparatively simple in structure. Because the residual is so simple in structure, our estimated heuristic propensities are insensitive to the parameterization of the residual misperception function. Reestimating equation 3 with the residual misperception function specified as any polynomial of orders 1 through 10 results in estimated heuristic propensities within two percentage points of the estimates reported in Table 2 (see Appendix Table A4). Estimates change more meaningfully when misperceptions are not included or are assumed to be constant, but these results suggest that a first-order polynomial is sufficient to remove the confounding influence of residual factors.

Care is needed in interpreting our estimates of residual misperceptions. When designing our estimation strategy, this component was included to capture any apparent non-schmeduling misperceptions, either real or spurious. As examples of the latter category, note that a spurious appearance of misunderstanding of progressivity could arise under certain structures of nonclassical

[22] Some readers may worry that our finding of near-zero spotlighting is driven by one extreme feature of a spotlighter's forecasts: predictions of negative tax liability for comparatively very low incomes. As we demonstrate in Appendix Table A3, this feature has little impact on our estimates. In that table, we recreate the analysis of Table 2, but replace spotlighters forecasts of negative tax liability with a forecast of zero tax liability. We find that this different coding has a negligible impact on our point estimates.

measurement error. Similarly, apparent forecasting bias may arise if respondents improperly treat our survey questions as if they ask about a tax filer more complex than the one considered. Similar bias would arise if respondents integrate tax burden from non-federal sources, such as state or FICA taxes. It is a crucial feature of our approach that such confounds result in perturbations of the form modeled by our residual misperception function, and thus are controlled for in a manner that removes their influence from the estimates of the schmeduling heuristics' propensity. However, interpretation of residual misperception function must be conducted under the caveat that its shape may be influenced by such non-externally-valid features. Consequently, out-of-sample predictive power of this component of the mixture model may fail.

Summary: When estimating our structural model of misperceptions, we find evidence of widespread reliance on the ironing heuristic and no evidence of reliance on the spotlighting heuristic. Our preferred specification implies 43% weight on ironing, no weight on spotlighting, and residual over-estimation of tax burdens that stays fairly constant throughout the income distribution but eventually turns to underestimation for very high incomes.

3.3.3 Ironing Explains Estimates of MTR Misperceptions

Our approach to estimating heuristic propensity relies on forecasts of taxes due over a wide range of incomes. While such data is useful for identification, optimal response to tax policy is often contingent only on "local" knowledge of the tax schedule, and specifically knowledge of one's own MTR. Because individuals have more reason to learn about taxes on incomes close to their own, a reasonable reader may worry that models estimated using tax forecasts over a wide range of income provide a misleading depiction of MTR misperceptions. In this subsection, we assess the degree to which our estimated model accounts for these local perceptions, and evaluate goodness of fit when the predictions are only based on our estimates of ironing propensity.

To address these questions, we assess the ability of our estimated model to fit the model-free estimate of own-bracket MTR underestimation reported in Table 1. We do this both by estimating our model using the full sample of income draws, and also using only the income draws that only lie outside of the respondent's tax bracket.

As repeated in the first column of Table 3, we found that perceptions of the slope of respondents' own tax bracket was scaled by a factor of 0.81, indicating an average underestimation of MTRs. In column 2, we reestimate this regression while replacing respondents' actual forecasts with the forecasts predicted by the preferred specification of our model. We find a resulting scaling parameter of 0.86, quantitatively similar to (and statistically indistinguishable from) the model-free baseline.

To what degree do the different components of the mixture model contribute to this fit? Moving across columns 2 through 4, we progressively remove model components and reassess the fit to MTR perceptions. In column 3, we replace our MTR forecasts with those based on the estimated ironing propensity and residual misperception function, but setting the propensity for spotlighting equal to zero. We find that this has no meaningful effect on the resulting scaling parameter, consistent with the statistical insignificance of our estimate of spotlighting's propensity. In column 4, we further restrict the model to remove the influence of the residual misperception function: we forecast MTRs based on a 43% propensity to use the ironing heuristic, with remaining weight placed on the true tax forecast. In this analysis, we find that the resulting scaling of MTRs precisely matches that in the model free baseline, with a statistically insignificant difference of less than one percentage point. In summary, accounting for ironing alone allows a quantitatively precise forecast of MTR perceptions.

The comparisons conducted above help assess the fit of our estimated model. However, this test of fit is explicitly within-sample: forecasts of taxes due in one's own bracket are used both for the estimation of the model and in the model-free validation. To provide a more stringent test of our model's ability to capture MTR perceptions, we construct an analogous out-of-sample test of fit. We reestimate our model while excluding all data from respondents' own tax brackets.[23] We then use the estimated model to forecast tax perceptions within respondents' own brackets, and again compare the scaling. As shown in columns 5-7, this analysis yields results nearly identical to those in the within-sample exercise. Importantly, a simple model consisting of only ironers and correct forecasters predicts the scaling of MTRs to within a percentage point.

Summary: A simple mixture model of ironers and correct forecasters accurately predicts MTR misperceptions.

3.3.4 Heterogeneity Analysis

The results of Table 2 suggest that aggregate tax misperceptions can be rationalized by placing significant weight on the ironing forecast. This is perhaps most naturally interpreted through a heterogeneous model in which some individuals have accurate beliefs (or accurate beliefs up to the perturbation of the residual misperception function) and some individuals employ the ironing heuristic. In such a model, our estimated coefficients may be interpreted as the propensity of use for each of the candidate forecasting rules. However, in principle our results could alternatively be rationalized by a homogeneous decision rule that places some weight on the truth and some weight on the ironing heuristic. While we believe that such a model would be difficult to psychologically motivate, we present individual-level estimates of our model to help rule out this possibility.

To begin, we estimate equation 2 at the individual level for each of the 3,552 respondents facing a non-zero tax rate.[24] Figure 5 plots a kernel-density estimate of the distribution of estimated individual classifications. In general, this distribution is quite diffuse. This is to be expected,

[23] Compared to column 4 of Table 2, this restriction results in extremely similar estimates of heuristic propensity. We estimate an ironing propensity of 0.46 (clustered s.e.= 0.101) and a spotlighting propensity -0.04 (clustered s.e. = 0.079), and thus see differences of no more than 0.03 from the values in our preferred specification.

[24] Notice that for individuals facing zero tax, the ironing and spotlighting heuristics yield the same forecast, and are thus not separately identified.

since estimating two parameters from only 14 tax forecasts would lead to a distribution of point-estimates that is influenced by both the true parameters and individual estimation error. Notice, however, that the resulting estimates yield a sharply bimodal distribution. As illustrated in the figure, the peaks of this distribution correspond to the parameter values describing those relying fully on the ironing forecast and those relying fully on the correct tax forecast. Consistent with earlier analyses, no excess mass is seen at the parameter values corresponding to full reliance on the spotlighting forecast. While this analysis cannot rule out the existence of some intermediate cases, this distribution is consistent with a substantial population of pure ironers, in the sense intended by Liebman and Zeckhauser's (2004) presentation of the heuristic.[25]

We can also perform the same estimation exercise while restricting individual-level parameter estimates to valid probabilities. To do so, we evaluate equation 2 over a grid of probability values and assign individuals to the grid point that minimizes the mean squared error of their forecasts. We again see strong support for the existence of pure types: 33% of respondents are classified as "true tax" forecasters, 33% are classified as pure ironing forecasters, 2% are classified as pure spotlighters, and the remainder are estimated at intermediate values. Similar results are obtained by repeating the analog of this exercise that applies the model and data restrictions of each column of Table 2 (all such analyses are reported in Appendix Tables A5-A8).

The individual classifications produced by this procedure help in determining which groups' tax perceptions respond to average tax rates. In Figure 6, we plot approximations of the average perceived tax schedule (analogous to Figure 2) separately for those who are "close" to being classified as ironers (specifically, $\hat{\gamma_I} \in [0.6, 1.4]$ and $\hat{\gamma_S} \in [-0.4, 0.4]$) and those who are not.[26] Within each group, we separately plot the perceived tax schedule for those above and below median income. Among those not classified as ironers, the average perceived tax schedules for the comparatively high and low income are nearly identical. In contrast, among those classified as ironers, perceptions are notably different across income groups. The perceived schedules are approximately linear, and come quite close to intersecting the true tax schedule evaluated at the average income for the group. In short, those labeled as ironers by our classification procedure have perceptions that satisfy the model's core predictions.

[25] Similar figures can be generated estimating equation 3, and by restricting the estimation sample to only the mid-range income distribution. In Appendix Figure A1, we reproduce Figure 5 while applying the restrictions for each of the columns in Table 2. Since we cannot estimate residual misperceptions at the individual level, we use the estimated residual misperception function from the analysis of Table 2. Across all specifications, we continue to find a large mass at $\gamma_I = 1, \gamma_S = 0$, indicating that our inference about the existence of pure ironers is not confounded by the residual misperception function. However, when residual misperceptions are included, the density around $\gamma_I = 0, \gamma_S = 0$ is reduced by approximately half. This can be rationalized by interpreting the residual misperception function not as a literal forecasting rule by a pure "type," but as an approximation to heterogeneous remaining misperceptions beyond the heuristics we study.

[26] Figure 6 presents results for married filing jointly filers, the most common filing status in our data. Analysis of single filers, with very similar results, is available in Appendix Figure A2.

3.3.5 Summary of Robustness Analyses

In Appendix B we conduct extensive robustness checks on the sensitivity of our estimates to a variety of sample restrictions. Our empirical results persist when analyzing taxpayers with comparatively high incentives for accurate tax knowledge such as the employed, those who complete their returns without outside assistance, and those of comparatively high income. Our results persist among respondents who might be suspected to be debiased, such as those who complete our survey after tax day, those of comparatively high age (and thus comparatively high experience with the tax system), or those of comparatively high confidence in their answers.

Our results are also robust to several possibilities for misunderstanding of the survey prompt. Allowing for the possibility that respondents' forecasts reflected their beliefs about not just federal income taxes but the sum of federal and state and/or payroll taxes does not change our estimates of ironing propensity. This is because such misunderstandings take a structural form accounted for with the residual misperception function. Our results are also robust to restrictions of the sample to those whose tax returns are and are not "simple" like those of the hypothetical taxpayer Fred.

Finally, our results are also robust to alternative treatments of inattentive respondents and Winsorization schemes.

Across the restrictions we consider, the estimated subgroup propensity to iron ranges from 24% to 58%; indicating widespread and cross-group prevalence. Across these robustness checks we continue to find no evidence of spotlighting.

4 Study 2: Inferring Ironing From Choice Data

A key feature of Study 1 is that it examines perceptions of the actual U.S. income tax schedule. However, an important limitation is that we cannot directly verify that individuals indeed act on the beliefs that they state. To partially address this limitation, we conducted a complementary online experiment in which participants' perceptions of the marginal tax rates can be directly revealed by their behavior. Beyond utilizing choice data, the online experiment has several other desirable features. First, the experiment uses a different source of identifying variation. Rather than utilizing naturally occurring variation in participants' tax rates, the experiment induced truly exogenous variation in average and marginal tax rates by varying the tax schedules that participants faced. Second, the experiment allows us to examine the extent to which schmeduling might be endogenous to the complexity of the tax schedule. Third, the experiment focuses participants on decisions about incremental changes to their taxable income, rather than focusing them on total tax paid—a framing manipulation that could in principle induce more attention to marginal tax rates. Fourth, the experiment also changes the frame by focusing individuals on after-tax income rather than on the taxes themselves.

Our experiment was preregistered on aspredicted.com. We report how we determined sample size, all data exclusions, all manipulations, and all measures below. The full experimental protocol is available in the survey appendix.

4.1 Experimental Design

Our experiment presented participants with a decision of whether to allocate a bonus to a taxable or non-taxable account. As we explain below, this decision allows us to infer individuals' perceived MTR from their choices.

Summary of experimental protocol: The experiment began with a standard elicitation of informed consent and a CAPTCHA task. After this introduction, participants were presented with a summary of the task they would face and an initial presentation of the tax schedule. The schedule is randomly generated according to criteria we describe in detail later in this section. Figure 7 presents this screen.

After this explanation, participants were asked a basic comprehension question that tested their understanding of which accounts were taxable. Participants were not able to proceed with the experiment if they failed either this comprehension test or the CAPTCHA.

Participants were then told that, as a baseline, they are given 40 cents in account A and 60 cents in account B. For each account, we presented the total tax on the earnings and the effective (average) tax rate on the earnings. In the example provided in Figure 7, this prompt would have indicated 20 cents of tax and a 50% effective tax rate on account A. The tax amount and rate for account B were always 0 and 0%, respectively. We used the language "effective tax rate" in lieu of "average tax rate" to be consistent with the language used by tax preparation software such as Turbotax. The presentation of both the absolute tax and the average tax rate also mimics the summary screen of such tax preparation software.

Immediately after the summary, we included a button that, if clicked, would display the tax table again on the screen. This provides an additional measure of the number of participants who thought that additional information beyond the ATR was necessary. Because some participants could have clicked the button without an intention to actually study the tax table, we think that it is best to regard the propensity to click on the button as an upper bound on how many participants felt it was necessary to consult the tax table to make their decision.

Participants were then asked to make 12 decisions about additional money to be placed in the taxable and nontaxable accounts. These decisions were made in a multiple-price-list (MPL) format. In each decision, one option was adding 20 cents pre-tax to account A. The other option was a non-constant amount of earnings in account B, with the payment (in cents) taking the values {0, 1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21}.

Following this decision, participants faced a common attention check and a battery of demographics questions. After these questions, one of the twelve MPL decisions was randomly selected and final compensation was determined based on that decision and reported to the subject.

Variation in tax schedules: We generated exogenous variation in the ATR and MTR, as well as in schedule complexity, by randomly assigning participants to one of fifty four tax schedules. Half of these schedules were *simple*: they consisted of two brackets for income ranges [0, 40] and [40, 100]. Half of the schedules were *complex*: they consisted of five brackets for income ranges [0, 20], [20, 40], [40, 60], [60, 80], and [80, 100].

For each MTR/ATR pair, there was a unique simple schedule and a unique complex schedule constructed to feature those rates. For simple schedules, the rate in the first bracket determines the ATR and the rate in the second bracket determines the MTR. For complex schedules, we set the rate in the first two brackets to ATR-0.05 and ATR+0.05, respectively, averaging to the ATR. The rate in the third bracket, which governs the marginal incentives for the relevant decision in the study, is set to the desired MTR. The fourth and fifth brackets have rates of MTR+0.05 and MTR+0.10, respectively.

Our (ATR, MTR) combinations were generated by sampling sets of three schedules. Each set of three schedules had the form $\{(ATR_0, MTR_0), (ATR_0 + 0.3, MTR_0), (ATR_0 + 0.3, MTR_0 + 0.3)\}$. Within each set, there are large and exogenous differences in both the ATR and the MTR, which facilitates high-powered estimation of reliance on average versus marginal tax rates. To make sure that our results were not tied to any particular set, we generated nine such sets by drawing a value of ATR_0 and MTR_0 from the set $\{0.2, 0.3, 0.4\}$, uniformly and with replacement. Because we sampled with replacement, 26% of the two-bracket schedules had the same tax rate in both brackets, and 15% of the five-bracket schedules had the same tax rate in the second and third bracket.

Quantifying revealed-perceived MTRs: We generate our measure of "revealed-perceived MTRs" using decisions in the MPL. If the participant chooses the option that he thinks has the highest financial payoff, answers in the MPL set-identify the perceived MTR. For example, if the participant perceived that the MTR was 0.2, he would expect to earn 16 cents post-tax if 20 cents are placed in account A. He should thus prefer to receive the bonus in account B if the amount offered is 17, 19, or 21 cents, and would take the bonus in account A otherwise. This pattern of decisions would imply that $(1-\widetilde{MTR})*20 \geq 15$ and $(1-\widetilde{MTR})*20 \leq 17$, allowing us to infer that $\widetilde{MTR} \in [0.15, 0.25]$. This measure of perceived MTR is our object of interest in this experiment. In principle, one could use approaches such as interval regression to analyze this dependent variable, as we do in Appendix C. However, the identified intervals are sufficiently narrow that assigning responses to their midpoint (e.g., to 0.2 in the example) and relying on OLS regressions yields essentially identical results. In the body of the paper, we report OLS regressions with midpoint coding of the dependent variable.[27]

This method of identifying revealed-perceived MTRs was previously applied in the experimental work of de Bartolome (1995). In that experiment, participants were presented with a single tax

[27] For participants who preferred 1 cent in the untaxed account over 20 cents in the taxable account, we code the revealed perceived MTR interval to be [0.95,1], and the midpoint to be 0.975. For participants who preferred 20 cents in the taxable account over 19 cents the untaxed account, we code the interval as [0, 0.05] and the midpoint as 0.025.

schedule and a comparatively coarse MPL, and the analysis revealed that a relatively large number of respondents had perceived MTRs in the bin that contained the ATR. This analysis is suggestive of ironing. However, participants who simply underestimate the MTR without relying on the ironing heuristic could similarly fall in that bin, as could participants who are inattentive to the tax schedule and either choose randomly or rely on some other rule of thumb. Variation in the ATR and MTR, as in our experiment, is crucial for separately identifying ironing from possibilities such as these.

4.2 Sample for Analysis and Dataset Preparation

Our study ran on Amazon Mechanical Turk in December of 2018 and January of 2019. During this time, we collected 4,582 complete responses. We preregistered three exclusion criteria for our sample. First, we exclude any subjects who failed the attention check included at the end of the study. This requirement screens 714 responses (15.58% of the initial sample). Second we exclude any subjects with MPL responses inconsistent with monotone preferences (i.e., not exhibiting a single switching point). This requirement screens an additional 179 responses (4.63% of the remaining sample). Finally, we included "trick" options at both ends of the MPL: one offering zero cents in the tax free account (which should not be preferred unless one expects a 100%+ tax rate) and one offering twenty one cents in the tax free account (which should always be chosen unless one expects negative tax rates). We exclude subjects who do not take the dominant option for each, screening an additional 559 responses (15.15% of the remaining sample). These exclusions yield a final sample of 3,130 responses, closely in line with our preregistered target of 3,000. Our primary sample is comprised of 53% women, has a median age of 35, and has a median income in the range of $40,000-$49,999.

We chose to preregister these stringent inclusion criteria because we are most interested in analyzing the responses of attentive subjects who show clear understanding of the task at hand. Of course, analyzing even those who fail to meet that bar can be of interest, although we note that inattention or lack of comprehension to our experiment need not be an externally valid indication of problems with understanding real tax systems. Appendix C reproduces our results with inattentive responses re-included.

4.3 Results

Our core, pre-registered analysis estimates the regression $\widetilde{MTR}_i = \alpha + \beta_1 MTR_i + \beta_2 ATR_i + \epsilon_i$. Viewed through the lens of our empirical model, β_2 identifies the fraction ironing and β_1 identifies the fraction responding to MTRs. As discussed in section 2.2, spotlighting and correct forecasting are not separately identifiable when examining only local forecasts, and thus β_1 estimates the pooled propensity of either behavior. The constant term captures the average revealed-perceived MTR of those not responding to the true MTR or ATR, scaled by $(1 - \beta_1 - \beta_2)$. The behavior of these remaining filers may be thought of as a rule-of-thumb that is relied on without reference to the schedule, perhaps motivated by simple inattention to taxes as in Chetty et al. (2009) and Taubinsky and Rees-Jones (2018).

Column 1 of Table 4 presents the results of this regression. Our estimate of the coefficient on ATR is consistent with 35% of our sample applying the ironing heuristic (s.e.=2 p.p.). In contrast, our estimate of the coefficient on MTR is consistent with only 9% of the sample responding to this feature of the schedule (s.e.=2 p.p.). Among the 44% of respondents who respond to one of the randomly varied features of the schedule, 79% respond to the ATR. Consistent with the survey study, these results suggests a remarkably high reliance on the ironing heuristic.

A strong prediction of our empirical model, consistent with the mixture model interpretation, is that the revealed-perceived marginal tax rates should increase linearly with the ATR and MTR. Consistent with this prediction, we find no evidence of nonlinear relationships. Comparing the fit of our preferred model to one with a dummy variable for each discrete level of ATR and MTR, we find no meaningful improvement (see Appendix Table 4). A likelihood ratio test fails to reject the null of the linear model ($p = 0.417$), providing support for the functional form assumptions of our estimating equation.[28]

Across the remaining columns of Table 4, we examine the sensitivity of the estimates to several different subsample restrictions. First, we provide separate estimates for "simple" (column 2) and "complex" schedules (column 3). Point estimates do not differ substantially across these samples; column 4 presents the estimated differences and standard errors from the fully interacted model and finds no significant differences. Furthermore, a likelihood ratio test fails to reject the null of the non-interacted model (p = 0.164). These results show no support for tax schedule complexity affecting the propensity to iron.

Next, we split our results by whether participants re-examined the tax table when given the opportunity. Interestingly, only 46% of of participants chose to do so. And we continue to find substantial prevalence of ironing even among those reexamining the tax table; indeed, these subjects are marginally statistically significantly *more* likely to iron. But consistent with some participants choosing to inattentively rely on rules of thumb instead of examining the tax schedules, we do see that those who clicked have a statistically significantly smaller constant term.[29] Overall, these results show that many participants rely on the ironing heuristic without taking the time to look up the tax table to make an optimal decision, even when doing so is nearly costless. This is despite the fact that in our experiment, examining the tax table to compute the tax on the incremental 20 cents is not more computationally costly than multiplying one's average tax rate by 20.

[28] Appendix Figure A3 plots the residualized relationship between average perceived tax rates and each of ATR (controlling for MTR) and MTR (controlling for ATR). The relationship appears remarkably linear when visually assessed.

[29] The fact that the coefficients on the ATR and MTR still do not sum to one even for the subsample that clicked on the tax schedule is consistent with some participants arbitrarily clicking on the tax table without an intention to really engage with it. It is similarly consistent with some respondents clicking on the tax table, deciding it is too complex to parse, and then choosing to rely on some rule of thumb instead.

Finally, and in analogue to our analysis in the survey study, we ask to what extent ironing provides a complete explanation of the individuals' misperceptions. To do so, we first estimate a benchmark of "maximum explainable variation" (Kleinberg et al., 2017). The explainable variation that is due to systematic behavioral responses to the differences in our tax schedules is simply the R^2 of the regression of perceived perceived MTRs on the 54 dummies for our tax tables. Given our large sample size, we have on average approximately 60 participants per tax table, and we are thus able to estimate this statistic with a reasonable degree of precision. The sampling error resulting from a finite number of participants allocated to any given tax table leads to (in our case, small) overfitting problem which makes the R^2 estimate to be a slight upper bound on the maximum explainable variation. We estimate that the maximum explainable variation corresponds to $R^2 = 0.1055$.

We contrast this with the R^2 of the regression model in column 1 of Table 4, which we find to be 0.0924. This implies that the simple linear model explains over 87% of the explainable variation in the data. The economic content of this result is that relaxing the assumptions of linearity, additive separability, and invariance to the complexity of the tax schedule does not meaningfully increase explanatory power. Consistent with this, we find that a likelihood ratio test fails to reject the null of our simple linear model (p =0.69).[30]

We provide a number of robustness checks of these results in Appendix C. In particular, we conduct analyses relaxing our inclusion criteria, we consider alternative definitions of "simple" and "complex" that correspond to exactly 1 or 4 kinks, respectively, and we examine interactions with demographic covariates. Across these tests, we find little deviation from the results of the primary specification reported here.

Summary: Consistent with with the results of Study 1, we find prevalent ironing and little evidence of other misperceptions.

5 Welfare Implications of Ironing

In this section, we explore the quantitative importance of ironing for welfare calculations in standard models of intensive-margin earnings decisions and distortionary taxation. Our model has the general features of the canonical frameworks of Mirrlees (1971) and Saez (2001), modified to allow for misperceptions as in Farhi and Gabaix (2018).[31] To keep exposition concise, we briefly provide intuition for the important theoretical consequences and we focus attention on quantitative estimates corresponding to our preferred model specification. In Appendix D, we provide formal analysis that corresponds to the theoretical claims summarized here.

[30] Appendix Table A10 examines how relaxing each of the three assumptions of our linear model improves model fit.

[31] These standard models are partial equilibrium models in the sense that they assume wages are fixed. Note that in a general equilibrium model, the presence of ironing could affect wages themselves, as well as the labor supply elasticity with respect to those wages.

5.1 Model and Assumptions

Economic Setting: Individuals have a utility function $U(c, l)$, where c is consumption and l is labor. Individuals produce $z = wl$ units of income for every l units of labor, where the wage w is drawn from an atomless distribution F. The government cannot observe earnings potential w, and is thus restricted to setting taxes $T(z)$ as a function of earnings z. More generally, and as in Feldstein (1995; 1999), this model serves as a model of taxable income choice. This encompasses, for example, decisions over tax-preferred activities such as charitable contributions.[32]

Optimizing individuals choose $z^* \in \text{argmax}\{U(z-T(z), z/w)\}$. Ironers choose $z^* \in \text{argmax}\{U(z-A(z^*)z, z/w)\}$, where $A(z^*) = T(z^*)/z^*$ is the average tax rate. Notice that for ironers, z^* is a fixed point of a decision process in which misperceptions are possibly shaped by z^*, while at the same time z^* is a perceived optimum given those misperceptions.[33] In Appendix D.1 we provide existence and uniqueness results as well as basic comparative statics for this solution concept, which have been implicitly assumed by Liebman and Zeckhauser (2004).

The social welfare function is given by

$$\int U(z^*(w, 1_\gamma) - T(z^*(w, 1_\gamma)), z^*(w, 1_\gamma)/w)dH + \lambda \int T(z^*(w, 1_\gamma))dH, \tag{4}$$

subject to $\int T(z^*(w, 1_\gamma))dH \geq 0$, where $1_\gamma \in \{0, 1\}$ is an indicator for ironing, H is the joint distribution over wage and ironing types, and λ is the marginal value of public funds. Denote the social marginal welfare weight at the current tax system by $g(w, 1_\gamma) = U'_c/\lambda$.

Assumptions for Numerical Results: We parameterize individual utility according to the functional form $U(z) = log(z - T(z) - \frac{(z/w)^{1+k}}{1+k})$, a commonly used specification in optimal tax studies (e.g., Atkinson 1990; Diamond 1998; Saez 2001). We parameterize the tax function $T(z)$ according to the U.S. tax code in 2014—the tax code relevant for forecasts in Study 1. In this model, assuming correct tax perceptions, the structural labor supply elasticity is determined by $\frac{1}{k}$. When tax rates are misperceived, the elasticity with respect to wages must be scaled by the term $\frac{1-\tilde{T}'}{1-T'}$, which takes an average value of 1.02 in our data—thus, while the formal calculation of elasticity is not identical, quantitatively the difference is negligible. In simulations, we will vary the parameter k across values from 2 to 5, capturing elasticities ranging from approximately 0.2 to 0.5. In a recent meta-analysis of labor-supply elasticity estimates, Chetty et al. (2011) report a preferred estimate

[32] See Chetty (2009) for some exceptions to this generalization, and Slemrod and Kopczuk (2002) for discussion of its impact on the interpretation of labor-supply elasticity.

[33] Farhi and Gabaix (2018) implicitly use this solution concept in their study of optimal income taxation. In Appendix D we formalize the solution concept and characterize existence or uniqueness for the types of misperceptions that we estimate. The solution concept may be reformulated as a special-case of Berk-Nash equilibrium (Esponda and Pouzo, 2016) and, as such, can be microfounded as a steady state of a dynamic process in which individuals follow a myopic best-response strategy while learning through a misspecified model. To formally embed our model in the Berk-Nash framework, we must re-interpret $\tilde{T}(z|z^*)$ as the mean of the individual's belief, while allowing the individual to have a sufficiently diffuse prior so that no outcomes are "surprises." See also Gabaix (2014) for a general approach to modeling boundedly rational misperceptions of incentive schemes.

of the intensive-margin Hicksian elasticity of 0.33, approximately corresponding to $k = 3$.

Contingent on these parameters, we calculate the wage parameter that rationalizes reported earnings for each individual in the dataset. An individual's wage parameter can be calculated according to the equation $w = (\frac{z^k}{1-\tilde{T}'})^{\frac{1}{1+k}}$. Since this calculation depends on the individual's perceived marginal tax rate $\tilde{T}$, we calculate this value both under the assumption of ironing and under the assumption of correct forecasting.

When considering the consequences of a policy reform, we forecast the aggregate consequences of the individual reactions implied by the specified utility model and individual utility parameters. We forecast each individual's predicted response contingent on being an ironer or a correct forecaster, applying the relevant estimated wage parameter in each case. We aggregate these individual responses by assuming that behavior is determined by the ironing model with probability $\gamma = 0.43$, and otherwise determined by the correct forecasting model. This estimate is drawn from our preferred specification in Study 1.[34] We generate analogs to 95% confidence intervals by alternatively assuming that $\gamma = 0.25$ or $\gamma = 0.62$—the boundaries of the 95% confidence interval on ironing propensity in our preferred regression model. Focusing on the ranges of effects that arise over this broad interval helps to illustrate the potential for welfare impact, even given a reasonable degree of uncertainty as to the precise propensity of ironing.

5.2 Results

Implication 1: Government revenue increases due to ironing, and this revenue is raised progressively. Thus, ironing increases social welfare. When the tax schedule is progressive, marginal tax rates will be higher than average tax rates, and thus ironing will lead to underestimation of marginal tax rates. As established in de Bartolome (1995) for a linear income tax, and in Liebman and Zeckhauser (2004) and Appendix D more generally, this implies that labor earnings and thus government revenue will increase in the presence of ironers. Numerically, we find that under the U.S. tax schedule, our estimated ironing propensity leads to a 1.6-3.7% increase in government revenue (reported in column 2 of Table 5), depending on the labor supply elasticity. For our preferred elasticity of $\frac{1}{3}$, the revenue gain is 2.2% (95% CI: 1.5%-3.7%).

While ironing raises revenue, it also generates individual misoptimization and therefore imposes costs on the affected individuals. Which individuals bear those costs significantly influences social welfare. In general, the cost of the mistake of ironing is proportional to a financial loss of $T'(z) - A(z)$, the difference between the marginal and average tax rates. To the extent that the difference between the marginal and average tax rate will increase with earned income, the financial burden of misoptimization will fall on the higher income individuals. For example, while the lowest income individuals will not misoptimize at all, the wedge will be substantial for the highest income

[34]While our primary estimate from Study 2 is quantitatively similar, we believe our Study 1 estimates are most appropriate for this exercise due to their focus specifically on perceptions of the U.S. tax code.

individuals of our sample.

Figure 8 illustrates the estimated burden of misoptimization across different income levels, taking as given the structure of the U.S. income tax code and our estimates of ironing propensity. For each individual, we first calculate the compensating variation of ironing: that is, the amount of money that the individual would have to receive to be as well off ironing as he would be if he optimized with correct forecasts. We then plot the average compensating variation for each level of earnings, normalized by the population wide average of revenue generated by ironing on the right y-axis. This figure illustrates that the individual cost of misoptimization induced by ironing is increasing in income, and is generally well below the average revenue generated by its presence, suggesting the potential to exploit this bias and redistribute the extra government revenue it provides in a welfare-enhancing manner. In short, the burden of ironing is starkly progressive.

The first two observations—that ironing counteracts the distortionary effects of taxation by raising earnings, and that it increases government revenue in a progressive fashion—lead to the implication that ironing leads to progressive revenue collection, which will typically be welfare-improving.

To quantify the impact of our estimated propensity of ironing on welfare, columns 3-5 of Table 5 present estimates of the fraction of current government revenue the social planner would pay to avoid reducing the rate of ironing to zero. Calculating social welfare requires specifying the value of public funds, λ. In our preferred specification, we adopt the standard assumption that λ is equal to the average marginal utility of consumption in the population.[35] To illustrate the sensitivity of results to that assumption, we additionally present results for a "low λ" regime (in which we set λ to be equal to the 50th percentile of marginal utilities in our population) and a "high λ" regime (in which we set λ to be equal to the 90th percentile of marginal utilities in our population). Across different assumed elasticities and values of public funds, we find that the improvement to social welfare realized from the presence of ironing is valued equivalently to an unfunded increase in government spending ranging from 1.4% to 3.6%, with 2.3% (95% CI: 1.4%-3.4%) corresponding to our preferred specification. As compared to the raw increase in government revenue presented in column 2, this indicates that the welfare costs of individual misoptimization have a minimal offsetting effect to the additional spending funded by ironing.

Implication 2: Ironing increases the welfare benefits of progressive taxation. As we have already shown, ironing results in progressive revenue generation when the income tax schedule is itself progressive. Note, however, that the presence of ironing becomes irrelevant under linear, "flat" taxes, since such systems equate marginal and average tax rates. In effect, then, a tax simplification scheme that bring the tax schedule closer to a flat tax will reduce the socially beneficial influence of this bias. In other words, tax simplification, in addition to changing material incentives, can

[35] When calculating this average, we assume an income floor of $6,000 to approximate the provision of social insurance that is outside of our current model.

indirectly generate a type of debiasing.

To illustrate this numerically for the U.S. income tax system, we consider the welfare effects of moving to a flat tax. We constrain the marginal tax rate of the flat tax to be such that the amount of revenue raised would be identical in the absence of substitution to or from leisure: 11.06%.[36] As shown in columns 2 and 3 of Table 6, perfect tax forecasters on average increase their labor in response to this tax reform, leading to a 1.9-5.3% increase in tax revenue. However, because this reform leads to a less progressive tax system, and because we have assumed a social welfare function valuing redistribution, the reform leads to a substantial reduction in welfare. Quantitatively, this welfare loss is equivalent to a loss of 9.9-12.9% of government revenue under the pre-reform tax system.

Turning to the estimates accounting for ironing in columns 4 and 5, we see that the revenue benefits of the flat tax are dampened by the less elastic response of ironers. This lower revenue, combined with the fact that ironing amplifies the decrease in progressivity generated by the flat tax, results in a more severe welfare loss than would be obtained under the assumption of perfect tax perceptions. The welfare losses range from 12.3-13.8% of pre-reform government revenue. In our preferred specification, the welfare loss associated with moving to a flat tax is 13.2% of pre-reform government revenue (95% CI: 12.5%-13.9%). Compared to the estimates that assume correct forecasting, this indicates that the welfare costs of this reform are 14% higher (95% CI: 8%-20%) when the effects of ironing are incorporated.

Another simple way to analyze the impact of ironing on the benefits of progressivity is to analyze the optimal top marginal tax rate. Saez (2002b) shows that the top marginal tax rate $\bar{\tau}$ satisfies $\frac{\bar{\tau}}{1-\bar{\tau}} = \frac{1-\bar{g}}{a\bar{\varepsilon}}$ where $\bar{g}$ is the average social marginal welfare weight on the top income earners, $\bar{\varepsilon}$ is the structural elasticity, and a is the Pareto parameter of the income distribution.[37] In the presence of ironing, the top marginal tax rate satisfies $\frac{\bar{\tau}}{1-\bar{\tau}} = \frac{1-\bar{g}}{[(1-\bar{\gamma}_I)a+\bar{\gamma}_I]\bar{\varepsilon}}$, where $\bar{\gamma}_I$ is the propensity to iron among high income earners.[38] Since the Pareto parameter $a > 1$, the optimal top marginal tax rate is higher in the presence of ironing. For a Pareto parameter $a = 2$ estimated by Saez (2002b), for example, $\frac{\bar{\tau}}{1-\bar{\tau}}$ is 27% higher with 43% propensity to iron as opposed to 0%.

Our analysis of tax schedule design highlights the importance of establishing the mechanisms underlying misperceptions of marginal tax rates. Based on our estimated model, we are able to forecast the changes to misperceptions (and the behaviors they dictate) after reforms to the tax schedule—a task for which reduced-form estimates of MTR perceptions are insufficient.

[36]Similar results obtain if we instead constrain the flat tax rate to raise the same amount of revenue after accounting for substitution effects (see Appendix E).

[37]For simplicity we consider the case where there are no income effects on labor supply.

[38]See Appendix D.3 for a derivation.

5.3 Interpretation and Caveats

Across these two classes of analyses—examining the welfare impact of introducing ironing under our current tax system, and examining the impact of ironing on a flat-tax reform—we find evidence that this bias serves a useful role for the social planner.[39] Ironing leads to additional progressive revenue collection compared to a baseline of perfect optimization, and this is consequential for welfare analysis of redistributive tax policy.

By focusing on the canonical income taxation framework, the welfare analysis here omits the cost of mistakes not related to choices about earnings and taxable income; for example, improper budgeting. All of our statements about welfare must therefore be interpreted as holding constant all other changes. While noting this limitation, we believe that the effect of ironing on earnings distortions would be of primary importance even in a more holistic welfare analysis that incorporates the costs of other potential mistakes.

We wish to caution readers against two implications one might erroneously draw from these results. First, since this heuristic use is beneficial in our model, one might infer that we endorse further obfuscation of the tax system with the goal of raising the population's use of heuristics. Note that the standard model of earnings choice that we have adopted abstracts entirely from political economy issues, which would be critically important in assessing such a proposed reform. We believe that there are meaningful, and potentially dramatic, costs associated with a populace coming to believe that a tax system is actively designed to mislead. Any policy recommendations on optimal obfuscation must derive from a model which incorporates those costs, which we do not.

As a second note of caution, we emphasize that our finding that heuristic use improves social welfare is specific to domain that we study: misperceptions of the tax rates of the federal income tax schedule for earned income. We have abstracted from misunderstanding regarding various other components of the schedule. For the wealthy, optimal decision making might additionally rely on knowledge of the alternative minimum tax or the estate tax.[40] For the comparatively low income, misunderstanding of the interaction of the tax schedule with assistance programs may generate additional misperceptions (Romich, 2006), and failure to optimize along these dimensions may be especially costly (Currie, 2006).[41] However, the likely presence of these other mistakes does not limit the validity of our specific claim: that ironing of the federal income tax schedule leads to progressive revenue collection and amplifies the benefits of progressivity.

[39] For other examples in which behavioral biases improve social welfare see Handel (2013); Handel and Kolstad (2015); Spinnewijn (2017); Handel et al. (forthcoming); Mullainathan et al. (2012).

[40] See Kuziemko et al. (2015) for documentation of misunderstanding of the estate tax.

[41] Concretely, a substantial literature documents misunderstanding specifically of the Earned Income Tax Credit (see, e.g., Liebman, 1998; Romich and Weisner, 2000; Chetty et al., 2013).

5.4 Summary of Additional Results

In Appendix E, we support the implications of our simulation results in a formal theoretical framework. We first provide an instructive step-by-step analysis of a two-bracket model in Appendix D.2, which elucidates the mathematical intuition behind our primary claims. We then generalize this in Appendix D.3, which provides formulas for the behavioral and welfare effects of raising tax rates on earners with incomes above a threshold. In Appendix D.3.3, we use that formula to calculate the top marginal tax rate, and in Appendix D.3.4 we use the formula to calculate the optimal income tax in a heterogeneous population of ironers. Our formula subsumes that of Liebman and Zeckhauser (2004) for the case of a population of only ironers, and complements the Farhi and Gabaix (2018) formula for the case of homogeneous partial ironers.

In Appendix E, we provide a series of robustness checks on our simulation results. We demonstrate that our highlighted implications still hold under weaker assumptions on the social planner's preference for redistribution, under different derivations of the flat-tax rate, and under different corrections for the absence of extremely high-income filers in our sample.

6 Discussion

A large and growing literature in behavioral economics shows that people rely on heuristics when facing complex incentives. We contribute to this literature by studying misperceptions of income taxes—a notoriously complex set of incentives with active public debate promoting simplification. We show that much of the systematic misperception of the income tax can be explained by widespread reliance on a single, simple heuristic: ironing. The ability to account for misperceptions with a single parsimonious model allows rigorous analysis of questions about tax reform. We provide such an illustrative analysis, in which the welfare effects of ironing are positive and economically significant.

Moving beyond applications specific to the design of optimal tax policy, we highlight that our empirical estimates are relevant for broader classes of tax incentives. When considering the adoption of tax-preferred behaviors, the investment in human capital in the hopes of raising future wages, or the pursuit of financial investments that will only accrue at a future date, our results provides a unique view into the tax perceptions that could shape such decisions.

Of course, we urge both caution and further research before relying on our experimental estimates to predict responses to actual tax reforms. As with many other heuristics, we suspect that ironing might be most prevalent for "quick" decisions; e.g., whether to work an extra shift or to make a tax-deductible charitable contribution. Moreover, to the extent that heuristic use is a deliberate means of reducing cognitive costs, it may be less frequent in high stakes labor-supply decisions than it is in our studies. At the same time, a countervailing force is that in practice, the decisions that rely on correct forecasts of tax rates involve many other dimensions that all require careful consideration; this additional complexity may leave little mental bandwidth for tax forecasting. For example, an individual choosing between two different jobs may, rightly or wrongly, be less concerned with correctly considering after-tax salaries than with workplace culture or livability of the different cities.

While further study of the elasticity of ironing propensity with respect to stakes is needed, some insights may be gleaned from research on analogous research on other heuristics. For example, while mental accounting, narrow bracketing, and the representativeness heuristic were originally documented in unincentivized survey studies, they have since been documented to play an important role in important field behaviors—e.g., in the marginal propensity to consume out of the supplemental nutrition assistance program in the U.S. (Hastings and Shapiro, 2018), take up of long-term care insurance (Gottlieb and Mitchell, 2015), and the link between analysts' earnings-growth forecasts and stock returns (Bordalo et al., forthcoming). Camerer and Hogarth (1999) provide a systematic analysis of whether incentives affect heuristic use in experiments, and find no evidence that it does. While some reservation is needed when assuming that field behavior will be governed by these heuristics, this existing literature strongly supports the possibility.

We additionally caution readers that the population used in our studies is likely non-representative. Despite Study 1 matching the U.S. population on several key observable demographics, unobserved characteristics could influence selection into our online survey platform. While we have shown that the results from both studies are robust to a variety of sample restrictions, further work is merited to validate our results in further populations of interest. However, were heuristics and biases besides ironing present in the general population, they would be found in subsamples such as ours as well; as such, we do not view these issues as a hindrance to a demonstration of ironing's comparative importance.

We are less reserved about our qualitative findings than we are about our point estimates. In particular, we do not see a clear reason for why our findings about the relative unimportance of residual misperceptions beyond ironing should be reversed when stakes are increased. This suggests that our finding that ironing provides a parsimonious account of MTR underestimation is less sensitive to concerns about external validity.

These results build towards a reasonably comprehensive account of income-tax-schedule misunderstanding that is much needed in the behavioral public finance literature. As illustrated in Section 5, our results are easily integrated into frameworks such as those of Farhi and Gabaix (2018), and in combination with their theoretical advances may broadly inform tax policy design. We believe that this line of inquiry has thus far received too little attention from behavioral scientists, perhaps because of its intrinsically tight connection to core economic questions rather than to core psychological questions. Beyond applications to tax policy, this line of inquiry can inform broader questions about mechanism design and misperceptions of complex incentives.

While ironing appears to be economically important, its psychological foundations remain ill understood. However, the experimental framework we have developed in Study 2 may be readily adapted for deeper study of this topic, both for concrete questions about taxes, and also for questions about mechanism design and complex incentives more broadly. For example, the extent to which ironing propensity depends on stakes or on the presence of attention-demanding stimuli can shed light on whether ironing is more of a deliberately chosen heuristic or if it's more of a strongly held belief about the typical nature of incentives schemes in the real world. Variation in cues that stimulate attention to marginal tax rates, or to the number of brackets in the tax code, can similarly shed light on whether ironing is more automatic or more deliberate. It would also be interesting to observe how individuals behave when they make decisions such as those in Study 2 repeatedly, with limited or full feedback. Beyond schmeduling, Farhi and Gabaix (2018) conjecture that the top marginal tax rate is particularly salient, a hypothesis that could be addressed by incorporating exogenous variation in the top marginal tax rate in experimentally generated schedules. Finally, it will be useful to explore the extent to which these heuristics are applied to complex incentive schemes other than tax codes. Exploring these and related questions is an important step toward a more psychologically-grounded mechanism design; we view our studies as developing some foundations for this promising line of research.

References

Aghion, Philippe, Ufuk Akcigit, Matthieu Lequien, and Stefanie Stantcheva, "Tax Simplicity and Heterogeneous Learning," Working Paper 24049, National Bureau of Economic Research November 2017.

Alesina, Alberto, Stefanie Stantcheva, and Edoardo Teso, "Intergenerational Mobility and Preferences for Redistribution," *American Economic Review*, 2018, *108* (2), 521–554.

Altomonte, Carlo, Alessandro Barattieri, and Susan Basu, "Average-Cost Pricing: Some Evidence and Implications," *European Economic Review*, 2015, *79*, 281–296.

Atkinson, Anthony Barnes, "Public Economics and the Economic Public," *European Economic Review*, 1990, *34* (2-3), 225–248.

Ballard, Charles L. and Sanjay Gupta, "Perceptions and Realities of Average Tax Rates in the Federal Income Tax: Evidence from Michigan," *National Tax Journal*, 2018, *71* (2), 263–294.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot, "Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference," *American Economic Review*, 2014, *104* (9), 2698–2735.

Benzarti, Youssef, "How Taxing Is Tax Filing? Using Revealed Preferences to Estimate Compliance Costs.," Working Paper 23903, National Bureau of Economic Research October 2017.

Bernheim, B. Douglas and Dmitry Taubinsky, "Behavioral Public Economics," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics - Foundations and Applications 1*, North Holland, 2018, pp. 381–516.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor, "Choose to Lose: Health Plan Choices from a Menu with Dominated Options," *The Quarterly Journal of Economics*, 2017, *132* (3), 1319–1372.

Blaufus, Kay, Jonathan Bob, Jochen Hundsdoerfer, Christian Sielaff, Dirk Kiesewetter, and Joachim Weimann, "Perception of Income Tax Rates: Evidence from Germany," *European Journal of Law and Economics*, 2015, *40* (3), 457–478.

Blomquist, Soren and Luca Micheletto, "Optimal Redistributive Taxation when Government's and Agents' Preferences Differ," *Journal of Public Economics*, 2006, *90* (6-7), 1215–1233.

Bordalo, Pedro, Nicola Gennaioli, Rafael La Porta, and Andrei Shleifer, "Diagnostic Expectations and Stock Returns," *Journal of Finance*, forthcoming.

Brown, C. V., "Misconceptions About Income Tax and Incentives," *Scottish Journal of Political Economy*, 1969, *16* (2), 1–21.

Camerer, Colin F. and Robin M. Hogarth, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 1999, *18*, 7–42.

Carter, David W. and J. Walter Milon, "Price Knowledge in Household Demand for Utility Services," *Land Economics*, 2005, *81* (2), 265–283.

Chetty, Raj, "Is the Taxable Income Elasticity Sufficient to Calculate Deadweight Loss? The Implications of Evasion and Avoidance," *American Economic Journal: Economic Policy*, 2009, *1* (2), 31–52.

_ , Adam Guren, Day Manoli, and Andrea Weber, "Are Micro and Macro Labor Supply Elasticities Consistent? A Review of Evidence on the Intensive and Extensive Margins," *American Economic Review*, May 2011, *101* (3), 471–475.

_ , Adam Looney, and Kory Kroft, "Salience and Taxation: Theory and Evidence," *American Economic Review*, 2009, *99* (4), 1145–1177.

_ and Emmanuel Saez, "Teaching the Tax Code: Earnings Responses to an Experiment with EITC Recipients," *American Economic Journal: Applied Economics*, 2013, *5* (1), 1–31.

_ , John N Friedman, and Emmanuel Saez, "Using Differences in Knowledge across Neighborhoods to Uncover the Impacts of the EITC on Earnings," *American Economic Review*, 2013, *103* (7), 2683–2721.

Congdon, William J., Jeffrey R. Kling, and Sendhil Mullainathan, "Behavioral Economics and Tax Policy," *National Tax Journal*, 2009, *62* (3), 375–386.

Congressional Budget Office, "Refundable Tax Credits," 2013.

Cruces, Guillermo, Ricardo Perez-Truglia, and Martin Tetaz, "Biased Perceptions of Income Distribution and Preferences for Redistribution: Evidence from a Survey Experiment," *Journal of Public Economics*, 2013, *98*, 100 – 112.

Currie, Janet, "The Take-Up of Social Benefits," in Alan J. Auerbach, David Card, and John M. Quigley, eds., *Public Policy and the Income Distribution*, Russell Sage Foundation, 2006, pp. 80–148.

de Bartolome, Charles A. M., "Which Tax Rate Do People Use: Average or Marginal?," *Journal of Public Economics*, 1995, *56* (1), 79–96.

Diamond, Peter A, "Optimal Income Taxation: An Example with a U-shaped Pattern of Optimal Marginal Tax rates," *American Economic Review*, 1998, *88* (1), 83–95.

Enrick, Nobert Lloyd, "A Pilot Study of Income Tax Consciousness," *National Tax Journal*, 1963, *16* (2), 169–173.

_ , "A Further Study of Income Tax Consciousness," *National Tax Journal*, 1964, *17* (3), 319–321.

Esponda, Ignacio and Demian Pouzo, "Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Agents,," *Econometrica*, 2016, *84* (3), 1093–1130.

Farhi, Emmanuel and Xavier Gabaix, "Optimal Taxation with Behavioral Agents," Working Paper 21524, National Bureau of Economic Research September 2018.

Feldman, Naomi E., Peter Katuscak, and Laura Kawano, "Taxpayer Confusion: Evidence from the Child Tax Credit," *American Economic Review*, 2016, *106* (3), 807–835.

Feldstein, Martin, "The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1986 Tax Reform Act," *Journal of Political Economy*, 1995, *103* (3).

— , "Tax Avoidance and the Deadweight Loss of the Income Tax," *Review of Economics and Statistics*, 1999, *81* (4), 674–680.

Fiedler, Klaus and Peter Juslin, "Taking the Interface Between Mind and Environment Seriously," *Information Sampling and Adaptive Cognition*, 2006, pp. 3–29.

Finkelstein, Amy, "E-ZTAX: Tax Salience and Tax Rates," *The Quarterly Journal of Economics*, 2009, *124* (3), 969–1010.

Fujii, Edwin T and Clifford Hawley, "On the Accuracy of Tax Perceptions," *The Review of Economics and Statistics*, 1988, *70* (2), 344–347.

Gabaix, Xavier, "A Sparsity Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 2014, *129*, 1661–1710.

Gennaioli, Nicola, Yueran Ma, and Andrei Shleifer, "Expectations and Investment," *NBER Macroeconomics Annual*, 2016.

Gensemer, Bruce, Jane Lean, and William Neenan, "Awareness of Marginal Income Tax Rates Among High-Income Taxpayers," *National Tax Journal*, 1965, *18* (3), 258–267.

Gerritsen, Aart, "Optimal Taxation When People Do Not Maximize Well-Being," *Journal of Public Economics*, 2016, pp. –.

Gideon, Michael, "Do Individuals Perceive Income Tax Rates Correctly?," *Public Finance Review*, 2017, *45* (1), 97–117.

Gottlieb, Daniel and Olivia Mitchell, "Narrow Framing and Long-Term Care Insurance," *Working Paper*, 2015.

Grubb, Michael D. and Matthew Osborne, "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock," *American Economic Review*, 2015, *105* (1), 234–71.

Handel, Ben, Jon Kolstad, and Johannes Spinnewijn, "Information Frictions and Adverse Selection: Policy Interventions in Health Insurance Markets," *Review of Economics and Statistics*, forthcoming.

Handel, Benjamin R, "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *The American Economic Review*, 2013, *103* (7), 2643–2682.

— **and Jonathan T Kolstad**, "Health Insurance for "Humans": Information Frictions, Plan Choice, and Consumer Welfare," *American Economic Review*, 2015, *105* (8), 2449–2500.

Hastings, Justine and Jesse Shapiro, "How Are SNAP Benefits Spent? Evidence from a Retail Panel," *American Economic Review*, 2018, *108* (12), 3493–3540.

Ito, Koichiro, "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing," *American Economic Review*, 2014, *104* (2), 537–563.

Juslin, Peter, Anders Winman, and Patrik Hansson, "The Naïve Intuitive Statistician: A Naïve Sampling Model of Intuitive Confidence Intervals.," *Psychological Review*, 2007, *114* (3), 678.

Kanbur, Ravi, Jukka Perttila, and Matti Tuomala, "Non-Welfarist Optimal Taxation Behavioural Public Economics," *Journal of Economic Surveys*, 2006, *20* (5), 849–868.

_ **, Jukka Pirttila, and Matti Tuomala**, "Moral Hazard, Income Taxation, and Prospect Theory," *Scandinavian Journal of Economics*, 2008, *110* (2), 321–337.

Kleinberg, Jon, Annie Liang, and Sendhil Mullainathan, "The Theory is Predictive, but is it Complete? An Application to Human Perception of Randomness," *Working Paper*, 2017.

Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva, "How Elastic are Preferences for Redistribution? Evidence from Randomized Survey Experiments," *American Economic Review*, 2015, *105* (4), 1478–1508.

Levvy, Matthew and Joshua Tasoff, "Exponential Growth Bias and Lifecycle Consumption," *Journal of the European Economic Association*, 2016, *14* (3), 545–583.

Lewis, Alan, "Perceptions of Tax Rates," *British Tax Review*, 1978, pp. 358–366.

Liebman, Jeffrey B., "The Impact of the Earned Income Tax Credit on Incentives and Income Distribution," in "Tax Policy and the Economy, Volume 12" NBER Chapters, National Bureau of Economic Research, Inc, June 1998, pp. 83–120.

_ **and Richard Zeckhauser**, "Schmeduling," *Working Paper*, 2004.

Lusardi, Annamaria and Olivia S. Mitchell, "The Economic Importance of Financial Literacy: Theory and Evidence," *Journal of Economic Literature*, 2014, *52* (1), 5–44.

Manski, Charles F., "Measuring Expectations," *Econometrica*, 2004, *72* (5), 1329–1376.

Miller, Benjamin and Kevin Mumford, "The Salience of Complex Tax Changes: Evidence From the Child and Dependent Care Credit Expansion," *National Tax Journal*, 2015, *68* (3), 477–510.

Mirrlees, James A, "An Exploration in the Theory of Optimum Income Taxation," *The Review of Economic Studies*, 1971, pp. 175–208.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J Congdon, "A Reduced-Form Approach to Behavioral Public Finance," *Annual Review of Economics*, 2012, *4*, 1–30.

Rees-Jones, Alex and Dmitry Taubinsky, "Taxing Humans: Pitfalls of the Mechanism Design Approach and Potential Resolutions," in Robert A. Moffitt, ed., *Tax Policy and the Economy, Volume 32*, University of Chicago Press, 2018.

Romich, Jennifer L., "Difficult Calculations: Low-Income Workers and Marginal Tax Rates," *Social Service Review*, 2006, *80* (1), 27–66.

_ **and Thomas Weisner**, "How Families View and Use the EITC: Advance Payment versus Lump Sum Delivery," *National Tax Journal*, 2000, *53* (4), 1245–1266.

Saez, Emmanuel, "Using Elasticities to Derive Optimal Income Tax Rates," *The Review of Economic Studies*, 2001, *68* (1), 205–229.

_ , "The Desirability of Commodity Taxation Under Non-Linear Income Taxation and Heterogeneous Tastes," *Journal of Public Economics*, 2002, *83* (2), 217–230.

_ , "Optimal Income Transfer Programs: Intensive versus Extensive Labor Supply Responses," *The Quarterly Journal of Economics*, 2002, *117* (3), 1039–1073.

Slemrod, Joel and Wojciech Kopczuk, "The Optimal Elasticity of Taxable Income," *Journal of Public Economics*, 2002, *84* (1), 91–112.

Spencer, J, "Estimating Averages," *Ergonomics*, 1961, *4* (4), 317–328.

—, "A Further Study of Estimating Averages," *Ergonomics*, 1963, *6* (3), 255–265.

Spinnewijn, Johannes, "Heterogeneity, Demand for Insurance and Adverse Selection," *American Economic Journal: Economic Policy*, 2017, *9* (1), 308–343.

Stango, Victor and Jonathan Zinman, "Exponential Growth Bias and Household Finance," *The Journal of Finance*, 2009, *64* (6), 2807–2849.

Taubinsky, Dmitry and Alex Rees-Jones, "Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment," *Review of Economic Studies*, 2018, *85* (4), 2462–2496.

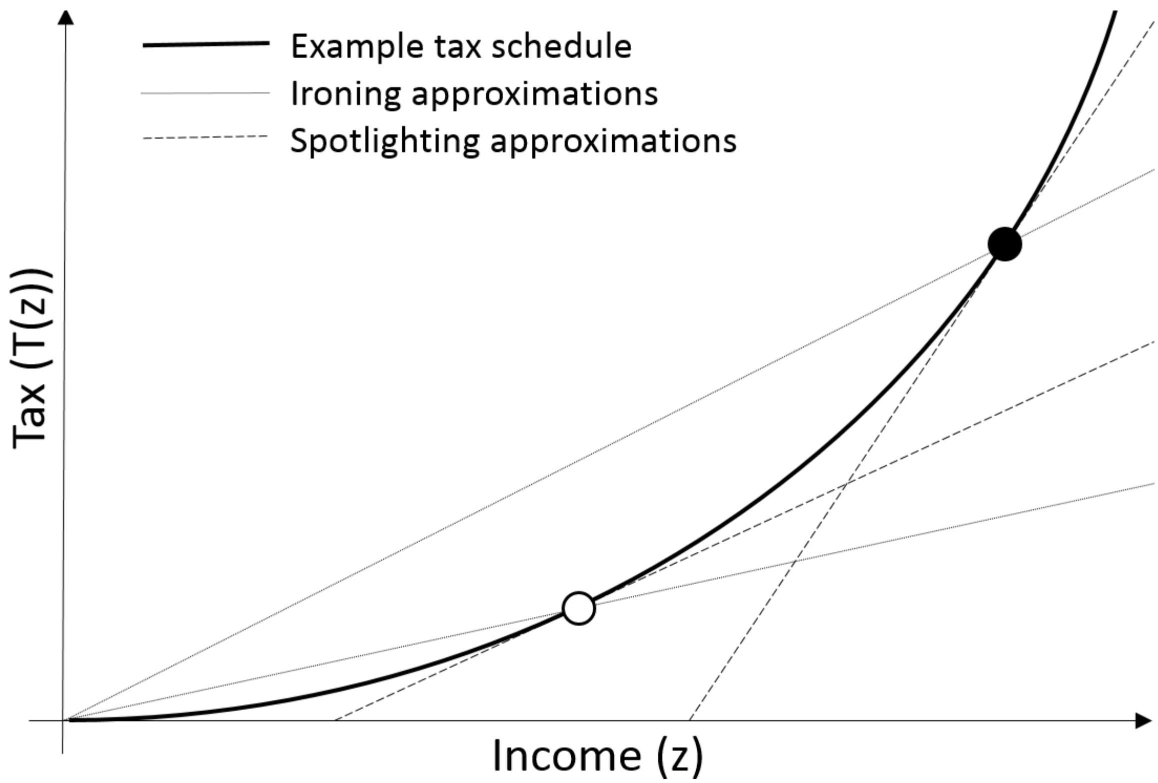
Villas-Boas, J. Miguel, "Comparative Statics of Fixed Points," *Journal of Economic Theory*, 1997, *73*, 183–198.

Wagstaff, Joseph Van, "Income Tax Consciousness under Withholding," *Southern Economic Journal*, 1965, *32* (1), 73–80.

Weinzierl, Matthew C, "The Promise of Positive Optimal Taxation," *Journal of Public Economics*, 2014, *118*, 128–142.

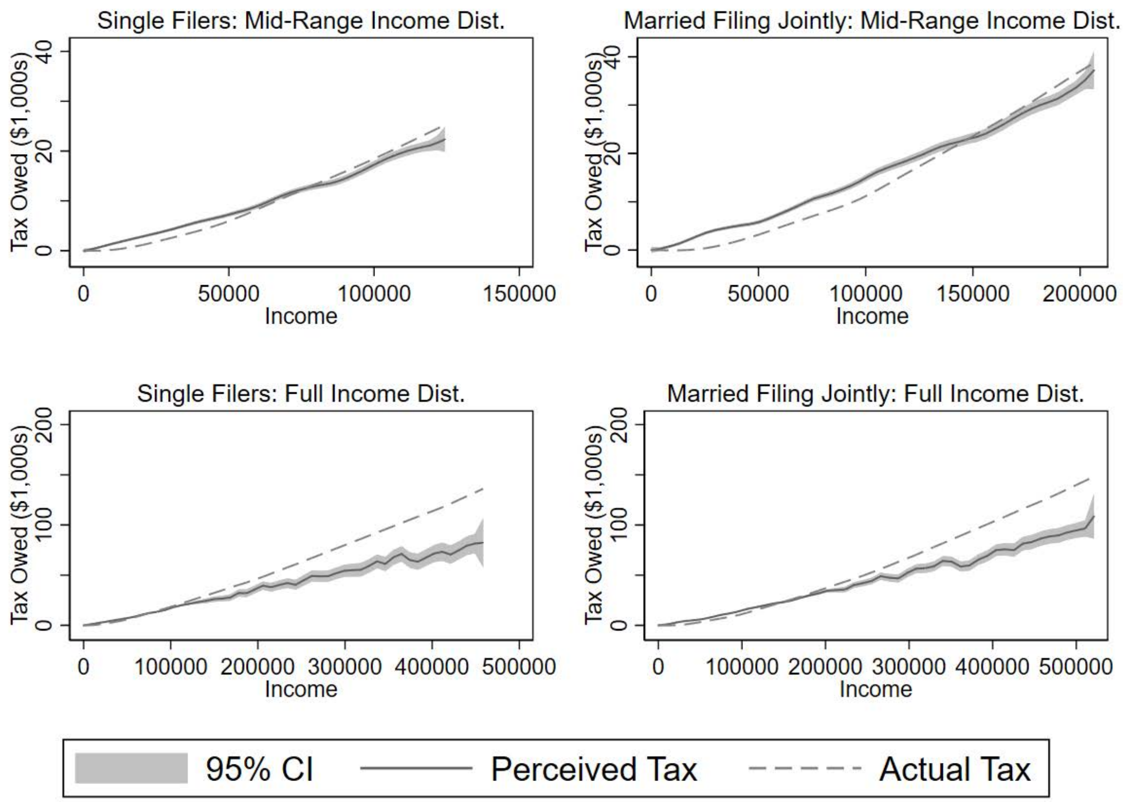
Weinzierl, Matthew C., "Popular Acceptance of Inequality Due to Innate Brute Luck and Support for Classical Benefit-based Taxation," *Journal of Public Economics*, 2017, *155*, 54–63.

Figure 1: Ironing and Spotlighting Heuristics



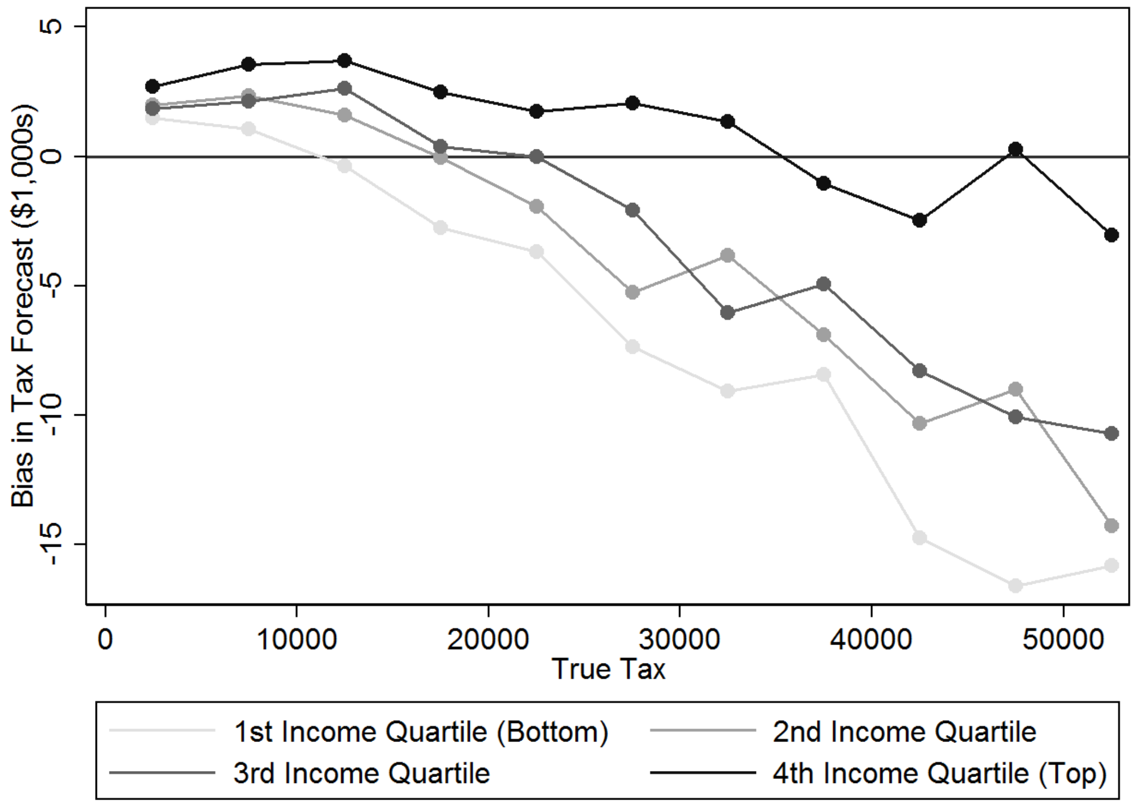
Notes: This figure presents an illustration of the ironing and spotlighting heuristics applied to a convex tax schedule. Taxpayers applying these heuristics approximate the schedule with linear forecasts that depend on their own position. We present two example positions, one with high income (the black dot) and one with comparatively low income (the white dot). Under the ironing heuristic, the taxpayer forecasts by applying his average tax rate at all points, resulting in the observed secant lines. Under the spotlighting heuristic, the taxpayer forecasts by applying his marginal tax rate to the change in income that would occur, resulting in the observed tangent lines.

Figure 2: Local-Polynomial Approximations of the Perceived Tax Schedule



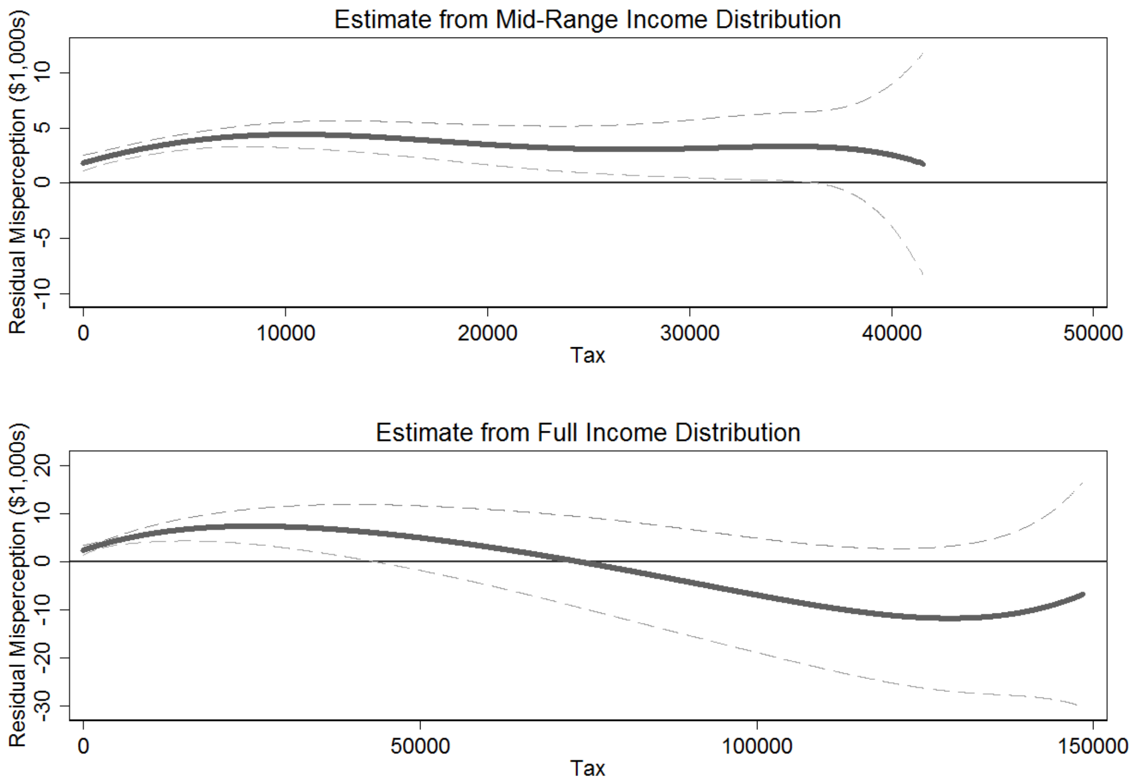
Notes: This figure presents local-polynomial approximations of the perceived relationship between the income earned and taxes owed. Results are plotted separately for single and married-filing-jointly tax filers, as incomes considered in the forecasting task were drawn from filing-status-specific distributions. The first row of figures presents estimates derived from only mid-range forecasts, while the second row presents estimates derived from the full sampling distribution. The shaded regions illustrate the 95% confidence intervals of the local-polynomial estimates. Bandwidth: 10,000. Degree of polynomial: 2. Kernel: Epanechnikov.

Figure 3: Bias in Tax Perceptions, by Respondents' Income Quartile



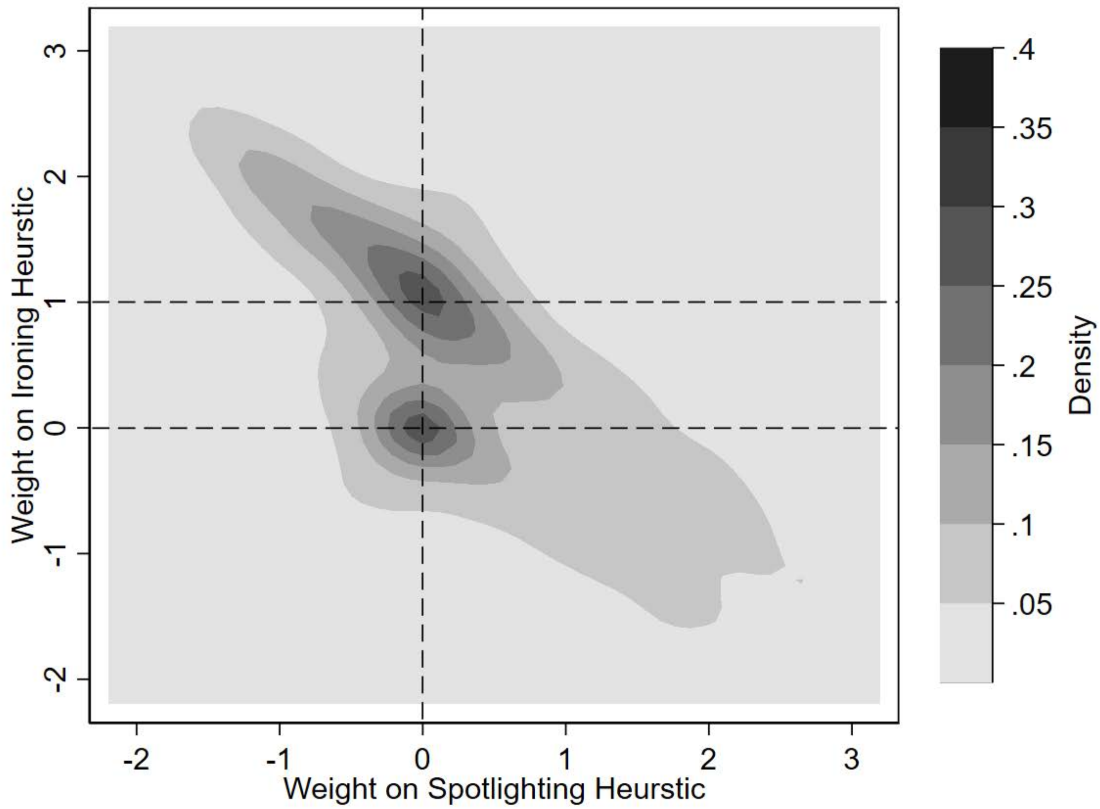
Notes: This figure plots the average bias in tax forecasts as a function of the true tax owed by the hypothetical tax payer. To explore how misperceptions of the tax schedule vary depending on the forecasters' own income, we plot this relationship separately by the income quartile of the respondent. Presented are the estimated coefficients from the regression $(\tilde{T} - T)_{i,f} = \sum_{b,q} \alpha_{b,q} * I(\text{income}_f \in \text{bin}_b) * I(\text{income}_i \in \text{quartile}_q) + \epsilon_{i,f}$, predicting average bias conditional on income quartile and the true tax owed, rounded into $5,000 bins.

Figure 4: Estimates of Residual Tax Misperception



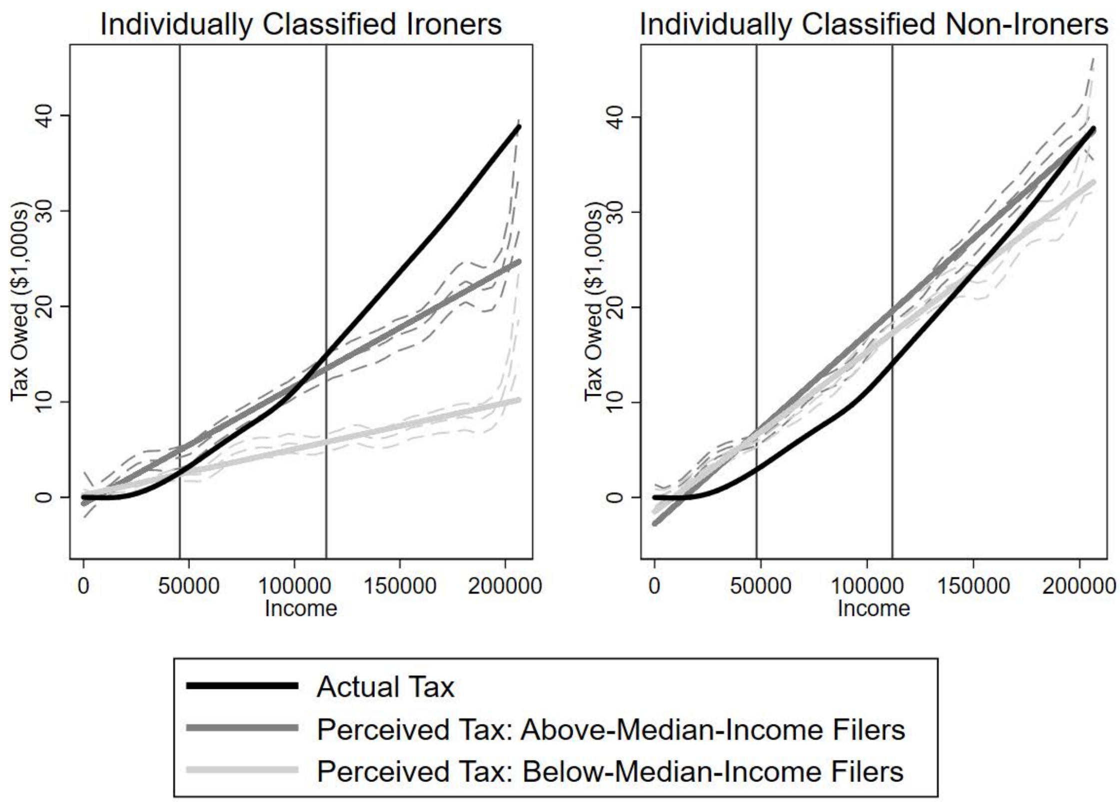
Notes: This figure plots the residual misperception functions estimated in columns 2 and 4 of Table 2. Dashed lines indicate the boundaries of 95% confidence intervals. These estimates indicate systematic overestimation of the taxes due when true taxes are comparatively small. For sufficiently large tax liabilities, this bias reverses into systematic underestimation of the taxes due.

Figure 5: Individual-Specific Estimates of Heuristic Propensity



Notes: This table presents a kernel-density estimate of the joint distribution of individual-specific ironing and spotlighting parameters, as estimated in the exercise described in section 3.3.2. Note that individual-level NLLS regressions failed to converge for 7 respondents. Bandwidth: .2. Kernel: Gaussian.

Figure 6: Examining Perceived Tax Schedules by Individual Classification



Notes: This figure presents approximations of the perceived relationship between the income earned and taxes owed, as previously shown in Figure 2, plotted by classification of "ironers" and "non-ironers" from Figure 5. Results shown are for married-filing-jointly tax filers. "Ironers" are classified as individuals with an ironing parameter within 0.4 of 1 and spotlighting parameter within 0.4 of 0. For each panel, we separately plot the perceived tax schedule for above and below median filers (conditional on having non-zero tax), with the vertical lines indicating the average income within each group. Dashed lines indicate the local polynomial fits with 95% confidence intervals (Bandwidth: 10,000. Degree of polynomial: 2. Kernel: Epanechnikov), and solid lines indicate fitted linear models.

Figure 7: Screenshot of Tax Schedules in Study 2

As a reward for completing this study, you will receive a bonus. Your reward comes in two accounts, A and B. **We impose a tax on your earnings in account A, but not on your earnings in account B.** Your final bonus will be the sum of the after-tax amounts from the two accounts. You will face choices on the next screen about how much we will put in each account. The tax that applies to account A is as follows:

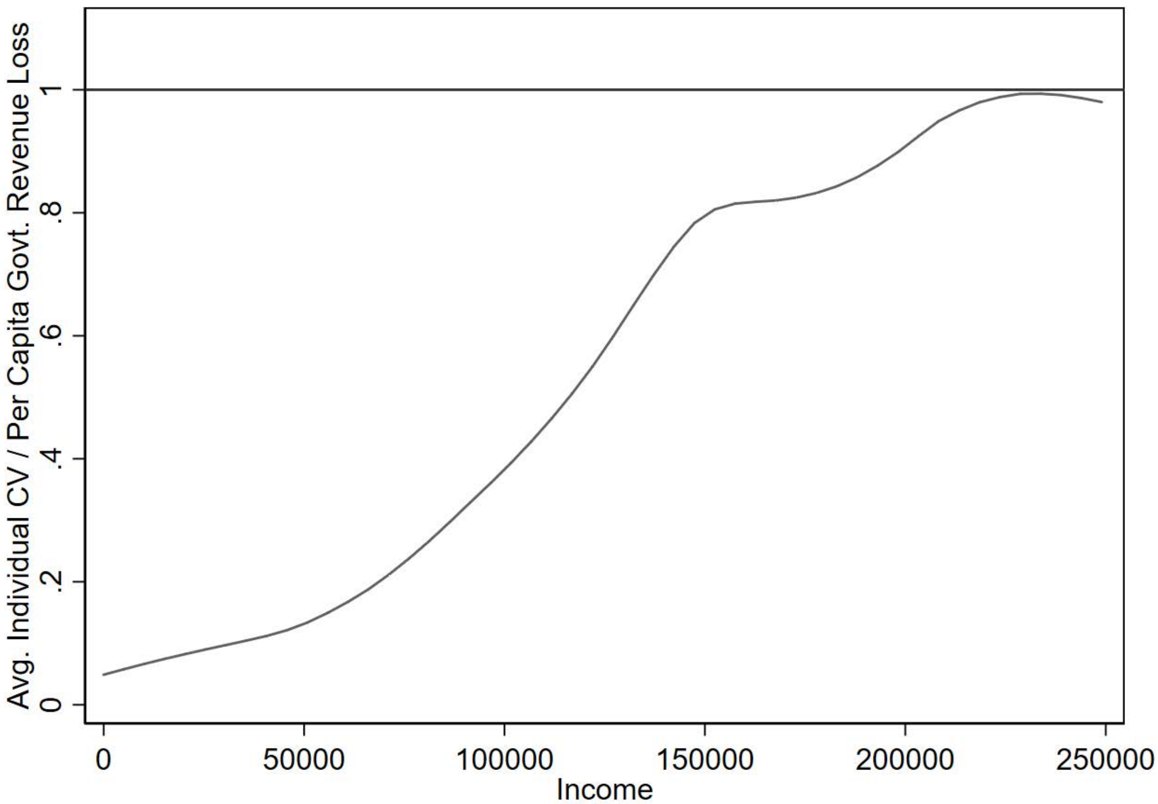
Taxes in Account A

# cents in A	10	20	30	40	50	60	70	80	90	100
Tax (in cents)	5	10	15	20	27	34	41	48	55	62

Once you have read these instructions, please answer the question below. You must answer this question correctly to proceed with this study.

Notes: Screenshot of the presentation of tax schedules in Study 2.

Figure 8: Consequences of Debiasing Across Income Distribution



Notes: This table presents local polynomial estimates that summarize the consequences of ironing for progressivity. For each taxpayer, we calculate the compensating variation in income that would lead an ironing taxpayer to have the same utility level as achieved by a correct forecaster. The y-axis reports this value as a fraction of the per-capita loss of government revenue that arises from correcting ironing. Bandwidth: 25000. Kernel: Epanechnikov.

Table 1: Testing for "Flattening" of the Tax Schedule

	All Incomes	Income Quartiles			
	Pooled	1	2	3	4
Estimation Sample: Local Draws					
Scale of slope (β)	0.81***	1.01***	1.07***	0.83***	0.78***
	(0.043)	(0.205)	(0.113)	(0.058)	(0.054)
P-value of H_0: $\beta = 1$	0.000	0.975	0.552	0.003	0.000
Respondents	4197	1050	1062	1040	1045
Forecasts	17937	3143	4074	5293	5427
Estimation Sample: Mid-Range Sampling Distribution					
Scale of slope (β)	0.82***	0.70***	0.78***	0.80***	0.94***
	(0.013)	(0.029)	(0.030)	(0.026)	(0.023)
P-value of H_0: $\beta = 1$	0.000	0.000	0.000	0.000	0.005
Respondents	4197	1050	1062	1040	1045
Forecasts	41970	10500	10620	10400	10450
Estimation Sample: Full Sampling Distribution					
Scale of slope (β)	0.62***	0.53***	0.56***	0.61***	0.76***
	(0.010)	(0.020)	(0.021)	(0.019)	(0.017)
P-value of H_0: $\beta = 1$	0.000	0.000	0.000	0.000	0.000
Respondents	4197	1050	1062	1040	1045
Forecasts	58758	14700	14868	14560	14630

Notes: Standard errors, clustered by respondent, in parentheses. Presented are coefficients from OLS fixed-effect regressions of the form $\tilde{T}_{i,f} = \beta * T_{i,f} + \nu_i + \epsilon_{i,f}$. The coefficient β can be interpreted as a scaling of the slope induced by the true tax schedule, where estimated values less than 1 indicate a "flattening" of the tax schedule. Two-sided Wald-test p-values, testing the hypothesis that $\beta = 1$, are presented below each regression. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 2: Parameter Estimates of Heuristic-Perception Model

	(1)	(2)	(3)	(4)
γ_I: weight on ironing forecast	0.21***	0.29***	0.47***	0.43***
	(0.037)	(0.052)	(0.048)	(0.095)
γ_S: weight on spotlighting forecast	-0.09*	-0.02	-0.03	-0.02
	(0.050)	(0.057)	(0.062)	(0.076)
Residual misperception function included	No	Yes	No	Yes
Income sampling distribution	Mid	Mid	Full	Full
Respondents	4197	4197	4197	4197
Forecasts	41970	41970	58758	58758

Notes: Standard errors, clustered by respondent, in parentheses. Presented are non-linear least squares estimates of ironing and spotlighting propensity. The estimated residual misperception function from columns 2 and 4 is plotted in Figure 4. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 3: Assessing Models' Fit Of MTR Perceptions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Model-Free	Preferred Specification			Out-of-Sample Ests.		
Ironing		X	X	X	X	X	X
Resid. misperceptions		X	X		X	X	
Spotlighting		X			X		
Scale of slope (β)	0.81	0.86	0.86	0.81	0.87	0.86	0.80
	(0.044)	(0.020)	(0.024)	(0.040)	(0.019)	(0.025)	(0.042)
Diff from model-free β		0.056	0.055	0.008	0.062	0.057	0.006
P-value of H_0: diff = 0		0.149	0.189	0.901	0.124	0.187	0.928

Notes: This table compares the the degree of MTR underestimation found empirically to that which would arise from our preferred specification of our empirical model. As a model-free baseline for comparison, column 1 reports the scaling parameter estimated in the top panel of table 1. Columns 2-4 present estimates of the scaling parameter predicted to arise under our estimated mixture model, progressively eliminating components of the model across columns. Columns 5-7 conduct an analogous exercise, but exclude local draws from the data used to estimate the forecasting model, and tests the ability of non-local forecasts to inform predictions of local tax understanding. Bootstrapped standard errors, resampled by subject with 1000 iterations, in parentheses.

Table 4: Ironing Propensity in Study 2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		By Complexity			Re-examined Tax Table		
	Full Sample	Simple	Complex	Diff.	Yes	No	Diff.
ATR	0.35***	0.34***	0.36***	-0.02	0.40***	0.31***	0.09*
Coefficients	(0.02)	(0.03)	(0.03)	(0.05)	(0.03)	(0.03)	(0.05)
MTR	0.09***	0.07**	0.12***	-0.05	0.09**	0.10***	-0.01
Coefficients	(0.02)	(0.03)	(0.03)	(0.05)	(0.03)	(0.03)	(0.05)
Constant	0.28***	0.30***	0.26***	0.04	0.25***	0.31***	-0.06**
	(0.01)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)
p: LR Test				0.164			0.007
Fraction	0.79***	0.83***	0.75***		0.82***	0.76***	
Ironing	(0.05)	(0.07)	(0.06)		(0.06)	(0.07)	
Fraction	0.21***	0.17**	0.25***		0.18***	0.24***	
using MTR	(0.05)	(0.07)	(0.06)		(0.06)	(0.07)	
N	3,130	1,571	1,559		1,431	1,699	

Notes: Standard errors in parentheses. Presented are OLS estimates of impact of random variation in ATR and MTR on perceived marginal tax rates. Column 1 presents our primary estimates. Columns 2 and 3 present estimates when the data are restricted to simple or complex schedules, with column 4 presenting the estimated differences in a fully interacted model. Columns 5 and 6 present estimates when the data are restricted to subjects who did or did not re-examine the tax table when given the opportunity, with column 7 presenting the estimated differences in a fully interacted model. Likelihood ratio tests of the fully interacted models, compared against the non-interacted models, are presented below the estimates. In the lower panel, we classify the fraction of "responsive" subjects (defined as those reacting to either the MTR or the ATR) who appear to be ironing versus using the MTR. Fractions are calculating by dividing the relevant coefficient by the sum of the two coefficients, with standard errors calculated with the delta method. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: Revenue and Welfare Effects of Ironing

Structural Elasticity ($\frac{1}{k}$)	Increase in Tax Rev. (%)	*Net Welfare Increase (%)* Low λ $\lambda = U'_{50}$	— $\lambda = \bar{U}'$	High λ $\lambda = U'_{90}$
1/2	3.7	3.2	3.4	3.6
	[2.1, 5.3]	[1.8, 4.6]	[2.0, 4.9]	[2.1, 5.1]
1/3	2.5	2.2	2.3	2.5
	[1.5, 3.7]	[1.3, 3.2]	[1.4, 3.4]	[1.4, 3.6]
1/4	1.9	1.7	1.8	1.9
	[1.1, 2.8]	[1.0, 2.4]	[1.0, 2.6]	[1.1, 2.7]
1/5	1.6	1.4	1.4	1.5
	[0.9, 2.3]	[0.8, 2.0]	[0.8, 2.1]	[0.9, 2.2]

Notes: The numbers presented contrast the revenue collected or welfare attained when comparing a population with perfect tax perceptions against one in which some portion of filers apply the ironing heuristic. The primary numbers presented are generated assuming that 43% of the population irons, as in our preferred specification. The bracketed numbers underneath present the calculation derived under the assumption that ironing propensity is either 25% or 62%—the boundaries of the 95% confidence interval of our preferred propensity estimate. The first column presents the structural elasticity in our assumed utility model: $U(z) = log(z - T(z) - \frac{(z/w)^{1+k}}{1+k})$. The second column presents the additional government revenue collected when the ironers are present. The final three columns present estimates of the increase in social welfare attained due to the presence of ironers, under alternative assumptions on the cost of public funds. Welfare effects are expressed as the percentage of total tax revenues that a social planner would pay to avoid converting all ironers to correct forecasters.

Table 6: Revenue and Welfare Effects Changing to Flat Tax

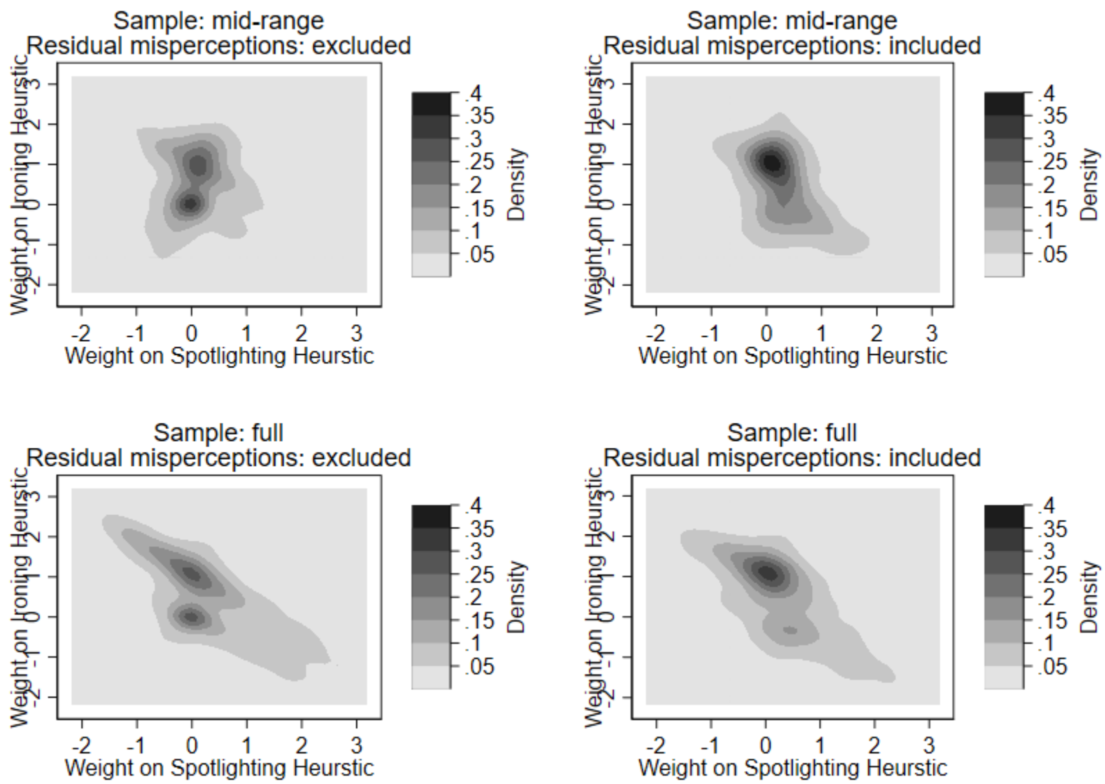
Structural	*All correct forecasters*		*With ironers*	
Elasticity	Δ Tax Rev.	Δ Welfare	Δ Tax Rev.	Δ Welfare
($\frac{1}{k}$)	(%)	(%)	(%)	(%)
1/2	5.2	-9.9	2.9	-12.3
			[3.9, 2.0]	[-11.3, -13.4]
1/3	3.3	-11.6	1.9	-13.2
			[2.5, 1.2]	[-12.5, -13.9]
1/4	2.5	-12.4	1.4	-13.6
			[1.8, 0.9]	[-13.1, -14.1]
1/5	1.9	-12.9	1.1	-13.8
			[1.4, 0.7]	[-13.5, -14.3]

Notes: This table summarizes the revenue collected or welfare attained as a result replacing the progressive tax schedule with a linear schedule that would be revenue-neutral assuming no change in behavior. The first column presents the structural elasticity in our assumed utility model: $U(z) = log(z - T(z) - \frac{(z/w)^{1+k}}{1+k})$. The second and third columns present the additional government revenue and welfare, respectively, resulting from the tax-rate change under the assumption of perfect tax perceptions. The fourth and fifth columns provide analogous calculations under the assumption that 43% of the population irons, as in our preferred specification. The bracketed numbers underneath present the calculation derived under the assumption that ironing propensity is either 25% or 62%, respectively—the boundaries of the 95% confidence interval of our preferred propensity estimate. Welfare effects are expressed as the percentage of total tax revenues that a social planner would pay to avoid going to the flat tax.

Appendices (not for publication)

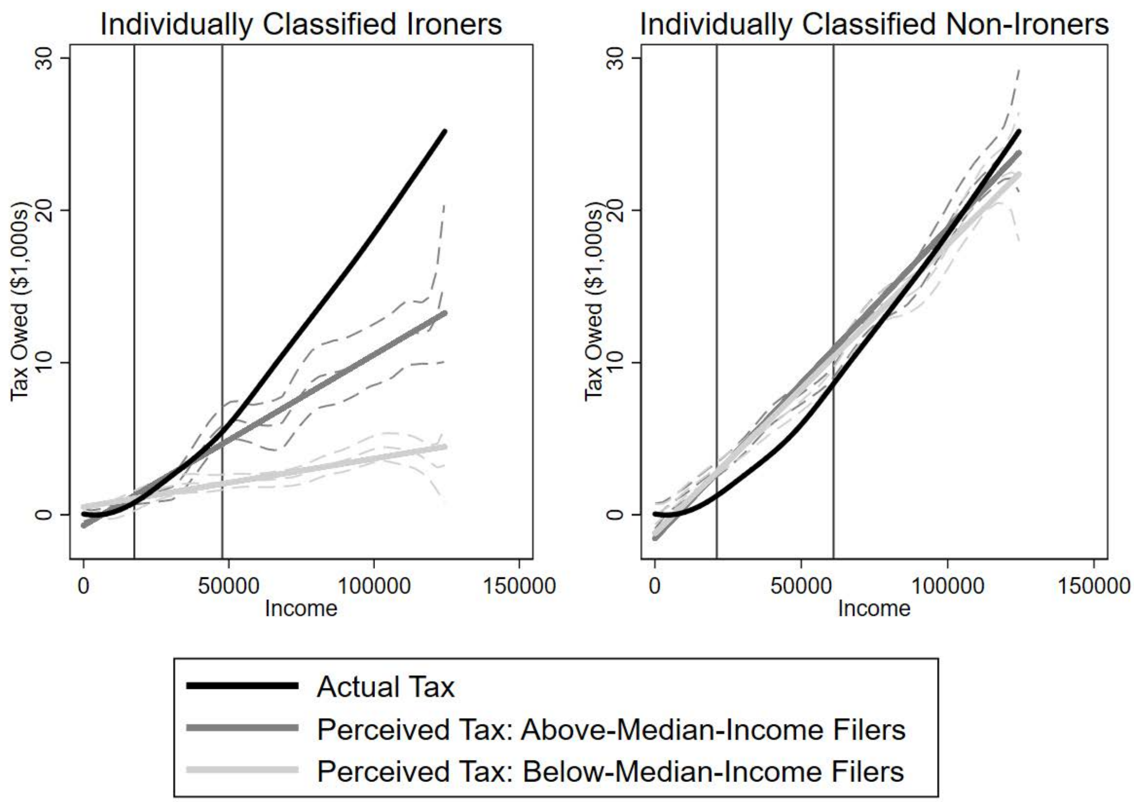
A Supplemental Figure and Tables

Figure A1: Alternative Versions of Figure 5



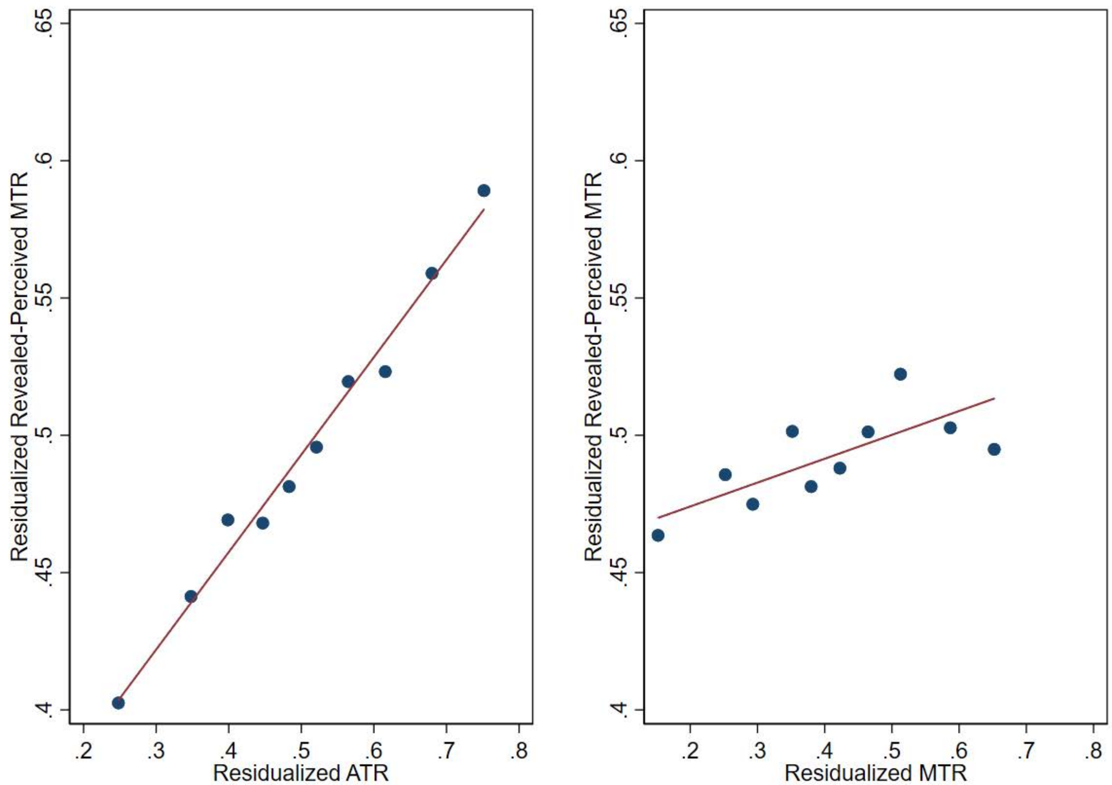
Notes: This figure plots alternative constructions of Figure 5, made to match the restrictions applied in each of the four columns of table 2. Note that individual-level NLLS regressions failed to converge for 13 respondents when using the mid-range sample, and for 7 respondents when using the full sample.

Figure A2: Examining Perceived Tax Schedules by Individual Classification: Single Filers



Notes: This figure reproduces figure 6, but uses data from single filers rather than married filing joingtly filers. Presented are approximations of the perceived relationship between the income earned and taxes owed, as previously shown in Figure 2, plotted by classification of "ironers" and "non-ironers" from Figure 5. "Ironers" are classified as individuals with an ironing parameter within 0.4 of 1 and spotlighting parameter within 0.4 of 0. For each panel, we separately plot the perceived tax schedule for above and below median filers (conditional on having non-zero tax), with the vertical lines indicating the average income within each group. Dashed lines indicate the local polynomial fits with 95% confidence intervals (Bandwidth: 10,000. Degree of polynomial: 2. Kernel: Epanechnikov), and solid lines indicate fitted linear models.

Figure A3: Conditional Linearity of Revealed-Perceived MTRs



Notes: This figure demonstrates the conditional linearity of revealed perceived MTRs elicited in Study 2, consistent with the predictions of our empirical model. Plotted are "binscatters" of the relationship between revealed-perceived MTR and the true ATR and MTR, respectively. In the left (right) panel, both the x- and y-axis variables are residualized by dummy variables for each discrete ATR (MTR) value. Plotted are the average values evaluated in each decile, and the best fit line.

Table A1: Findings Consistent with "Schmeduling" Predictions in Survey Literature

		Lewis (1978)	*Auld (1979)*	*Fujii & Hawley (1988)*	*Blaufus et al (2015)*
	Predictions:				
1	Taxes on low- vs high-income		I		I
2	Taxes on low- vs high-income, by own income				
3	Perceptions of MTRs	I		I	
4	Slope of tax schedule				
5	Slope of tax schedule, by own income				
	Sample Size	200	1,294	3,197	1,009
	Country	UK	Canada	USA	Germany

Notes: This table summarizes the available results relevant to predictions 1-5 in the existing tax misperception literature. A result consistent with ironing or spotlighting is indicated with an I or S, respectively.

Table A2: Demographics of Sample Compared to Census Data

	In-sample distribution	Census distribution
Gender		
Male	49%	49%
Female	51%	51%
Age		
18-44	39%	48%
45-64	44%	35%
65+	17%	17%
Income		
Under $15,000	16%	12%
$15,000 to $24,999	12%	10%
$25,000 to $34,999	11%	10%
$35,000 to $49,999	15%	13%
$50,000 to $74,999	19%	17%
$75,000 to $99,999	13%	12%
$100,000 to $149,999	10%	14%
$150,000 to $199,999	3%	6%
$200,000 +	1%	6%

Notes: This table presents tabulations of the gender, age, and income distributions reported in our sample for analysis, compared against the distributions reported in the census. Age distributions condition on being 18+.
Source: http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf and
https://www.census.gov/data/tables/2016/demo/income-poverty/p60-256.html.

Table A3: Table 2 with Modified Definition of Spotlighting

	(1)	(2)	(3)	(4)
γ_I: weight on ironing forecast	0.19***	0.31***	0.46***	0.44***
	(0.054)	(0.066)	(0.052)	(0.101)
γ_S: weight on spotlighting forecast	-0.07	-0.03	-0.03	-0.03
	(0.078)	(0.080)	(0.067)	(0.082)
Residual misperception function included	No	Yes	No	Yes
Income sampling distribution	Mid	Mid	Full	Full
Respondents	4197	4197	4197	4197
Forecasts	41970	41970	58758	58758

Notes: Standard errors, clustered by respondent, in parentheses. Presented are non-linear least squares estimates of ironing and spotlighting propensity, constructed as in Table 2. The sole difference from the analysis in Table 2 is a different coding of the spotlighting forecast. Rather than allowing the heuristic to predict negative tax liability for low incomes, we instead assume that a spotlighter would predict zero tax liability in such circumstances. As these results illustrate, our results are minimally affected by these differences in coding. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A4: Parameter Estimates of Heuristic-Perception Model: Alt. Degrees of Polynomial

	(1)	(2)	(3)	(4)
γ_I: weight on ironing forecast	0.28***	0.41***	0.29***	0.43***
	(0.052)	(0.094)	(0.052)	(0.095)
γ_S: weight on spotlighting forecast	0.01	0.01	-0.00	-0.01
	(0.055)	(0.075)	(0.056)	(0.076)
Degree of $r(t)$ polynomial	1	1	2	2
γ_I: weight on ironing forecast	0.29***	0.43***	0.29***	0.43***
	(0.052)	(0.095)	(0.052)	(0.095)
γ_S: weight on spotlighting forecast	-0.01	-0.02	-0.02	-0.02
	(0.057)	(0.076)	(0.057)	(0.076)
Degree of $r(t)$ polynomial	3	3	4	4
γ_I: weight on ironing forecast	0.29***	0.43***	0.30***	0.43***
	(0.052)	(0.095)	(0.052)	(0.095)
γ_S: weight on spotlighting forecast	-0.02	-0.02	-0.02	-0.02
	(0.057)	(0.076)	(0.057)	(0.076)
Degree of $r(t)$ polynomial	5	5	6	6
γ_I: weight on ironing forecast	0.30***	0.43***	0.30***	0.43***
	(0.052)	(0.095)	(0.052)	(0.095)
γ_S: weight on spotlighting forecast	-0.02	-0.03	-0.02	-0.02
	(0.057)	(0.076)	(0.057)	(0.076)
Degree of $r(t)$ polynomial	7	7	8	8
γ_I: weight on ironing forecast	0.30***	0.43***	0.30***	0.43***
	(0.052)	(0.095)	(0.052)	(0.095)
γ_S: weight on spotlighting forecast	-0.02	-0.03	-0.02	-0.03
	(0.057)	(0.076)	(0.057)	(0.076)
Degree of $r(t)$ polynomial	9	9	10	10
Income Sampling Distribution	Mid	Mid	Full	Full
Respondents	4197	4197	4197	4197
Forecasts	41970	58758	41970	58758

Notes: Standard errors, clustered by respondent, in parentheses. This table reproduces the estimates from columns 2 and 4 of table 2, while varying the degree of the polynomial used to approximate residual misperception. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A5: Classification of Individuals to Ironing Parameters (Table 2 column 1 analog)

Weight on Ironing Heuristic		**Weight on Spotlighting Heuristic**											
		0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Total
	0%	1407	55	35	23	20	22	16	19	17	31	41	1686
	10%	58	10	7	1	2	1	3	4	1	21	0	108
	20%	36	11	2	4	3	3	0	3	12	0	0	74
	30%	39	6	2	1	3	2	3	19	0	0	0	75
	40%	38	12	4	1	1	4	21	0	0	0	0	81
	50%	20	8	4	5	2	28	0	0	0	0	0	67
	60%	27	16	5	7	24	0	0	0	0	0	0	79
	70%	36	13	13	35	0	0	0	0	0	0	0	97
	80%	29	18	59	0	0	0	0	0	0	0	0	106
	90%	34	91	0	0	0	0	0	0	0	0	0	125
	100%	1054	0	0	0	0	0	0	0	0	0	0	1054
	Total	2778	240	131	77	55	60	43	45	30	52	41	3552

Notes: This table presents the distribution of individual-level classifications of heuristic-use parameters for all respondents with positive tax liability. For each respondent, we compared their 10 mid-range sample tax forecasts to the forecast of the model $\tilde{T_{f,i}} = (1 - \gamma_I - \gamma_S)T(z_{f,i}|\theta_i) + \gamma_I \tilde{T}_I(z_{f,i}|z_i^*, \theta_i) + \gamma_S \tilde{T}_S(z_{f,i}|z_i^*, \theta_i) + \epsilon_{f,i}$. We calculated this forecast for the grid of values of (γ_I, γ_S) indicated in the table above, and assigned each respondent to the parameter values which minimized the mean squared error of the difference.

Table A6: Classification of Individuals to Ironing Parameters (Table 2 column 2 analog)

		Weight on Spotlighting Heuristic											
		0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Total
Weight on Ironing Heuristic	0%	1101	84	72	70	77	56	42	26	14	24	28	1594
	10%	28	13	13	11	4	8	6	8	5	10	0	106
	20%	26	11	5	13	7	6	10	7	12	0	0	97
	30%	30	8	11	4	9	2	6	18	0	0	0	88
	40%	24	19	9	8	7	8	15	0	0	0	0	90
	50%	25	16	13	9	11	29	0	0	0	0	0	103
	60%	34	19	17	10	28	0	0	0	0	0	0	108
	70%	36	24	19	45	0	0	0	0	0	0	0	124
	80%	39	21	68	0	0	0	0	0	0	0	0	128
	90%	49	88	0	0	0	0	0	0	0	0	0	137
	100%	977	0	0	0	0	0	0	0	0	0	0	977
	Total	2369	303	227	170	143	109	79	59	31	34	28	3552

Notes: This table presents the distribution of individual-level classifications of heuristic-use parameters for all respondents with positive tax liability. For each respondent, we compared their 10 mid-range tax forecasts to the forecast of the model $\tilde{T_{f,i}} = (1-\gamma_I-\gamma_S)(T(z_{f,i}|\theta_i)+\hat{r}(T(z_{f,i}|\theta_i))) + \gamma_I\tilde{T}_I(z_{f,i}|z_i^*,\theta_i)+\gamma_S\tilde{T}_S(z_{f,i}|z_i^*,\theta_i)+\epsilon_{f,i}$, where $\hat{r}$ represents the fitted residual misperception function estimated in column 2 of table 2. We calculated this forecast for the grid of values of (γ_I,γ_S) indicated in the table above, and assigned each respondent to the parameter values which minimized the mean squared error of the difference.

Table A7: Classification of Individuals to Ironing Parameters

Weight on Ironing Heuristic	Weight on Spotlighting Heuristic											
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Total
0%	1155	77	36	31	42	42	42	34	26	39	60	1584
10%	38	13	8	3	6	3	7	5	13	16	0	112
20%	24	8	4	4	3	4	2	2	22	0	0	73
30%	24	11	7	0	1	4	5	30	0	0	0	82
40%	20	12	5	4	3	1	34	0	0	0	0	79
50%	20	10	3	4	3	38	0	0	0	0	0	78
60%	20	11	9	4	40	0	0	0	0	0	0	84
70%	29	14	6	47	0	0	0	0	0	0	0	96
80%	20	12	64	0	0	0	0	0	0	0	0	96
90%	32	81	0	0	0	0	0	0	0	0	0	113
100%	1155	0	0	0	0	0	0	0	0	0	0	1155
Total	2537	249	142	97	98	92	90	71	61	55	60	3552

Notes: This table presents the distribution of individual-level classifications of heuristic-use parameters for all respondents with positive tax liability. For each respondent, we compared their 14 tax forecasts to the forecast of the model $\tilde{T_{f,i}} = (1 - \gamma_I - \gamma_S)T(z_{f,i}|\theta_i) + \gamma_I \tilde{T}_I(z_{f,i}|z_i^*, \theta_i) + \gamma_S \tilde{T}_S(z_{f,i}|z_i^*, \theta_i) + \epsilon_{f,i}$. We calculated this forecast for the grid of values of (γ_I, γ_S) indicated in the table above, and assigned each respondent to the parameter values which minimized the mean squared error of the difference.

Table A8: Classification of Individuals to Ironing Parameters (Table 2 column 4 analog)

	Weight on Spotlighting Heuristic											
Weight on Ironing Heuristic	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Total
0%	1118	97	47	33	49	27	35	20	23	15	25	1489
10%	29	16	11	3	9	6	5	5	11	10	0	105
20%	36	17	7	8	3	3	7	4	8	0	0	93
30%	32	9	6	3	1	2	3	20	0	0	0	76
40%	37	11	9	1	0	2	23	0	0	0	0	83
50%	26	12	10	9	7	17	0	0	0	0	0	81
60%	34	20	17	6	24	0	0	0	0	0	0	101
70%	39	27	18	25	0	0	0	0	0	0	0	109
80%	50	30	56	0	0	0	0	0	0	0	0	136
90%	56	77	0	0	0	0	0	0	0	0	0	133
100%	1146	0	0	0	0	0	0	0	0	0	0	1146
Total	2603	316	181	88	93	57	73	49	42	25	25	3552

Notes: This table presents the distribution of individual-level classifications of heuristic-use parameters for all respondents with positive tax liability. For each respondent, we compared their 14 tax forecasts to the forecast of the model $\tilde{T_{f,i}} = (1-\gamma_I-\gamma_S)(T(z_{f,i}|\theta_i)+\hat{r}(T(z_{f,i}|\theta_i)))+\gamma_I\tilde{T}_I(z_{f,i}|z_i^*,\theta_i)+\gamma_S\tilde{T}_S(z_{f,i}|z_i^*,\theta_i)+\epsilon_{f,i}$, where $\hat{r}$ represents the fitted residual misperception function estimated in column 4 of table 2. We calculated this forecast for the grid of values of (γ_I,γ_S) indicated in the table above, and assigned each respondent to the parameter values which minimized the mean squared error of the difference.

B Robustness Analyses for Study 1

In this Appendix we present additional results and robustness analyses associated with Study 1.

B.1 Persistence of Heuristic Use Among Subgroups

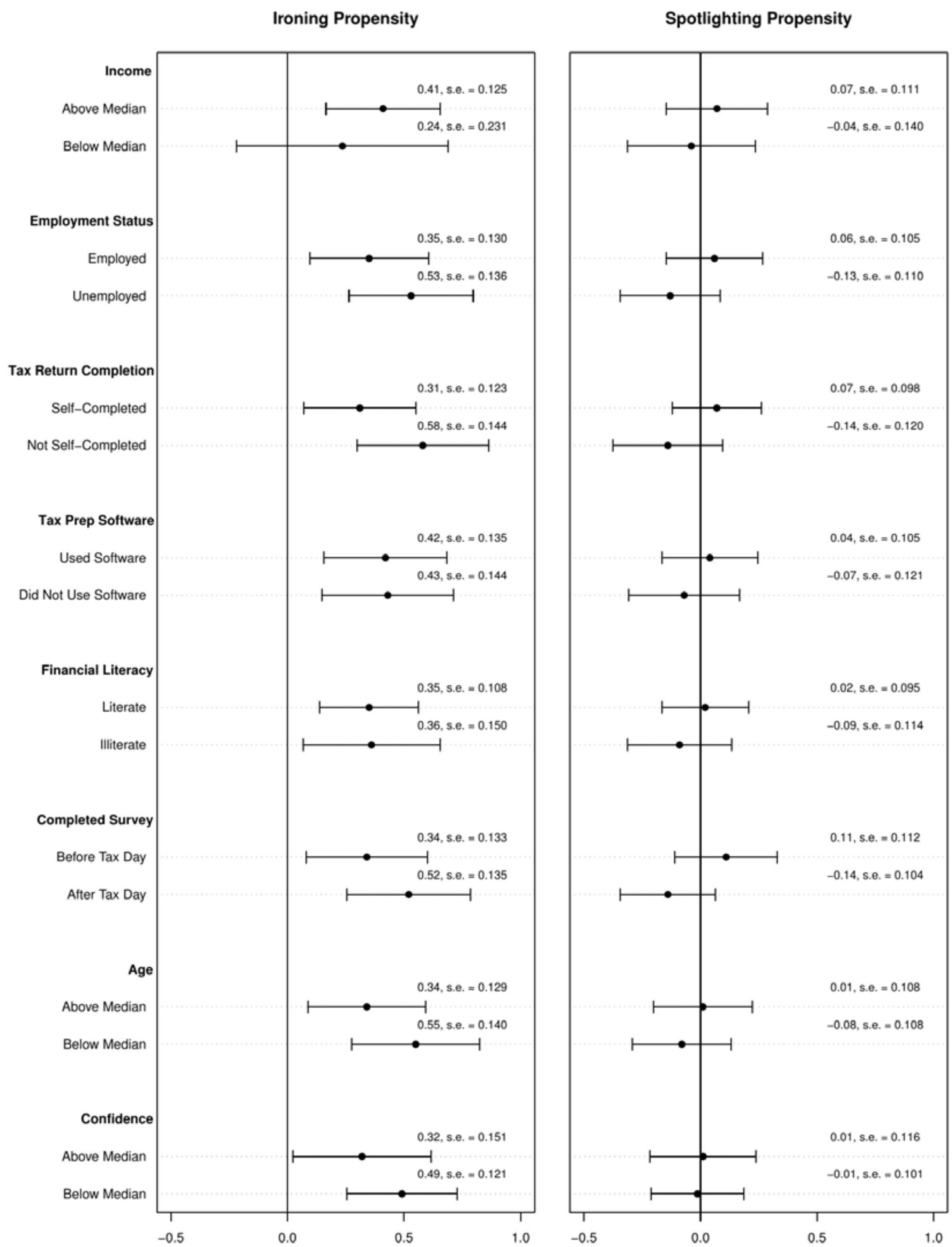
While tax knowledge is important to all people for, e.g., determining which policies and politicians to support or for budgeting spending, economic analyses often hinge on knowledge among specific groups. Groups of particular interest include the rich (in models of redistribution), workers (in models of labor-supply), or individuals completing their own tax returns (in models of compliance). Figure A4 presents estimates of ironing and spotlighting propensity created by applying our primary regression specification to various sample splits of interest. We continue to estimate prevalent ironing among both above- and below-median income respondents (41% vs 24%), the employed and the unemployed (35% vs 53%), those who completed their own tax return and those who did not (31% vs 58%), and those who use tax preparation software and those who do not (42% vs 43%). Furthermore, we find substantial prevalence of the ironing heuristic among both financially literate and financially illiterate tax filers, as classified by whether they do or do not correctly answer all of the "Big Three" financial literacy measures (35% vs 36%). We find that reliance on the ironing heuristic persists among those who completed the survey before or after tax day (34% vs 52%) and among both above- and below-median age respondents (34% vs 55%), suggesting that the misperceptions we document are neither temporarily eliminated by the experience of completing a tax return nor permanently eliminated by the cumulative experience with tax payments incurred over a lifetime. Finally, reliance on ironing persists among those with above- and below-median rates of indicating confidence in their given forecast (32% vs 49%).[42]

Across these sample splits, the propensity to iron is statistically significantly different from zero at least at the 5% α-level in all but one case.[43] The propensity to spotlight is statistically insignificant, evaluated at the 10% α-level, across all sample splits.

[42]Confidence in forecasts was elicited with the question "How confident are you that your answer is within $500 of the correct answer?" Available responses were "not confident at all," "somewhat confident," and "very confident." We conducted our median split by counting the number of forecasts for which the respondent indicated they were very confident. Note that the median respondent was very confident in zero of their forecasts. 40% of respondents were very confident in at least one forecast, and 5% were very confident in all 16 forecasts.

[43]The ironing propensity estimate of 0.24 among below-median income respondents has a clustered standard error of 0.231, generating an extremely large confidence interval including zero. This unusually large standard error is generated in this analysis due to multicolinearity: since average tax rates and marginal tax rates are nearly identical for low income filers, with their difference increasing in income on a convex tax schedule, the ironing and spotlighting predictions become highly correlated (ρ=0.91) if attention is restricted to low income respondents. The resulting correlation of the ironing and spotlighting forecasts significantly limits the statistical power of our approach.

Figure A4: Estimates of Heuristic Propensity: Robustness Analyses 1



Notes: The figure summarizes the estimated propensity of ironing and spotlighting across a variety of sample restrictions. The estimates presented correspond to our preferred specification (column 4 of table 2), but are estimated according to the sample definitions described in the left of the figure.

B.2 Inclusion of Other Taxes

In practice, the federal income tax is not the only tax on income; for most respondents, state taxes and FICA taxes also apply. Our experimental exercise specifically asked respondents to make forecasts about their federal income tax. However, a confused respondent could make forecasts that incorporate additional tax components. Since the inclusion of these extra taxes increases both the aggregate MTR and ATR, the presence of confusion of this sort would render our estimates of the degree of underestimation of the steepness of the tax schedule conservative. Thus, this confusion cannot account for our central reduce-form results. Moreover, this confusion could not account for our reduced-form evidence of ironing, since it would not explain why a respondent's estimate of Fred's tax liability is increasing in his own income.

In principle, such confusion could affect point estimates of ironing propensity. To examine the sensitivity of estimates to these concerns, we reestimate our primary heuristic model presented in Table 2 under three alternative assumptions: that the true tax, ATR, and MTR are all based on an aggregate tax schedule that additionally includes FICA tax, state tax, or both.[44]

Results are presented the top panel of Figure A5. We find that our conclusions regarding heuristic propensity are broadly similar across these alternative specifications. Estimated rates of ironing range from 37% to 55% across these specifications, whereas spotlighting is indistinguishable from zero (or marginally significantly negative in one case). The minimal influence of these alternative assumptions demonstrates an advantage of our empirical approach. The apparent misperception of tax amounts that would result from the contraindicated inclusion of additional taxes takes a form that can be approximated by the residual misperception function. Absent the presence of a residual misperception function, this type of confusion could be incorrectly attributed to heuristic forecasting. With a residual misperception function included, this class of forecasting errors is correctly classified as alternative phenomena, resulting in similar schmeduling propensity estimates.

B.3 Similarity of Actual and Hypothetical Tax Filers

Our experiment focused on a hypothetical taxpayer constructed to approximate the respondent. While the hypothetical taxpayer had the same filing status and number of exemptions as the respondent, he was built with intentionally simple taxable behavior: only wage income, and no additional schedules, credits, or deductions. This design element resolves an important difficulty present in other surveys of tax knowledge: uncertainty about the complete details shaping the respondents' own tax liability. While this design eliminates the measurement error inherent from that lack of knowledge, and thus allows us to incentivize experimental forecasts, it has one undesirable feature: respondents with filing behavior more complex than pure wage income are making forecasts regarding a tax schedule that imperfectly approximates their own. Our description of Fred precisely matches the returns submitted by 1,357 (32%) of our respondents, and the remaining 2,840 respondents have some element of their tax return—such as schedule B-F, an itemized deduction, or a claim to the EITC—that renders the approximation imperfect. In Figure A5, we conduct our main analysis restricted to each group of respondents. Both demonstrate substantial ironing (dissimilar filers: 34%; similar filers: 57%), and statistically insignificant spotlighting.

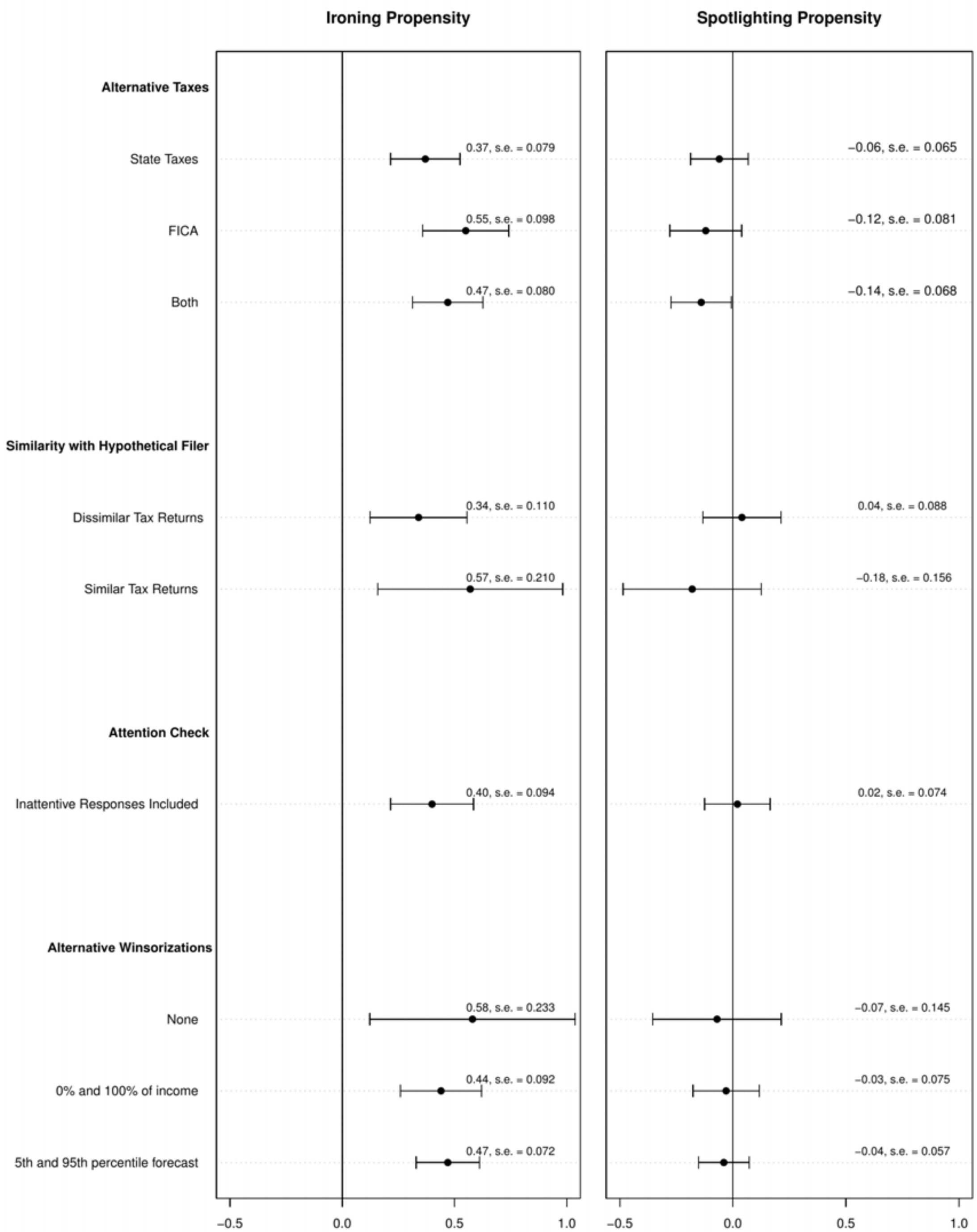
[44] We approximate state tax liability by applying the state's single or married-filing-jointly schedule to the federal adjusted gross income. Note that across states there are often small differences in the calculation of the tax base, which we necessarily abstract from due to data limitations. In analysis including state tax approximations, we exclude 34 respondents that we are unable to match to a state.

B.4 Importance of Data Restrictions

While most of our data restrictions described in section 3.2 are standard and affect few responses, two decisions may be contentious. First, note that we exclude 436 respondents (9% of our initial sample) who failed the attention check included in the miscellaneous questions module. As illustrated in Figure A5, reincluding these respondents has little effect on our estimated heuristic propensities. While this exclusion has little effect on the final results, we implement it as a matter of principle. Prior to running analyses, we worried that forecasts of respondents that do not carefully read instructions would necessarily be imperfect, and that the imperfection resulting from their inattention would not generate an externally valid measurement of the misperceptions of interest.

Second, we employ a Winsorization strategy as a means of controlling extreme forecasts. When deploying a unconstrained-response survey to thousands of respondents, at least a small number of wildly unreasonable forecasts are to be expected. To present an illustrative example, one respondent indicated that the tax due for an income of $823 is $96,321, when in fact it is zero. Even if most respondents have reasonably accurate tax perceptions, a small number of such extreme forecasts can significantly impact both parameter estimates and power. Furthermore, we believe the extremity of such forecasts does not approximate any externally valid forecasting problem, but rather is an indication of unusual confusion or experimental noncompliance. This motivated our choice to Winsorize tax forecasts at the 1st and 99th percentile forecasts within each $10,000 bin. As we demonstrate in Figure A5, alternative means of Winsorization have little impact on our quantitative estimates. Furthermore, our basic results persist even with the complete omission of outlier control, although estimates become notably less precise.

Figure A5: Estimates of Heuristic Propensity: Robustness Analyses 2



Notes: The figure summarizes the estimated propensity of ironing and spotlighting across a variety of sample restrictions. The estimates presented correspond to our preferred specification (column 4 of table 2), but are estimated according to the sample definitions described in the left of the figure.

C Robustness Analyses for Study 2

In this Appendix we present additional results and robustness analyses associated with Study 2.

C.1 Interval Regression

As discussed in section 4.1, our MPL identifies a narrow range of perceived marginal tax rates that would rationalize a given subjects choices. In our primary regressions, we mapped each interval to its midpoint and applied OLS. Alternatively, one could use techniques such as interval regression, which yield effectively identical results due to the fine partitioning that we adopted. Column 2 of Appendix Table A9 presents point estimates from this approach, yielding extremely similar results.

Table A9: Robustness Checks on Primary Specification in Study 2

	(1) OLS	(2) Intreg	(3) OLS	(4) OLS	(5) OLS
ATR	0.350***	0.350***	0.336***	0.333***	0.308***
	(0.0236)	(0.0237)	(0.0225)	(0.0288)	(0.0274)
MTR	0.0934***	0.0934***	0.0855***	0.0850***	0.0879***
	(0.0238)	(0.0239)	(0.0226)	(0.0288)	(0.0276)
Constant	0.281***	0.281***	0.290***	0.348***	0.362***
	(0.0126)	(0.0126)	(0.0119)	(0.0154)	(0.0146)
Failed Responses Included					
MPL Attention Check			X		X
Final Attention Check				X	X
N	3130	3130	3603	3689	4314

Notes: Standard errors in parentheses. This table reproduces the estimates from column 1 of table 4. In column 2, estimates are generated by applying interval regression instead of OLS. In columns 3-5, groups of inattentive respondents are re-included in the sample, as indicated by the lower panel. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

C.2 Reinclusion of Excluded Inattentive Respondents

We preregistered exclusion of respondents who failed to satisfy three basic criteria indicating good attention to, and understanding of, our experimental environment. These were: 1) excluding respondents with MPL responses inconsistent with well-behaved, monotone utility, 2) excluding respondents who chose the effectively dominated options at each end of the MPL list, and 3) excluding respondents who failed a simple attention check at the end of the survey. Group 1 cannot be re-included into our analysis in a very principled way, but groups 2 and 3 can be. Columns 3-5 of Appendix Table A9 estimate our primary specification with column 3 re-including group 2, column 4 re-including group 3, and column 5 re-including both groups. Overall, the impact on point estimates is relatively minimal, with both estimates becoming slightly smaller. The attenuation of point estimates is consistent with these respondents being confused and not reacting in any way to the tax schedule in front them.

C.3 Predictive Power of Alternative Models

Table A10: Improvements in Predictive Power from Alternative Models

Interactions in Model				
None	X			
ATR x MTR		X		
ATR x Complexity			X	
MTR x Complexity			X	
ATR x MTR x Complexity				X
Variables in Model				
ATR, MTR	0.0924	0.0954		
ATR, MTR, i.Complexity	0.0934	0.0963	0.0939	0.0980
i.ATR, i.MTR	0.0948	0.0989		
i.ATR, i.MTR, i.Complexity	0.0959	0.0998	0.0993	0.1055

Notes: This table presents the estimated R^2 arising from different versions of the primary analysis of Study 2. We consider models in which the ATR and MTR are included linearly as well as cases with an indicator variable for each discrete value (denoted i.ATR and i.MTR). We additionally vary whether we include an indicator variable for the complexity condition, as well as the presence of interactions between various subsets of included variables.

Table A10 summarizes the changes in the explanatory power of the model resulting from adding different degrees of non-linearity or interactions to the model. Recall from the text that the difference between the simple two-type model and the full schedule-specific-dummy model is not statistically detectable, and as such none of these differences are statistically significant. As this table shows, however, the differences are quantitatively insignificant, with the vast majority of explanatory power being achieved by the our parsimonious baseline.

C.4 Alternative Definition of Simple and Complex Schedules

Table A11: Robustness to exclusion of 0- or 3-kink schedules

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		By Complexity			Re-examined Tax Table		
	Full Sample	Simple	Complex	Diff.	Yes	No	Diff.
ATR	0.35***	0.35***	0.34***	0.01	0.41***	0.29***	0.11**
Coefficients	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
MTR	0.11***	0.08**	0.14***	-0.05	0.12***	0.10***	0.02
Coefficients	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)	(0.04)	(0.05)
Constant	0.28***	0.29***	0.27***	0.03	0.24***	0.32***	-0.08***
	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.03)
p: LR Test				0.526			0.015
N	2,454	1,239	1,215		1,142	1,312	

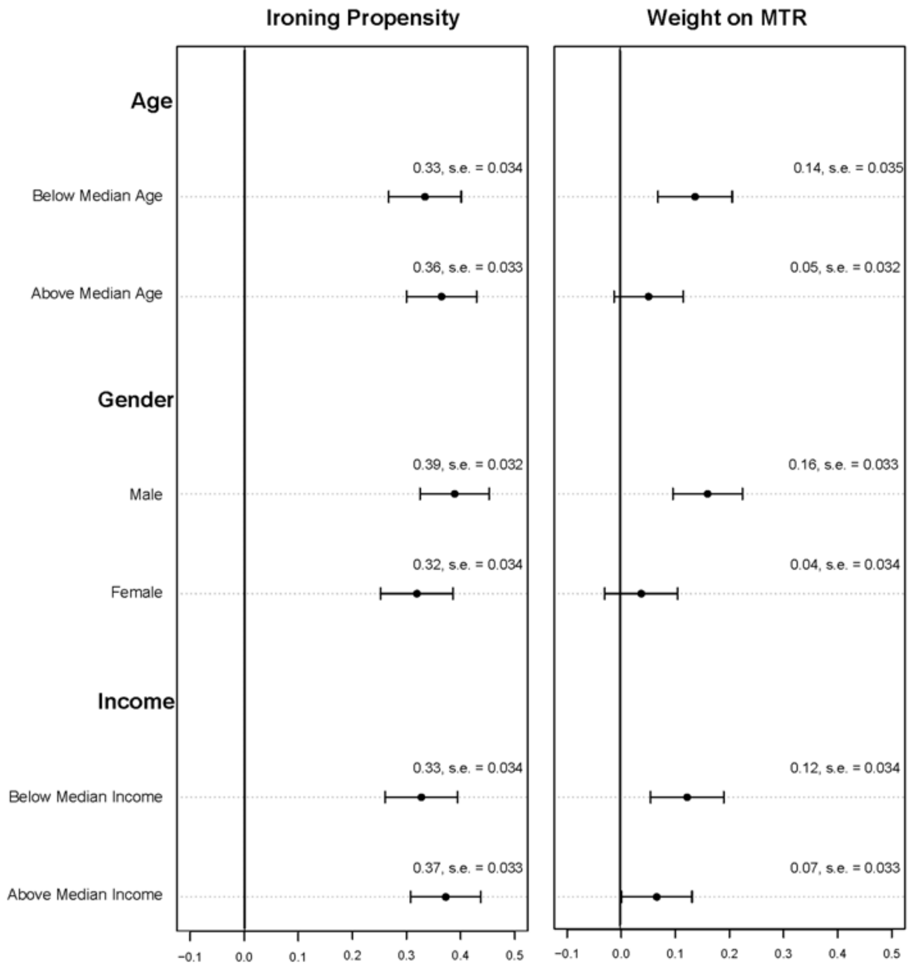
Notes: Standard errors in parentheses. This table reproduces the analysis of table 4, excluding data associated with schedules where the kink at 40 was "smoothed out." $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We designed our sampling scheme over schedules to generate exogenous variation in ATR and MTR in a schedule with either 2 or 5 brackets. However, based on the sampling scheme, at times the marginal rate in the bracket immediately above and immediately below the subject's 40 cent income were the same, effectively merging those two brackets. For simple schedules, this results in a linear tax. For complex schedules, this results in a schedule with 3 kinks rather than 4. In table A11, we recreate our main results restricting the sample to only true 1-kink or 4-kink schedules. This restriction proves to have no meaningful effect on our estimates.

A related feature of our sampling scheme offers an opportunity to study the degree of "effective debiasing" achieved in schedules that equate MTR and ATR. In our data, ATR=MTR for 22% of the schedules . When we regress revealed-percievd MTRs on true MTRs/ATRs among schedules for which the marginal and average tax rates are equal at 40 cents, the point estimate on tax rate on the incremental 20 cents is 0.41 (s.e. = 0.04). This is almost exactly equal to the sum of the coefficients on the ATR and MTR in Table 4, and highlights how far fewer individuals make a mistake when their average and marginal tax rates are equal. This illustrates a value of linear schedules: because they equate MTR and ATR, they guide ironers to behave optimally.

C.5 Differences in Behavior by Demographics

Figure A6: ATR and MTR Reliance by Demographics



Notes: Using data from Study 2, this figure summarizes the estimated propensity of MTR and ATR utilization across a variety of sample restrictions. All regressions correspond to the specification in column 1 of figure 4 with the sample restricted as indicated.

We explore differences in the estimated coefficients by subjects reported gender, age, and annual income. As illustrated in Appendix Figure A6, we find minimal and insignificant variation across these demographic groups in the propensity to rely on the ironing heuristic. In contrast, we find more meaningful differences in the weight placed on the MTR, with respondents of age greater than or equal to the median age of 35 placing less weight on the MTR (p-value of interaction=0.07) and with female respondents placing less weight on the MTR (p-value of interaction=0.01).

D Theory Appendix

We assume that utility takes the form $G(u(c) - \psi(z/w))$, where u is smooth, increasing and concave, ψ is smooth, strictly increasing and convex, and $\psi(0) = 0$ and $\lim_{z\to\infty} \psi'(z) = \infty$. We also assume that $-xu''(x)/u'(x) < 1$ to ensure that substitution effects dominate income effects; that is, so that an increase in a flat tax rate decreases the marginal benefits of consumption.

D.1 Existence and Uniqueness of the Solution Concept

Definition 3. Choice $z^*(w)$ is a *Ironing Equilibrium (IE)*, if

$$z^*(w) \in \text{argmax}\{U(z - \tilde{T}(z|z^*(w), \gamma), z/w)\}$$

where $\tilde{T}(z|z^*) = (1-\gamma)T(z^*) + \gamma z A(z^*)$ and $A(z^*)$ is the average tax rate at z^*.

Proposition 1. *Suppose that $T(z)$ is continuous. Then*

1. *There exists a IE $z^*(w)$.*
2. *z^* is continuous and increasing in w.*
3. *z^* is continuous and increasing in γ.*

Proof of Proposition 1 Assume that $G(x) = x$; which is without loss of generality since monotonic transformations of utility functions preserve behaviors.

Part 1. Let $B_{w,\gamma}(z')$ denote an optimal choice of z by an individual facing tax schedule $\tilde{T}(z|z')$. We first establish the following

1. $\tilde{T}(z|z')$ is convex for each z' because $T(z)$ is convex. Because u is concave and ψ is strictly convex, this means that $u(z - \tilde{T}(z|z')) - \psi(z/w)$ is strictly concave in z. Thus $B_w(z')$ is uniquely defined.
2. $B_{w,\gamma}(z')$ is continuous in z' because $u(z - \tilde{T}(z|z')) - \psi(z/w)$ is continuous in z' and is strictly concave in z.
3. $B_{w,\gamma}(z')$ is decreasing in z'. To show this, first note that

$$\frac{d}{dz}u(z - \tilde{T}(z|z')) = [1 - (1-\gamma)T'(z) - \gamma A(z')]u'(\cdot)$$

and

$$\begin{aligned}\frac{d}{dz'}\frac{d}{dz}u(z - \tilde{T}(z|z')) &= -\gamma A'(z')u'(\cdot) + [1 - (1-\gamma)T'(z) - \gamma A(z')](-\gamma z A'(z')u''(\cdot)] \\ &< -\gamma A'(z')u'(\cdot) + (-\gamma z A'(z')u''(\cdot)] \\ &= -\gamma A'(z')u'(\cdot)[1 + zu''(\cdot)/u'(\cdot)] \\ &< 0\end{aligned}$$

This implies that the perceived marginal benefits of increasing z are decreasing in z', and thus $B_{w,\gamma}(z')$ must be decreasing in z'.

4. $B_{w,\gamma}(0) > 0$, since the assumption that $\psi(0) = 0$ guarantees that the optimal choice of z is interior for any perceived tax schedule. Also, $B_{w,\gamma}(z') < z'$ for large enough z' by the assumption that $\lim_{z\to\infty} \psi'(z) = \infty$.

The above four facts show that $B_{w,\gamma}(z')$ is a continuous and decreasing function, that $B_{w,\gamma}(0) > 0$ and that there exists a $\bar{z}$ large enough such that $B_{w,\gamma}(z) \in [0, \bar{z}]$ for every $z \in [0, \bar{z}]$. Brower's theorem guarantees that a fixed point exists. It must also be unique: If $B_{w,\gamma}(x) = x$ and $B_{w,\gamma}(x') = x'$ for $x < x'$ then because $B_{w,\gamma}(x)$ is a decreasing function of x, it must follow that $0 < B_{w,\gamma}(x) - B_{w,\gamma}(x') = x - x'$, which is a contradiction.

Part 2. Because $u(z - \tilde{T}(z|z')) - \psi(z/w)$ is continuous in w and is strictly concave in z, it follows that $B_{w,\gamma}(z')$ is continuous in w. Because $B_{w,\gamma}$ is continuous in w and has a unique fixed point, its fixed point must be continuous as well. If this were not the case, there would be a $\delta > 0$ such that for any $\epsilon > 0$, the fixed points z_ϵ of $B_{w+\epsilon,\gamma}$ and z of $B_{w,\gamma}$ would always satisfy $|z_\epsilon - z| > \delta$. But $\lim_{\epsilon\to 0} B_{w+\epsilon,\gamma}(z) = B_{w,\gamma}(z) = z$ because B is continuous in w. Now that implies that there exists a series $\epsilon_i \to 0$ such that, without loss of generality, $z_{\epsilon_i} > z$ for all i. But then $B_{w+\epsilon_i,\gamma}(z_{\epsilon_i}) < B_{w+\epsilon_i,\gamma}(z)$ for all i, while $z_{\epsilon_i} - z > \delta$ for all i. This leads to the contradiction that $0 \geq \lim_{\epsilon_i\to 0}(B_{w+\epsilon_i,\gamma}(z_{\epsilon_i}) - B_{w+\epsilon_i,\gamma}(z)) = (z_{\epsilon_i} - z) > \delta$,

Next, we show that for any $z_1 > z_2$ and z', $\Big(u(z_1 - \tilde{T}(z_1|z')) - \psi(z_1/w)\Big) - \Big(u(z_2 - \tilde{T}(z_2|z')) - \psi(z_2/w)\Big)$ is strictly increasing in w. To see this, take the derivative with respect to w:

$$\frac{1}{w^2}\psi'(z_1/w) - \frac{1}{w^2}\psi'(z_2/w)$$

The above equation is positive because ψ' is increasing. Thus for $w_1 < w_2$, $B_{w_1,\gamma}(z) < B_{w_2,\gamma}(z)$ for all z. Moreover, the assumption that $\lim_{z\to\infty} \psi'(z) = \infty$ guarantees that there exists a $\bar{z}$ such that $B_{w_i,\gamma}(z) \in [0, \bar{z}]$ for all $z \in [0, \bar{z}]$ and all $i \in \{1, 2\}$. The statement in the proposition is thus a standard comparative static on fixed points (e.g., Theorem 1 of Villas-Boas, 1997).

Part 3. Continuity in γ follows as in part 2. Next, it follows that for $z_1 > z_2$, $\tilde{T}(z_1|z') - \tilde{T}(z_2|z')$ is decreasing in γ. Thus, for any $z_1 > z_2$ and z', $\Big(c - \psi(z_1/w) - \tilde{T}(z_1|z')\Big) - \Big(c - \psi(z_2/w) - \tilde{T}(z_2|z')\Big)$ is increasing in γ. The result then follows as in Part 2 by Villas-Boas (1997). ■

An observation: It is useful to note that convexity of T plays two important roles in the proof of Proposition 1. First, it ensures that the individual's optimization problem is convex, and thus that B_w is single-valued. In particular, this then ensures that B_w has a closed graph, a property that would not hold for all possible T. Second, convexity of T ensures that B_w is a decreasing function. If T were concave, however, B_w would be an increasing function; and more generally, B_w could be increasing in some regions and decreasing in others for some tax schedules T. Existence and uniqueness are thus not guaranteed for all possible T. To ensure existence, the ME concept would need to be extended to allow for "mixed strategies."

D.2 An Instructive Two-Bracket Model

For purposes of crisp and simple exposition, we will illustrate the main qualitative implications of ironing using a model in which individuals are either low-income earners ($w = w_L$) or high-income earners ($w = w_H$). We assume utility takes the form $G(c - \psi(l))$, where ψ is isoelastic with structural elasticity $\varepsilon < 1$. Motivated by our empirical results, we also assume that workers either correctly perceive taxes or are pure ironers ($\gamma = 0$ or $\gamma = 1$), with $Pr(\gamma = 1) \equiv \gamma_I$ for both wage types.

The policymaker sets a two-bracket income tax given by $T(z) = \tau_1 z$ for $z \leq z^\dagger$ and $T(z) = \tau_1 z^\dagger + \tau_2(z - z^\dagger)$ for $z > z^\dagger$. We assume that the parameters are such that low-income earners fall in the bottom bracket while high-income earners fall in the top bracket. For the low-income earners, we assume that $g(w_L, \gamma) > 1$; that is, the policymaker would transfer additional resources to them if he could do it in a non-distortionary way. For the high-income earner, we assume that w_H is high enough that $(\lambda - G'(z - T(z) - \psi(z/w_H)))\, z$ is increasing in z for all $z \in [z^*(w_H, 0), z^*(w_H, 1)]$. This is a slightly stronger version of the assumption that $g(w_H, \gamma) < 1$ for the high income earners, and must be true for high enough w_H. Throughout, we also assume that τ_2 is lower than the revenue-maximizing tax rate.

Preliminaries

We begin with some preliminary observations we use repeatedly in other proofs.

- For high types, the ATR is $A(z) = \frac{\tau_1 z^\dagger + \tau_2(z - z^\dagger)}{z} = \tau_2 - (\tau_2 - \tau_1)z^\dagger/z$.
- Thus $T'(z) - A(z) = (\tau_2 - \tau_1)z^\dagger/z$ and the perceived MTR by ironing H types is $\tilde{\tau}_2^H(z) = (1 - \gamma)\tau_2 + \gamma A(z) = \tau_2 - \gamma(\tau_2 - \tau_1)z^\dagger/z$.
- $\frac{\partial A}{\partial \tau_2} = (1 - z^\dagger/z)$ and $\frac{\partial \tilde{\tau}_2}{\partial \tau_2} = 1 - \gamma z^\dagger/z$.
- $\frac{\partial A}{\partial z} = (\tau_2 - \tau_1)z^\dagger/z^2$, and $\frac{\partial \tilde{\tau}_2}{\partial z} = \gamma(\tau_2 - \tau_1)z^\dagger/z^2$.
- The structural elasticity ε is given by $\varepsilon = \frac{1}{(1/w^2)\psi''(z/w)} \cdot \frac{(1/w)\psi'(z/w)}{z} = \frac{\psi'(z/w)}{(z/w)\psi''(z/w)}$.

Lemma 1. *For the high types,* $\frac{dz}{d\gamma} = \frac{\varepsilon z^\dagger(\tau_2 - \tau_1)}{1 - \tau_2 + \gamma(1+\varepsilon)(\tau_2 - \tau_1)(z^\dagger/z)}$.

Proof. The high types' first-order condition for choice of z is

$$(1/w)\psi'(z/w) = 1 - \tilde{\tau}_2^H(z) = 1 - \tau_2 + \gamma(\tau_2 - \tau_1)z^\dagger/z.$$

Differentiating implicitly with respect to γ yields

$$(1/w^2)\psi''(z/w)\frac{dz}{d\gamma} = -\frac{\gamma(\tau_2 - \tau_1)z^\dagger}{z^2}\frac{dz}{d\gamma} + (\tau_2 - \tau_1)(z^\dagger/z)$$

and thus

$$\frac{dz}{d\gamma} = \frac{(\tau_2 - \tau_1)(z^\dagger/z)}{(1/w^2)\psi''(z/w) + \frac{\gamma(\tau_2 - \tau_1)z^\dagger}{z^2}} > 0.$$

This establishes that high-income ironers (those with $\gamma = 1$) choose higher labor supply than high-income non-ironers (those with $\gamma = 0$).

We now have

$$\begin{aligned}
\frac{dz}{d\gamma} &= \frac{(\tau_2 - \tau_1)(z^\dagger/z)}{(1/w^2)\psi''(z/w) + \frac{\gamma(\tau_2-\tau_1)z^\dagger}{z^2}} \\
&= \frac{(\tau_2 - \tau_1)(z^\dagger/z)}{(1/\varepsilon)(1/w)(1/z)\psi'(z/w) + \frac{\gamma(\tau_2-\tau_1)z^\dagger}{z^2}} \\
&= \frac{\varepsilon z^\dagger(\tau_2 - \tau_1)}{(1/w)\psi'(z/w) + \varepsilon\gamma(\tau_2 - \tau_1)(z^\dagger/z)} \\
&= \frac{\varepsilon z^\dagger(\tau_2 - \tau_1)}{1 - \tilde{\tau}_2 + \varepsilon\gamma(\tau_2 - \tau_1)(z^\dagger/z)} \\
&= \frac{\varepsilon z^\dagger(\tau_2 - \tau_1)}{1 - \tau_2 + \gamma(\tau_2 - \tau_1)z^\dagger/z + \varepsilon\gamma(\tau_2 - \tau_1)(z^\dagger/z)} \\
&= \frac{\varepsilon z^\dagger(\tau_2 - \tau_1)}{1 - \tau_2 + \gamma(1+\varepsilon)(\tau_2 - \tau_1)(z^\dagger/z)}
\end{aligned}$$

□

Lemma 2. *For the high types,* $\frac{dz}{d\tau_2} = -\frac{z\varepsilon - \gamma z^\dagger \varepsilon}{1-\tau_2+\gamma(1+\varepsilon)(\tau_2-\tau_1)z^\dagger/z} < 0$

Proof. We have

$$\begin{aligned}
\frac{dz}{d\tau_2} &= -\frac{1 - \gamma z^\dagger/z}{\frac{1}{w^2}\psi''(z/w) + \gamma(\tau_2 - \tau_1)z^\dagger/z^2} \\
&= -\frac{1 - \gamma z^\dagger/z}{\frac{1-\tilde{\tau}_2}{z\varepsilon} + \gamma(\tau_2 - \tau_1)z^\dagger/z^2} \\
&= -\frac{z\varepsilon - \gamma z^\dagger\varepsilon}{1 - \tilde{\tau}_2 + \gamma\varepsilon(\tau_2 - \tau_1)z^\dagger/z} \\
&= -\frac{z\varepsilon - \gamma z^\dagger\varepsilon}{1 - \tau_2 + \gamma(\tau_2 - \tau_1)z^\dagger/z + \gamma\varepsilon(\tau_2 - \tau_1)z^\dagger/z} \\
&= -\frac{z\varepsilon - \gamma z^\dagger\varepsilon}{1 - \tau_2 + \gamma(1+\varepsilon)(\tau_2 - \tau_1)z^\dagger/z}
\end{aligned}$$

This is negative because $\gamma z^\dagger < z$. □

Lemma 3. *For the high types,* $\frac{d\tilde{\tau}_2}{d\tau_2} > 0$

Proof. Start with the FOC $1 - \tilde{\tau}_2 = \psi'(z/w)/w$. Now by Lemma 2, z is decreasing in τ_2, and thus the right-hand-side of the FOC is decreasing in τ_2 (by convexity of ψ). Since the right-hand-side is decreasing in τ_2 , $\tilde{\tau}_2$ must be increasing in τ_2. □

Main results:

Claim 1. Labor supply and thus government revenue increase in the propensity to iron.

Proof. Follows by Lemma 1. □

Claim 2. The extra revenue raised due to ironing is raised progressively.

Proof. Notice that in the two-bracket model, the term $T'(z) - A(z)$ is zero for the low-earning types and is positive for high-earning types, indicating that the entire burden of misoptimization falls on the comparatively rich. Combined with the earlier implication that ironing increases government revenue, this additional result establishes that the additional revenue is raised in a manner that is desirable for redistributive purposes. □

Claim 3. Ironing increases social welfare.

Proof. The first two observations—that ironing counteracts the distortionary affects of taxation by raising earnings, and that it increases government revenue in a progressive fashion—lead to the implication that ironing leads to progressive revenue collection. To see this simply in our two-bracket model, notice that ironing has no effect on the behavior of the low-income earners, for whom the marginal and the average tax rate both equal τ_1. The social welfare effect of increasing the γ of a high-income earner (normalized by the marginal value of public funds), for whom the difference between marginal and average tax rates is $(\tau_2 - \tau_1)\frac{z^\dagger}{z}$, is

$$\underbrace{T'(z)\frac{dz}{d\gamma}}_{\text{Gov revenue}} + \underbrace{\frac{d}{dz}G(z - T(z) - \psi(z/w))\frac{dz}{d\gamma}/\lambda}_{\text{Individual utility cost}} = (T'(z) - g(w_H, \gamma)\gamma(T'(z) - A(z)))\frac{dz}{d\gamma} \tag{5}$$

$$= \left(\tau_2 - g(w_H, \gamma)\gamma(\tau_2 - \tau_1)\frac{z^\dagger}{z}\right)\frac{dz}{d\gamma}$$

Now since $g < 1$ for the high income earners, and since $\gamma(\tau_2 - \tau_1)\frac{z^\dagger}{z} < \tau_2$ for all $z \geq z^\dagger$, it follows that the social welfare impact of increasing the $\gamma$of a high income earner is positive. This directly implies that social welfare is increasing in the propensity to iron. □

Claim 4. The revenue and welfare effects of raising tax rates on high incomes are increasing in the propensity to iron.

We prove the result via a series of instructive lemmas that establish intermediate results that further help flesh out the intuition behind how ironers respond to tax rate perturbations. In the first two lemmas we first show that the impact of ironing on earnings is strongest the more convex the the tax schedule is-–that is, the higher is τ_2.

Lemma 4. *For the high type, earnings, $\frac{d}{d\tau_2}\frac{d}{d\gamma}z > 0$ as long as τ_2 is not so high that raising it further would decrease revenue collected from ironers.*

Proof. That τ_2 is lower than the revenue-maximizing tax-rate for ironers implies that $z + \tau_2\frac{dz}{d\tau_2} \geq 0$, and thus that $\frac{1}{z}\frac{dz}{d\tau_2} \geq -\frac{1}{\tau_2}$. Thus

$$
\begin{aligned}
\frac{d}{d\tau_2}\frac{d}{d\gamma}z = &\frac{d}{d\tau_2}\frac{(\tau_2-\tau_1)}{1-\tilde{\tau}_2+\gamma\varepsilon(\tau_2-\tau_1)(z^\dagger/z)} \\
&\propto (1-\tilde{\tau}_2)+\gamma\varepsilon(\tau_2-\tau_1)(z^\dagger/z) \\
&-(\tau_2-\tau_1)\left[-\frac{d\tilde{\tau}_2}{d\tau_2}-\gamma\varepsilon(\tau_2-\tau_1)(z^\dagger/z^2)\frac{dz}{d\tau_2}\right] \\
&\geq (1-\tilde{\tau}_2)+\gamma\varepsilon(\tau_2-\tau_1)(z^\dagger/z)+(\tau_2-\tau_1)\frac{d\tilde{\tau}_2}{d\tau_2} \\
&-\gamma\varepsilon(\tau_2-\tau_1)^2(z^\dagger/z)(1/\tau_2) \\
&=(\tau_2-\tau_1)\frac{d\tilde{\tau}_2}{d\tau_2}+(1-\tilde{\tau}_2)+\gamma\varepsilon(\tau_2-\tau_1)(z^\dagger/z)(1-(\tau_2-\tau_1)/\tau_2) \\
&=(\tau_2-\tau_1)\frac{d\tilde{\tau}_2}{d\tau_2}+(1-\tilde{\tau}_2)+\gamma\varepsilon\left(\tau_2-\tau_1\right)\left(z^\dagger/z\right)(\tau_1/\tau_2) \\
&>0
\end{aligned}
$$

□

To complete our result claim that the impact of ironing on earnings is increasing with τ_2 we now show that as long as τ_2 is below the revenue-maximizing tax-rate, the revenue from ironers will increase in τ_2, a condition of Lemma 4.

Lemma 5. *The tax rate $\bar{\tau}_2^I$ that maximizes revenue from the ironing individuals is higher than the tax rate $\bar{\tau}_2^{NI}$ that maximizes revenue from the non-ironing individuals.*

Proof. Suppose, for the sake of contradiction, that $\bar{\tau}_2^I < \bar{\tau}_2^{NI}$. Then by the previous lemma, $z+\tau_2\frac{dz}{d\tau_2}=0$ for the ironers at $\tau=\bar{\tau}_2^I$, while $z+\tau_2\frac{dz}{d\tau_2}<0$ for the non-ironers at $\tau=\bar{\tau}_2^I$. We will now reach a contradiction if we can show that the revenue extracted from non-ironers is a concave function of τ_2. To that end, note that for the non-ironers, $\frac{dz}{d\tau_2}=-\frac{z\varepsilon}{1-\tau_2}$, and thus

$$
\begin{aligned}
\frac{d}{d\tau_2}\left(z+\tau_2\frac{dz}{d\tau_2}\right) &= 2\frac{dz}{d\tau_2}-\tau_2\frac{d}{d\tau_2}\frac{z\varepsilon}{1-\tau_2} \\
&=2\frac{dz}{d\tau_2}-\tau_2\frac{\varepsilon(1-\tau_2)\frac{dz}{d\tau_2}+z\varepsilon}{(1-\tau_2)^2} \\
&=2\frac{dz}{d\tau_2}-\tau_2\frac{-z\varepsilon^2+z\varepsilon}{(1-\tau_2)^2} \\
&=2\frac{dz}{d\tau_2}-z\varepsilon\tau_2\frac{1-\varepsilon}{(1-\tau_2)^2}<0
\end{aligned}
$$

□

Lemma 6. *Under the assumption that τ_2 is lower than the tax-rate that maximizes revenue, $\frac{d}{d\tau_2}\frac{d}{d\gamma}z>0$ for the high types.*

Proof. Follows directly from the previous two lemmas. □

Having characterized the revenue effects of ironing on increasing τ_2, we now proceed to analyze the welfare effects. We begin by characterizing just the effect of increasing τ_2 on an ironer's welfare:

Lemma 7. *An increase in the tax rate impacts a high type ironer's utility by* $-\frac{dz}{d\tau_2}\gamma(\tau_2-\tau_1)z^\dagger/z$

Proof. We have

$$\begin{aligned}\frac{d}{d\tau_2}(z-T(z)-\psi(z/w)) &= -z+(1-T'(z)-\psi'(z/w)/w)\frac{dz}{d\tau_2}\\ &= -z+\left(1-T'(z)-(1-\tilde{T}'(z))\right)\frac{dz}{d\tau_2}\\ &= -z-\gamma(T'(z)-A(z))\frac{dz}{d\tau_2}\\ &= -z-\gamma(\tau_2-\tau_1)z^\dagger/z\frac{dz}{d\tau_2}\end{aligned}$$

□

We now compute the social marginal welfare effect of increasing τ_2, taking into account the revenue effects.

Lemma 8. *The welfare impact of increasing the tax rate on high types with ironing weight* γ *and social marginal welfare weight* g *is given by* $\frac{dW}{d\tau_2}=\frac{dz}{d\tau_2}\left(\tau_2-g\gamma(\tau_2-\tau_1)z^\dagger/z\right)+(1-g)z$

Proof. Increasing τ_2 mechanically increases revenue by z. This is offset by the substitution to leisure, which leads to a revenue loss of $-\frac{dz}{d\tau_2}\tau_2$. Putting the revenue effects, which are weighted by λ, together with the impact on individual welfare as computed in Lemma 8, which is weighted by $g(z)$ leads to the statement in the proposition. □

We are now ready complete the proof of Claim 4. Lemma 6 implies that a tax rate change impacts ironers less than it does non-ironers. For the welfare effect, note that because $\frac{dz}{d\tau_2}$ is increasing in γ, and because $g\gamma(\tau_2-\tau_1)z^\dagger/z$ is plainly higher for $\gamma=1$ than for $\gamma=0$, the term $\frac{dz}{d\tau_2}\left(\tau_2-g\gamma(\tau_2-\tau_1)z^\dagger/z\right)$ is higher for ironers than for non-ironers. Moreover, because z is higher for ironers than non-ironers by Implication 1, our assumptions imply that $(1-g)z$ is higher for ironers than for non-ironers. This completes the proof of Implication 4.

Claim 5. The revenue and welfare effects of raising tax rates on low incomes are decreasing in the propensity to iron.

Proof: Reasoning analogous to Lemma 7 shows that the impact of increasing τ_1 on the utility of high-income full ironers is given by $-\frac{dz}{d\tau_1}\gamma(\tau_2-\tau_1)z^\dagger/z$. The direct impact on public funds is $z^\dagger$. The indirect substitution effect generates revenue losses given by $-\frac{dz}{d\tau_1}\tau_2$. Putting this together, the social marginal welfare effect stemming from high-income full ironers is given by

$$\begin{aligned}&\frac{dz}{d\tau_1}\left(\tau_2-g(w_H,1)(\tau_2-\tau_1)z^\dagger/z\right)+(1-g(w_H,1))z^\dagger\\ &=\frac{dz}{d\tau_1}\left((1-g(w_H,1))\tau_2+g(w_H,1)A(z)\right)+(1-g(w_H,0))z^\dagger\end{aligned}$$

By comparison, the social marginal welfare effect stemming from non-ironers is simply $(1 - g(w_H, 0))z^\dagger$. Because , ironing leads to lower individual utility $g(w_H, 0) > g(w_H, 1)$ and thus $(1 - g(w_H, 1))z^\dagger < (1 - g(w_H, 0))z^\dagger$. Moreover, $\frac{dz}{d\tau_1} < 0$ for ironers. Thus the social marginal welfare effect from increasing τ_1 is decreasing in the number of (full) ironers.

Claim 6. Ironing increases the welfare consequences of making taxes more progressive.

Proof. This is a direct corollary of Implications 5 and 6. □

D.3 Results for a General Income Tax

We now consider perturbations of any smooth income tax $T(z)$ in a model with a continuum of types. We first solve for the effects of increasing the marginal tax rate by some amount $d\tau$ on all incomes above $z(w^\dagger, 0)$—the earnings of non-ironers with wage $w^\dagger$. We then use this to characterize the optimal nonlinear income tax. We assume that the fraction of ironers is γ_I, which is independent of w. We consider a social welfare function $W = \int \alpha(z, w, 1_\gamma)U(c, z/w)dF(w)$, with α denoting the social welfare weights and $U(c, l) = G(c - \psi(l))$. We let λ denote the social marginal value of public funds. We assume that welfare weights α are such that the social marginal welfare weights $g = \alpha U_c/\lambda$ depend only on z. This assumption follows the Saez (2002a) treatment of multidimensional heterogeneity.

D.3.1 Preliminary Results

As is standard, we define the structural elasticity to be $\varepsilon(z, w) := \frac{\psi'(z/w)/w}{z\psi''(z/w)/w^2}$. This is the elasticity with respect to a linear tax rate of an individual with wage w earning income z. Note that for a utility function $U(c, l) = c - \frac{l^{1+k}}{1+k}$, the elasticity is $\varepsilon \equiv 1/k$.

We next quantify how non-ironers change their earnings in response to a small decrease η in their marginal tax rate. Their FOC is $\psi'(z/w)/w = (1 - T'(z)) + \eta$. The derivative with respect to η is $\psi''(z/w)/w^2 \frac{dz}{d\eta} = (-T''(z))\frac{dz}{d\eta} + 1$. Thus

$$\begin{aligned}
\frac{dz}{d\eta} &= \frac{1}{\psi''(z/w)/w^2 + T''(z)} \\
&= \frac{1}{(1 - T')/(z\varepsilon) + T''} \\
&= \frac{z\varepsilon}{1 - T' + z\varepsilon T''}
\end{aligned}$$

We now analogously compute how ironers respond to a small decrease η in their average tax rate. Consider the ironer's FOC $\psi'(z/w) = w(1 - A(z)) + \eta$. Differentiating that with respect to η yields $\psi''(z/w)/w^2 \frac{dz}{d\eta} = (-A'(z))\frac{dz}{d\eta} + 1$. Now $A = T(z)/z$ and thus $A'(z) = \frac{T'z - T}{z^2} = \frac{T' - A}{z}$. Thus

$$\begin{aligned}
\frac{dz}{d\eta} &= \frac{1}{\psi''/w^2 + \frac{T' - A}{z}} \\
&= \frac{1}{(1 - A)/(z\varepsilon) + (T' - A)/z} \\
&= \frac{z\varepsilon}{1 - A + \varepsilon(T' - A)}
\end{aligned}$$

D.3.2 Welfare Gains of Raising Tax Rates

Let $\bar{\gamma}_I(z)$ be the fraction of ironers with incomes above z. Consider increasing the marginal tax rate by some amount $d\tau$ on all incomes above $z^\dagger$. This has the following effects:

1. A mechanical revenue effect, net of welfare loss, given by $d\tau Pr(z \geq z^\dagger)E\left[(z-z^\dagger)(1-g(z))|z \geq z^\dagger\right]$
2. Substitution toward leisure by the non-ironers. For a given individual, this is $\frac{dz}{d(1-\tau)} = \frac{-z\varepsilon}{1-T'+z\varepsilon T''}$. This leads to an overall loss to public funds given by $d\tau Pr(z \geq z^\dagger|1_\gamma(z)=0)E\left[\frac{z\varepsilon T'(z)}{1-T'(z)+z\varepsilon T''(z)}|z \geq z^\dagger, 1_\gamma(z)=0\right]$.
3. Substitution toward leisures by the ironers. Note that the ironers set $(1-A)-\psi'(z/w)/w=0$, and thus the impact on a given ironer's welfare from a change dz in earnings is $((1-T'(z))-(1-A(z)))dz = (A(z)-T'(z))dz = (A(z)-T'(z))dz$. The impact on public funds is again $T'(z)dz$. The change dz is $\frac{dz}{d(1-A)} \cdot \left(\frac{z-z^\dagger}{z}\right) d\tau = -\frac{z\varepsilon}{1-A+\varepsilon(T'-A)}\left(\frac{z-z^\dagger}{z}\right) d\tau$. This leads to an overall welfare impact of $d\tau Pr(z \geq z^\dagger|1_\gamma(z)=1)E\left[\frac{z\varepsilon\tilde{\tau}(z)}{1-A(z)+\varepsilon(T'-A)}\frac{z-z^\dagger}{z}|z \geq z^\dagger, 1_\gamma(z)=1\right]$, where $\tilde{\tau}(z) = T'(z)+g(z)(A(z)-T'(z)) = (1-g(z))T'(z)+g(z)A(z)$.

Putting this together, the overall effect of an increase $d\tau$ in the marginal tax rate on all incomes above $z^\dagger$ is:

$$\begin{aligned}
&Pr(z \geq z^\dagger)E\left[(z-z^\dagger)(1-g(z))|z \geq z^\dagger\right] \\
&-Pr(z \geq z^\dagger)(1-\bar{\gamma}_I(z^\dagger))E\left[\frac{z\varepsilon T'(z)}{1-T'(z)+z\varepsilon T''(z)}|z \geq z^\dagger, 1_\gamma(z)=0\right] \\
&-Pr(z \geq z^\dagger)\bar{\gamma}_I(z^\dagger)E\left[\frac{z\varepsilon\tilde{\tau}(z)}{1-A(z)+\varepsilon(T'-A)}\frac{z-z^\dagger}{z}|z \geq z^\dagger, 1_\gamma(z)=1\right]
\end{aligned}$$

Note that $\tilde{\tau}(z) \leq T'(z)$ when $g(z)>0$ and $T'(z) \geq A(z)$. Thus,

$$\frac{z\varepsilon T'(z)}{1-T'(z)+z\varepsilon T''(z)} = -\frac{dz}{d(1-T')}|_{1_\gamma=1} > \frac{dz}{d(1-A)}|_{1_\gamma=0} = \frac{z\varepsilon\tilde{\tau}(z)}{1-A(z)+\varepsilon(T'(z)-A(z))}$$

whenever $1-T'(z)+z\varepsilon T''(z) < 1-A+\varepsilon(T'(z)-A(z))$. This occurs at each point z at which T is not too convex. In particular, this inequality holds for any point z on a linear part of the schedule. This establishes that increasing marginal tax rates, particularly in the top bracket, generates higher welfare gains in the presence of more ironers when the tax schedule is progressive ($T'(z) > A(z)$ over the income range under consideration).

D.3.3 Top Marginal Tax Rate

We now follow Saez (2001) to derive the top marginal tax rate. We assume that $\lim_{z\to\infty} T'(z)$ exists and is finite. This implies that $\lim_{z\to\infty} T'(z)-A(z)=0$ and that $\lim_{z\to\infty} T''(z)=0$. We also assume that the elasticity $\varepsilon(z)$ converges to $\bar{\varepsilon}$. Finally, we assume that the social marginal welfare weights for the top converge to $\bar{g}$ and that the propensity to iron is uncorrelated with earnings ability at the top.

We use the Saez (2001) result that $\lim_{z^\dagger\to\infty} E[z|z \geq z^\dagger]/z^\dagger = a/(a-1)$, where a is the pareto parameter of the income distribution. In the limit, the effect of an increase $d\tau$ in the marginal tax rate on all incomes above $z^\dagger$ is then

$$
\begin{aligned}
&Pr(z \geq z^\dagger)E\left[(z - z^\dagger)(1-\bar{g})|z \geq z^\dagger\right] \\
-&Pr(z \geq z^\dagger)(1-\bar{\gamma}_I(z^\dagger))E\left[\frac{z\bar{\varepsilon}T'}{1-T'}|z \geq z^\dagger\right] \\
-&Pr(z \geq z^\dagger)\bar{\gamma}_I(z^\dagger)E\left[\frac{z\bar{\varepsilon}T'}{1-T'}\frac{z-z^\dagger}{z}|z \geq z^\dagger\right]
\end{aligned}
$$

Note that we do not condition on 1_γ in the second and third lines because the schedule is approximately linear at the top, and thus ironers and non-ironers have the same distribution of earnings as long as the propensity to iron is not correlated with earnings ability at the top. Thus, for $z_m := E[z|z \geq z^\dagger]$

$$
(1-\bar{g})(z_m - z^\dagger) - \frac{T'}{1-T'}\varepsilon\left(z_m - \bar{\gamma}_I z^\dagger\right) = 0
$$

from which it follows that

$$
\begin{aligned}
\frac{T'}{1-T'} &= \frac{(1-\bar{g})(z_m - z^\dagger)}{\bar{\varepsilon}(z_m - \bar{\gamma}_I z^\dagger)} \\
&= \frac{(1-\bar{g})(z_m/z^\dagger - 1)}{\bar{\varepsilon}(z_m/z^\dagger - \bar{\gamma}_I)} \\
&= \frac{(1-\bar{g})\left(\frac{a}{a-1} - 1\right)}{\bar{\varepsilon}\left(\frac{a}{a-1} - \bar{\gamma}_I\right)} \\
&= \frac{(1-\bar{g})}{\bar{\varepsilon}(a - \bar{\gamma}_I a + \bar{\gamma}_I)} \\
&= \frac{1-\bar{g}}{[(1-\bar{\gamma}_I)a + \bar{\gamma}_I]\bar{\varepsilon}}
\end{aligned}
$$

Note that since the pareto parameter $a > 1$, the optimal top tax rate is increasing in the propensity to iron. Liebman and Zeckhauser (2004) prove a special case of this result for $\bar{\gamma}_I = 1$: in this case, $\frac{T'}{1-T'} = \frac{1-\bar{g}}{\bar{\varepsilon}}$ at the top.

D.3.4 Optimal Income Tax Derivation

Let $w_N^\dagger$ be the wage of the non-ironers earning $z^\dagger$ and let $w_I^\dagger$ be the wage of the ironers earning $z^\dagger$. Let $z(w, 1_\gamma)$ denote the income chosen by a type $(w, 1_\gamma)$.

For simplicity, we assume here that the propensity to iron is independent of earnings ability w. Let f be the conditional density function of w and let F be the cumulative density function. Let H be the distribution over types $(w, 1_\gamma)$. In terms of wages, the welfare impact of increasing the tax rates by $d\tau$ on all incomes $z \geq z$ is

$$
\begin{aligned}
dW = &-(1-\gamma_I)\int_{w \geq w_N^\dagger} T'(z(w))\frac{dz(w)}{d(1-\tau)}f(w)dw \\
&-(1-\gamma_I)\int_{w \geq w_I^\dagger} T'(z(w))\frac{dz(w)}{d(1-\tau)}f(w)dw \\
&-\int_{z(w,1_\gamma)\geq z(w_N^\dagger,0)} (1-g(z)))(z-z(w_N^\dagger,0))dH(w)
\end{aligned}
$$

The above has to be equal to zero at the optimum for all $w^\dagger$. Thus the derivative of the above with respect to $w^\dagger$ must also equal zero. Differentiating it with respect to $w^\dagger$ leads to

$$
\begin{aligned}
0 = &(1-\gamma_I)T'(z(w_N^\dagger))\frac{dz(w,0)}{d(1-\tau)}|_{w=w_N^\dagger} f(w) + \gamma_I \int_{w\geq w_I^\dagger} \tilde{\tau}(w)\left(\frac{dz(w,0)}{dw}\right)\frac{dz}{d(1-A)}\frac{1}{z(w,1)}f(w) \\
&-\int_{z(w,\gamma)\geq z(w_N^\dagger,0)}(1-g(w,\gamma)))\left(\frac{dz(w_N^\dagger,0)}{dw_N^\dagger}\right)dF \\
&=(1-\gamma_I)T'(z(w))\frac{dz(w,0)}{d(1-\tau)}|_{w=w_N^\dagger} f(w_I^\dagger) + \gamma_I(1-T')\frac{\varepsilon+1}{\varepsilon w_I^\dagger}\frac{dz(w,0)}{d(1-\tau)}|_{w=w_N^\dagger}\int_{w\geq w_I^\dagger}\tilde{\tau}(w)\frac{dz}{d(1-A)}\frac{1}{z(w,1)}f(w)dw \\
&+-(1-T')\frac{\varepsilon+1}{\varepsilon w_N^\dagger}\frac{dz(w,0)}{d(1-\tau)}|_{w=w_N^\dagger}\int_{z(w,1_\gamma)\geq z(w^\dagger,0)}(1-g(w,\gamma)))dH(w) \qquad (6)
\end{aligned}
$$

For $\gamma_I < 1$, rearranging yields

$$
\begin{aligned}
\frac{T'(z^\dagger)}{1-T'(z^\dagger)} = &-\frac{\gamma_I}{1-\gamma_I}\frac{\varepsilon+1}{\varepsilon}\frac{1-F(w_I^\dagger)}{w_N^\dagger f(w_N^\dagger)}E\left[\tilde{\tau}(w)\frac{dz(w,1)}{d(1-A)}\frac{1}{z(w,1)}|z(w,1)\geq z^\dagger\right] \\
&+\frac{1}{1-\gamma_I}\frac{\varepsilon+1}{\varepsilon}\frac{1-\gamma_I F(w_N^\dagger)-(1-\gamma_I)F(w_I^\dagger)}{w_N^\dagger f(w_N^\dagger)}E\left[(1-g(z)|z\geq z^\dagger\right].
\end{aligned}
$$

Instead, when $\gamma_I = 1$, equation (6) reduces to

$$
\int_{w\geq w_I^\dagger}\tilde{\tau}(w)\frac{dz}{d(1-A)}\frac{1}{z(w,1)}f(w)dw - \int_{w\geq w_I^\dagger}(1-g(z(w)))f(w)dw = 0
$$

Differentiating with respect to $w_I^\dagger$ yields

$$
\frac{\varepsilon\tilde{\tau}(w_I)}{1-A+\varepsilon(T'-A)} = (1-g(z^\dagger))
$$

Rearranging generates $\frac{A}{1-A} = \frac{1-g(z)}{\varepsilon}$.

E Welfare Simulations: Robustness Analyses

Alternative Strengths of Redistributive Preferences: Our simulations assume individual utility takes the form $U(z) = log(z - T(z) - \frac{(z/w)^{1+k}}{1+k})$, referred to as "Type 1" utility functions in Saez (2001). While this functional form is common in the public finance literature, one might argue that the assumption of log curvature imposes greater redistributive preferences than may exist in practice. To explore the sensitivity of our conclusions to weaker demand for redistribution, we reconduct our simulation with utility of the form $U(z) = (z - T(z) - \frac{(z/w)^{1+k}}{1+k})^{(1-\rho)}/(1-\rho)$. Log utility corresponds to the case where $\rho = 1$, we reestimate our primary tables under the assumptions that $\rho = 0.5$ or $\rho = 0.25$. As illustrated by these tables, the qualitative importance of both the presence of ironing and its interaction with simplification policies remains.

Table A12: Revenue and Welfare Effects of Ironing: Alt. Redistributive Preferences

Structural Elasticity ($\frac{1}{k}$)	Increase in Tax Rev. (%)	*Net Welfare Increase (%)* Low λ $\lambda = U'_{50}$	— $\lambda = \bar{U}'$	High λ $\lambda = U'_{90}$
Lower Redistributive Preferences: $\rho = 0.5$				
1/2	3.7	3.1	3.2	3.4
1/3	2.5	2.1	2.2	2.3
1/4	1.9	1.6	1.7	1.8
1/5	1.6	1.3	1.4	1.5
Lowest Redistributive Preferences: $\rho = 0.25$				
1/2	3.7	3.0	3.0	3.2
1/3	2.5	2.1	2.1	2.2
1/4	1.9	1.6	1.6	1.7
1/5	1.6	1.3	1.3	1.4

Notes: The numbers presented contrast the revenue collected or welfare attained when comparing a population with perfect tax perceptions against one in which 43% of filers apply the ironing heuristic. Assumed utility model: $U(z) = (z - T(z) - \frac{(z/w)^{1+k}}{1+k})^{(1-\rho)}/(1-\rho)$. The top panel sets $\rho = 0.5$ and the bottom panel sets $\rho = 0.25$. The first column presents the structural elasticity (1/k). The second column presents the additional government revenue collected when the ironers are present. The final three columns present estimates of the increase in social welfare attained due to the presence of ironers, under alternative assumptions on the cost of public funds. Welfare effects are expressed as the percentage of total tax revenues that a social planner would pay to avoid converting all ironers to correct forecasters.

Table A13: Revenue and Welfare Effects Changing to Flat Tax: Alt. Redistributive Preferences

Structural Elasticity ($\frac{1}{k}$)	*All correct forecasters* Δ Tax Rev. (%)	Δ Welfare (%)	*43% ironers* Δ Tax Rev. (%)	Δ Welfare (%)
Lower Redistributive Preferences: $\rho = 0.5$				
1/2	5.2	-2.6	2.9	-5.4
1/3	3.3	-4.9	1.9	-6.7
1/4	2.5	-6.0	1.4	-7.3
1/5	1.9	-6.6	1.1	-7.7
Lowest Redistributive Preferences: $\rho = 0.25$				
1/2	5.2	2.3	2.9	-0.7
1/3	3.3	-0.3	1.9	-2.3
1/4	2.5	-1.5	1.4	-3.0
1/5	1.9	-2.3	1.1	-3.5

Notes: This table summarizes the revenue collected or welfare attained as a result replacing the progressive tax schedule with a linear schedule that would be revenue-neutral assuming no change in behavior. Assumed utility model: $U(z) = (z - T(z) - \frac{(z/w)^{1+k}}{1+k})^{(1-\rho)}/(1-\rho)$. The top panel sets $\rho = 0.5$ and the bottom panel sets $\rho = 0.25$. The first column presents the structural elasticity (1/k). The second and third columns present the additional government revenue and welfare, respectively, resulting from the tax-rate change under the assumption of perfect tax perceptions. The fourth and fifth columns provide analogous calculations under the assumption that 43% of the population irons. Welfare effects are expressed as the percentage of total tax revenues that a social planner would pay to avoid going to the flat tax.

Alternative Flat-Tax Rates: Table 6 analyzes the welfare consequences of moving to a flat tax. The imposed tax rate of 11.06% would be revenue neutral *assuming no behavioral response*. In practice, a policymaker aiming to implement a revenue-neutral flat tax may tailor the rate to account for elastic labor supply. We analyze the sensitivity of our conclusions to rates tailored for these purposes in Table A14. The top panel analyzes the welfare consequences of moving to a flat tax with a rate of 10.49%—the rate that would be revenue neutral assuming optimal response governed by a structural elasticity of $\frac{1}{2}$, our most elastic specification. The bottom panel analyzes the welfare consequences of moving to a flat tax with a rate of 10.85%—the rate that would be revenue neutral assuming optimal response governed by a structural elasticity of $\frac{1}{5}$, our least elastic specification. Across both exercises, we continue to substantially larger welfare costs of moving to the flat tax in the presence of ironing. Under our preferred elasticity of $\frac{1}{3}$, the presence of ironing increases the welfare costs of the flat tax by 11% and 13%, respectively. For comparison, the analysis in Table 6 suggests that the presence of ironing increases welfare costs by 14%.

Table A14: Revenue and Welfare Effects Changing to Flat Tax: Alternative Rates

Structural Elasticity ($\frac{1}{k}$)	*All correct forecasters* Δ Tax Rev. (%)	*All correct forecasters* Δ Welfare (%)	*43% ironers* Δ Tax Rev. (%)	*43% ironers* Δ Welfare (%)
Tax rate: 10.49% (revenue neutral when elasticity = $\frac{1}{2}$)				
1/2	0.0	-12.3	-2.1	-14.6
1/3	-1.8	-14.1	-3.2	-15.6
1/4	-2.7	-14.9	-3.7	-16.1
1/5	-3.2	-15.4	-4.0	-16.3
Tax rate: 10.85% (revenue neutral when elasticity = $\frac{1}{5}$)				
1/2	3.2	-10.8	1.1	-13.2
1/3	1.4	-12.5	0.0	-14.1
1/4	0.5	-13.4	-0.5	-14.5
1/5	0.0	-13.8	-0.8	-14.8

Notes: This table reproduces the analysis of table 6 under alternative assumptions on the rate of the flat tax imposed. Whereas table 6 analyzes a flat tax that would be revenue neutral assuming no behavioral response, this table considers reforms that would be revenue neutral assuming optimal behavior governed by the maximum and minimum elasticities of our considered range.

Omission of Very-High-Income Filers: Due to our sampling structure, our within-sample income distribution closely approximates the U.S. income distribution, with the caveat of being truncated at $250,000. While filers above this income threshold account for only 2% of tax returns, they pay 46% of all federal income tax revenue.[45] Their exclusion influences our estimates in two important ways.

First, if top tax filers exhibit the propensity to iron documented in this paper, the welfare gains associated with ironing become more dramatic. Since the social planner down-weights individual taxpayers' misoptimization costs by their social marginal welfare weights, which are typically assumed to tend to zero for sufficiently rich filers. The welfare-relevant consequence of debiasing a top-2-percent filer would therefore be nearly entirely driven by the fiscal externality component of the equation, guaranteeing that this taxpayers' individual contribution to the welfare effect of debiasing would be negative. We believe that our focus

[45] See https://www.irs.gov/uac/soi-tax-stats-individual-income-tax-returns#prelim.

on within-sample analysis provides the most principled and conservative approach to approximating welfare costs, as it does not rely on untested assumptions that the absolute richest filers exhibit the same misperceptions measured in our population. However, if they do, their effect would only increase the quantitative importance of accounting for ironing.

Second, however, notice that in several of our calculations in Table 5, we benchmark revenue losses or welfare effects against total government revenue. The lack of top-2-percent tax filers in our sample would naturally lead our within-sample revenue forecasts to underestimate true total revenue. Since the omitted range of returns pays 46% of total taxes, rescaling columns 2-4 of Table 5 by 0.54 corrects for their omitted revenue. After this correction, our preferred estimate of the welfare benefit of ironing implies an equivalence with a 1.2% government revenue windfall, and thus still represents a large welfare consideration relative to commonly-studied interventions.